Schema and value: Characterizing the role of the rostral and ventral medial prefrontal cortex in episodic future thinking

Von der Fakultät für Lebenswissenschaften
der Universität Leipzig
genehmigte

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

doctor rerum naturalium
Dr. rer. nat.

vorgelegt von

Dipl.-Psych. Philipp C. Paulus
geboren am 16. Juli 1989 in Stuttgart

Dekan:	Prof. Dr. Marc Schönwiesner
Gutachter:	Prof. Dr. Dr. h.c. Angela D. Friederici
	Prof. Dr. Alexander Strobel
Tag der Verteidigung:	14. Juli 2022

BIBLIOGRAPHISCHE DARSTELLUNG

Philipp Chrysostomos Paulus

Schema and value: Characterizing the role of the rostral and ventral medial prefrontal cortex in episodic future thinking

Fakultät für Lebenswissenschaften

Universität Leipzig

Dissertation

199 Seiten, 320 Literaturangaben, 19 Abbildungen, 8 Tabellen

As humans we are not stuck in an everlasting present. Instead, we can project ourselves into both our personal past and future. Remembering the past and simulating the future are strongly interrelated processes. They are both supported by largely the same brain regions including the rostral and ventral medial prefrontal cortex (mPFC) but also the hippocampus, the posterior cingulate cortex (PCC), as well as other regions in the parietal and temporal cortices. Interestingly, this core network for episodic simulation and episodic memory partially overlaps with a brain network for evaluation and value-based decision making. This is particularly the case for the mPFC. This part of the brain has been associated both with a large number of different cognitive functions ranging from the representation of memory schemas and self-referential processing to the representation of value and affect. As a consequence, a unifying account of mPFC functioning has remained elusive. The present thesis investigates the unique contribution of the mPFC to episodic simulation by highlighting its role in the representation of memory schemas and value. In a first functional MRI and pre-registered behavioral replication study, we demonstrate that the mPFC encodes representations of known people as well as of known locations from participants' everyday life. We demonstrate that merely imagined encounters with liked vs. disliked people at these locations can change our attitude toward the locations. The magnitude of this simulation-induced attitude change was predicted by activation in the mPFC during the simulations. Specifically, locations simulated with liked people exhibited significantly larger increases in liking than those simulated with disliked people. In a second behavioral study, we examined the mechanisms of simulation-based learning more closely. To this end, participants also simulated encounters with neutral people at neutral locations. Using repeated behavioral assessments of participants' memory representations, we reveal that simulations cause an integration of memory representations for jointly simulated people and locations. Moreover, compared to the neutral baseline condition we demonstrate a transfer of positive valence from liked and of negative valence from disliked people to their paired locations. We also provide evidence that simulations induce an affective

experience that aligns with the valence of the person and that this experience can account for the observed attitude change toward the location. In a final fMRI study, we examine the structure of memory representations encoded in the mPFC. Specifically, we provide evidence for the hypothesis that the mPFC encodes schematic representations of our social and physical environment. We demonstrate that representations of individual exemplars of these environments (i.e., individual people and locations) are closely intertwined with a representation of their value. In sum, our findings show that we can learn from imagined experience much as we learn from actual past experience and that the mPFC plays a key role in simulation-based learning. The mPFC encodes information about our environment in value-weighted schematic representations. These representations can account for the overlap of mnemonic and evaluative functions in the mPFC and might play a key role in simulation-based learning. Our results are in line with a view that our memories of the past serve us in ways that are oriented toward the future. Our ability to simulate potential scenarios allows us to anticipate the future consequences of our choices and thereby fosters farsighted decision making. Thus, our findings help to better characterize the functional role of the mPFC in episodic future simulation and valuation.

Contents

Summary

As humans we are not stuck in an everlasting present. Instead, we are able to mentally project ourselves back and forth in time. This ability for mental time travel has great adaptive value as it allows us to imagine potential scenarios that might happen in the future and *pre*-experience what it would feel like if these events actually occurred. By this, episodic simulation can provide strong motivational cues that can render our decisions more farsighted.

The human abilities to remember the past and imagine the future share many similarities: Both functions are based on our episodic memories (i.e., memory of unique experiences that took place at a particular place and time) and semantic knowledge (i.e., generalized knowledge about the typical features of our environment, also referred to as schemas). Moreover, both abilities are characterized by parallel developmental trajectories and are equally affected in ageing. Remembering and simulating are supported by largely the same network of brain regions that include the rostral and ventral medial prefrontal cortex (mPFC), the hippocampus, and the posterior cingulate cortex (PCC). The hippocampus plays a key role in the representation of episodic memories, whereas the mPFC has been implicated in the mediation of schematic knowledge. Theoretical accounts of the human ability for mental time travel have argued that remembering and simulating might both be supported by the same constructive memory system. This memory system adaptively uses memory of past events to construct simulations of potential future happenings. It has thus been argued that our memories of the past primarily serve us in ways that are oriented toward the future.

Interestingly, the core network of brain regions that supports remembering and simulating, partially overlaps with a brain system for evaluation and value-based decision making. Particularly the mPFC has been implicated in both mnemonic processes as well as in the representation of a domain general value signal and affect. The question how the mPFC might support such seemingly disparate functions has remained unsolved.

In the present thesis I have taken two complementary perspectives on the human ability for episodic simulation and attempted to extend our understanding of the role of the mPFC in mnemonic and evaluative processes: (i) I examined the potential of episodic simulations to serve as an imaginary parallel to actual experiences. Specifically, I tested whether we can learn from simulated experiences, much as we learn from actual past experience. (ii) I examined the neural mechanisms that support episodic simulation. Specifically, I investigated the structure of neural memory representations that are activated whenever we imagine a hypothetical event. The individual projects may be summarized as follows.

Study 1: Forming real-life attitudes via episodic simulations

Humans can vividly imagine hypothetical events. Episodic simulation is supported by a core network of brain regions that includes the mPFC. Episodic simulations are based on knowledge about our environment (e.g., of known people) to construct simulations of potential happenings (e.g., meeting a known person at a familiar location). Participants provided names of people they personally know and locations from their everyday environment. Participants rated how well they know and how much they liked these people and locations. We then selected liked and disliked people and paired each one with a neutral location. Participants returned for the simulation session and repeatedly simulated location specific interactions with the people at their respective paired neutral location while being scanned with fMRI. After the simulations, participants rated their liking of the simulated people and the locations again.

The behavioral results revealed a general increase in liking of the locations that was significantly larger for those locations that had set the stage for imaginary encounters with liked people. The neuroimaging results revealed a critical contribution of the mPFC to this simulation-induced attitude change: Activation in the mPFC scaled with the liking of the simulated person and was predictive of the subsequent change in attitude toward the location. Moreover, multivariate analyses using representational similarity analysis (RSA) provided evidence that more similar representations emerged in the mPFC whenever participants simulated the same person or location as compared to different exemplars of the same category. This latter finding suggests that the mPFC encodes representations of these known people and locations. Given that the selected sets of liked and disliked people in the fMRI study also differed with regard to their familiarity, we had to rule out the unlikely possibility that the observed effects were caused by differences in familiarity. We therefore conducted a preregistered replication study where we carefully matched familiarity in both sets and replicated the original findings in a larger behavioral sample.

Together these results suggest that episodic simulations can shape attitudes toward the very elements that these simulations had been based on. Episodic simulations yielded a transfer of positive valence from the simulated people toward their paired locations. This transfer of affective valence was mediated by the mPFC.

Study 2: Simulation-based learning influences real-life attitudes

This behavioral study aimed at clarifying the mechanisms that support simulation-induced attitude changes. The previous study had demonstrated evidence for a *positive* transfer of valence from the people toward their paired locations. Here, we extended the previous design with a neutral baseline condition to examine whether simulations can also induce a transfer of

negative affective valence. Moreover, we instructed participants to repeatedly arrange the names of the people and the locations on two-dimensional surfaces to indicate how much they associate them. We also recorded skin conductance responses during the episodic simulations as a measure of the emotional arousal of our participants.

The results of the behavioral arrangement tasks revealed an overall integration of the memory representations of the jointly simulated people and locations that was not dependent on the valence of the simulated people. We also obtained evidence for the anticipated transfer of valence effects: Compared to the neutral baseline condition, locations that were simulated with liked people were characterized by a positive shift in liking and locations that were simulated with disliked people were characterized by a negative shift in liking. Analyses of the emotional responses revealed overall stronger emotional arousal both when participants simulated scenarios with liked as well as with disliked people as compared to the neutral baseline condition. Participants rated their simulations with disliked people as unpleasant, simulations featuring neutral people as neutral, and simulations featuring liked people as pleasant. A causal mediation analysis revealed that the transfer of valence from the person toward the location was mediated via this perceived pleasantness. Thus, contingent on the valence of the simulated person, episodic simulations induced an affective experience that caused a change in attitude toward the location.

In sum, merely imagined experiences can induce affective states that shape real-life attitudes much like actual experiences can. Under some circumstances, episodic simulations might therefore be regarded an imaginary parallel to actual experiences.

Study 3: Value shapes the structure of schematic representations in the mPFC

This fMRI study examined the structure of neural memory representations that are activated whenever we imagine hypothetical events. The study is based on the observation that the mPFC is involved both in the representation of memory schemas and in the computation of a domain general value signal. Here, we hypothesized that the mPFC might subserve these seemingly disparate functions by encoding schematic memory representations where representations of individual exemplars are closely intertwined with a representation of their value: value-weighted schematic representations.

Participants provided names of people and locations they personally know from their everyday life. They then provided fine-grained behavioral assessments of their relationships by arranging their names on two-dimensional surfaces. From these arrangements we determined the *centrality* of each person and location to their respective environment. Moreover, we instructed participants to indicate how well they know (as a measure of *experience*) and how

much they like (as a measure of *affective value*) each individual person and location. Participants then reinstated each exemplar's neural memory representation by vividly simulating a typical scenario of interacting with the people or being at the locations. Using functional MRI, we measured the ensuing multi-voxel activation pattern in each of these simulation trials as a proxy measure of participants' neural memory representations.

Using RSA, we replicate our earlier finding and demonstrate that more similar representations emerge in the mPFC whenever we simulate the same person or location as compared to simulations featuring different exemplars of the same category. Moreover, we demonstrate that the structure of neural memory representations in the mPFC reflects a combination of how *central* a person is to the respective environment, how much *experience* we have with it, as well as how much we *like* it. Thus, people that are central to our social network, that we know well, and like much are also overall more strongly embedded in the neural memory representation in the mPFC. Critically, only the structure of neural memory representations in the mPFC was best accounted for by the combination of these three features. This was not only true for the simulated people, but also for the simulated locations. In contrast, representations in the hippocampus and PCC – two regions that have also been implicated in the mediation of mnemonic and evaluative processes – other models were better suited to account for the structure of neural representations.

In sum, the findings of this study indicate that the mPFC encodes schematic memory representations where knowledge about individual exemplars is closely intertwined with a representation of their value. These value-weighted schematic representations may provide an account for the overlapping involvement of the mPFC in both mnemonic and evaluative functions.

Discussion

How we perceive our environment and the people that live in it is shaped by our memories of past events. Across unique experiences we gradually learn what our environment is typically like. This knowledge is encoded in generalized schematic memory representations. The results of the three projects reported in this thesis demonstrate that we can adaptively use this knowledge about our environment to simulate events that might happen in the future.

The mPFC is a central node in both a brain network for remembering and simulating as well as in a brain network for valuation and value-based decision making. The present thesis has provided evidence that the mPFC might support these seemingly disparate functions by encoding schematic representations within which knowledge about individual exemplars is

closely intertwined with a representation of their value. These representations, in turn, might support a simulation-based learning mechanism that can shape real-life attitudes.

To conclude, the present thesis has demonstrated how mnemonic and evaluative processes interact when we imagine hypothetical events that may happen in our personal future. By this the presented results have provided evidence for the central argument of the constructive episodic simulation hypothesis: Our memories of the past serve us in adaptive ways that are oriented toward the future.

Deutsche Zusammenfassung

Als Menschen sind wir nicht in einer sich ewig wiederholenden Gegenwart gefangen. Vielmehr sind wir in der Lage, uns mental sowohl in unsere persönliche Vergangenheit als auch in unsere Zukunft zu projizieren. Diese Fähigkeit zur mentalen Zeitreise hat großen adaptiven Nutzen: Sie erlaubt uns Ereignisse in der Zukunft zu simulieren und so bereits im Hier und Jetzt zu antizipieren, wie es sich anfühlen würde, wenn diese Ereignisse sich tatsächlich ereignen würden. Episodische Simulationen ermöglichen uns so Zugriff auf motivationale Anreize, die uns unsere Entscheidungen stärker an langfristigen Zielen ausrichten lassen.

Die menschliche Fähigkeit sich an die Vergangenheit zu erinnern und die Fähigkeit Ereignisse in der Zukunft zu simulieren sind einander sehr ähnlich: Beide Funktionen beruhen auf unseren episodischen Erinnerungen (d.h. Erinnerungen an einzelne Ereignisse, die an einem bestimmten Ort und zu einer bestimmten Zeit stattfanden) und unserem semantischen Wissen (d.h. generalisiertem Wissen, das die typischen Eigenschaften unserer Umgebung abbildet und auch als Schema bezeichnet wird). Darüber hinaus weisen beide Fähigkeiten ähnliche Entwicklungen über die Lebensspanne auf und sind im hohen Alter gleichermaßen beeinträchtigt. Unsere Fähigkeiten Vergangenes zu erinnern und uns Zukünftiges vorzustellen, werden darüber hinaus von einem gemeinsamen Netzwerk verschiedener Gehirnregionen unterstützt. Dieses neuronale Netzwerk besteht unter anderem aus dem rostralen und ventralen Teil des medialen präfrontalen Kortex (mPFC), dem Hippocampus und dem posterioren cingulären Kortex (PCC). Während der Hippocampus vorrangig im Zusammenhang mit der Repräsentation episodischer Gedächtnisinhalte assoziiert ist, wird der mPFC mit schematischen Gedächtnisinhalten in Zusammenhang gebracht. Formale Theorien über die Ähnlichkeiten zwischen der Fähigkeit zum Erinnern und der Fähigkeit zur episodischen Simulation der Zukunft argumentieren, dass beide Fähigkeiten unterschiedliche Manifestationen eines gemeinsamen konstruktiven Gedächtnissystems sind. Dieses Gedächtnissystem nutzt unsere Erinnerungen an vergangene Erlebnisse auf adaptive Weise, um Vorhersagen über die Zukunft zu treffen. Wir können uns also vermutlich an die Vergangenheit erinnern, damit wir in der Zukunft die gleichen Fehler nicht erneut machen müssen.

Interessanterweise überlappt das neuronale Netzwerk, das es uns ermöglicht uns an die Vergangenheit zu erinnern und die Zukunft zu simulieren, mit einem neuronalen Netzwerk, das evaluative Prozesse und Entscheidungsprozesse unterstützt. Insbesondere der mPFC wird sowohl mit Gedächtnisprozessen als auch mit der Endkodierung des subjektiven Werts unterschiedlicher Verhaltensalternativen, sowie mit der Repräsentation affektiver Zustände in Zusammenhang gebracht. Die Frage *wie* der mPFC solche auf den ersten Blick vollkommen verschiedenartigen Funktionen unterstützt, ist weitgehend unbeantwortet.

In der vorliegenden Dissertation habe ich anhand zweier komplementärer Ansätze versucht unser Verständnis von der Rolle des mPFC sowohl für Gedächtnis- als auch für Bewertungsprozesse zu erweitern: (i) Ich habe untersucht, inwiefern episodische Simulationen als Ersatz für tatsächliche Erlebnisse dienen können und (ii) ich habe die neuronalen Mechanismen episodischer Simulation untersucht. Mein besonderes Interesse galt in diesem Zusammenhang denjenigen Gedächtnisrepräsentationen, die aktiviert werden, wann immer wir uns hypothetische Ereignisse in der Zukunft vorstellen. Die einzelnen Projekte können folgendermaßen zusammengefasst werden:

Studie 1: Einstellungsänderungen durch episodische Simulationen

Menschen können sich eine Vielzahl möglicher Ereignisse, die sich eventuell in der Zukunft ereignen könnten, lebhaft vorstellen. Diese Fähigkeit wird von einem neuronalen Netzwerk unterstützt, das den mPFC umfasst. Episodische Simulationen basieren auf unserem Wissen über unsere Umgebung (z.B. über Personen, die wir persönlich kennen) und ermöglichen es uns, potenzielle Szenarien zu simulieren (z.B. diese bekannte Person an einem spezifischen Ort zu treffen, den wir aus unserem Alltag kennen). In Studie 1 fertigten Versuchsteilnehmer Listen von Personen und Orten an, die sie aus ihrem Alltag kennen. Anschließend schätzten die Versuchsteilnehmer ein, wie gut sie diese Personen und Orte kennen und wie sehr sie diese mögen. Auf Basis dieser Einschätzungen wählten wir besonders beliebte sowie besonders unbeliebte Personen aus und paarten diese mit Orten, die als neutral bewertet wurden. In einer folgenden Sitzung simulierten die Versuchsteilnehmer spezifische Situationen, in denen sie mit den Personen auf eine Art interagierten, die typisch für den jeweiligen Ort war. Während dieser Phase wurde die Gehirnaktivität der Versuchsteilnehmer mittels funktioneller MRT gemessen. Nach diesen Simulationen schätzten die Versuchsteilnehmer erneut ein, wie sehr sie die simulierten Personen und Orte mögen.

Der Vergleich der beiden Bewertungen der Orte hinsichtlich der Beliebtheit zeigte einen generellen Anstieg in der Bewertung aller Orte. Dieser Anstieg in der Beliebtheit war signifikant größer für diejenigen Orte, die mit besonders beliebten Personen simuliert worden waren. Die Ergebnisse der bildgebenden Verfahren weisen auf eine zentrale Beteiligung des mPFC für diese Einstellungsänderung hin: Aktivität im mPFC sagte die Stärke der Einstellungsänderung gegenüber dem Ort vorher. Darüber hinaus zeigten multivariate Analysen mittels *Representational Similarity Analysis (RSA)*, dass ähnliche Repräsentationen im mPFC reaktiviert wurden, wann immer die gleiche Person oder der gleiche Ort simuliert wurde. Diese Ähnlichkeit überstieg die zu erwartende Ähnlichkeit bei der Simulation von Exemplaren der gleichen Kategorie (Personen oder Orte). Dieser Befund legt die Interpretation

nahe, dass der mPFC selbst Repräsentationen bekannter Personen und Orte enkodiert. Die ausgewählten beliebten und unbeliebten Personen unterschieden sich jedoch auch hinsichtlich ihrer Bekanntheit voneinander. Um auszuschließen, dass die beobachteten Effekte auf diesem Unterschied beruhten, führten wir eine präregistrierte Replikationsstudie durch. In dieser Studie stellten wir sicher, dass es keine Unterschiede in der Bekanntheit der ausgewählten Personen gab. Diese Studie konnte die ursprünglichen Ergebnisse in einer größeren Stichprobe replizieren.

Gemeinsam legen diese beiden Experimente den Schluss nahe, dass episodische Simulationen Einstellungen gegenüber den Exemplaren verändern können auf denen diese Simulationen ursprünglich beruhten. Episodische Simulationen verursachen einen Transfer positiver Valenz von der Person zum jeweiligen gepaarten Ort. Dieser Valenztransfer wird vom mPFC vermittelt.

Studie 2: Simulationsbasiertes Lernen von Einstellungen

Diese behaviorale Studie wurde mit dem Ziel durchgeführt, zu klären, welche Mechanismen der beobachteten Einstellungsänderung in Studie 1 zugrunde liegen. Studie 1 hatte bereits Evidenz für einen Transfer *positiver* Valenz geliefert. Hier erweiterten wir unser ursprüngliches Studiendesign mit einer neutralen Baselinebedingung, um zu überprüfen ob episodische Simulationen auch zu einem Transfer *negativer* Valenz führen können. Darüber hinaus instruierten wir die Versuchsteilnehmer auch dazu, wiederholt die Namen der Personen und der Orte auf zweidimensionalen Flächen so anzuordnen, dass ihre Positionen widerspiegeln, wie die Personen und Orte zusammengehören. Während den episodischen Simulationen erhoben wir auch die Hautleitfähigkeit unserer Versuchsteilnehmer als Maß emotionaler Erregung.

Die Ergebnisse der Anordnungsaufgabe wiesen auf eine allgemeine Integration der Gedächtnisrepräsentationen gemeinsam simulierter Personen und Orte hin. Diese Integration war nicht abhängig von der emotionalen Valenz der simulierten Personen. Darüber hinaus erhielten wir auch Evidenz für die vorhergesagten Veränderungen der Einstellungen unserer Probanden: Verglichen mit der neutralen Baselinebedingung wiesen Orte, die mit beliebten Personen simuliert wurden, eine Zunahme in der Beliebtheit auf und Orte, die mit unbeliebten Personen simuliert wurden, wiesen hingegen eine Abnahme in der Beliebtheit auf. Analysen der emotionalen Reaktionen unserer Versuchsteilnehmer wiesen ein höheres Maß emotionaler Erregung in den beiden emotionalen Bedingungen im Verhältnis zur neutralen Bedingung auf. Darüber hinaus gaben die Versuchsteilnehmer an, dass Simulationen mit beliebten Personen angenehm, Simulationen mit neutralen Personen neutral und Simulationen mit unbeliebten Personen als unangenehm erlebt wurden. Mittels einer kausalen Mediationsanalyse konnten wir

aufdecken, dass diese erlebte emotionale Qualität den Transfer affektiver Valenz von der Person zum jeweiligen gepaarten Ort erklärt. Dies legt nahe, dass die beobachteten Veränderungen hinsichtlich der Einstellung unserer Versuchsteilnehmer gegenüber den Orten im emotionalen Erleben während der Simulationen begründet sind.

Zusammenfassend lässt sich festhalten, dass ausschließlich vorgestellte Erlebnisse emotionale Zustände induzieren können, die Einstellungen gegenüber Orten aus unserem tagtäglichen Leben verändern. Unter den beschriebenen Umständen können episodische Simulationen als imaginärer Ersatz für tatsächlich Erlebtes betrachtet werden.

Studie 3: Die Struktur schematischer Repräsentationen im mPFC bildet Bewertungsprozesse ab

Diese funktionelle Bildgebungsstudie untersuchte die neuronalen Repräsentationen, die reaktiviert werden, wann immer wir uns hypothetische Szenarien vorstellen. Die Studie beruht auf der Beobachtung, dass der mPFC sowohl in der Enkodierung generalisierter Gedächtnisrepräsentationen, als auch in Bewertungsprozesse involviert ist. Hier stellten wir die Hypothese auf, dass der mPFC diese unvereinbar erscheinenden Funktionen unterstützt, indem diese Gehirnregion Schemata enkodiert in denen Wissen über unsere unmittelbare Umgebung (z.B. bekannte Personen oder Orte) eng mit Merkmalen ihrer Beliebtheit verwoben ist. Diese Repräsentationen nennen wir Schemata mit Bewertungskomponente.

In dieser Studie fertigten Versuchsteilnehmer Listen persönlich bekannter Personen und Orte an. Anschließend ordneten die Versuchsteilnehmer diese Personen und Orte hinsichtlich mehrerer Merkmale an. Zunächst gaben sie mittels Anordnungen auf zweidimensionalen Flächen an, wie Personen und Orte zusammengehören. Aus diesen Anordnungen bestimmten wir anschließend, wie *zentral* die Personen für das soziale Netzwerk bzw. die Orte für das Umfeld der Versuchsteilnehmer sind. Anschließend gaben die Versuchsteilnehmer an, wie gut sie die Personen und Orte *kennen* und wie sehr sie diese *mögen*. An einem folgenden Tag simulierten die Versuchsteilnehmer bei einer Sitzung im Magnetresonanztomographen lebhafte Episoden mit den Personen oder an den Orten. Diese Aufgabe reaktivierte die neuronalen Repräsentationen der jeweiligen Personen oder Orte. Wir bestimmten die daraus resultierenden neuronalen Aktivierungsmuster mittels fMRT.

Mit Hilfe von RSA konnten wir auch in dieser Studie Evidenz dafür finden, dass ähnliche Repräsentationen im mPFC reaktiviert werden, wann immer wir die gleiche Person oder den gleichen Ort simulieren. Darüber hinaus konnten wir zeigen, dass sich die Ähnlichkeiten neuronaler Repräsentationen im mPFC mittels einer Kombination der *Zentralität*, der *Vertrautheit*, sowie der *Beliebtheit* der jeweiligen Personen und Orte

vorhersagen ließ. Dies bedeutet z.B., dass Repräsentationen von Personen, die besonders zentral für unser soziales Netzwerk sind, die wir besonders gut kennen und die wir besonders mögen auch besonders stark in die neuronale Repräsentation unseres sozialen Netzwerks im mPFC eingebunden sind. Interessanterweise war dies nur im mPFC der Fall. Im Hippocampus und PCC, zwei anderen Regionen, die ebenfalls mit Gedächtnis- und Bewertungsprozessen in Zusammenhang gebracht werden, waren andere Merkmale besser geeignet die neuronalen Repräsentationen zu beschreiben.

Zusammenfassend lässt sich aus den vorliegenden Ergebnissen schlussfolgern, dass der mPFC Schemata repräsentiert, in denen Wissen über unsere Umgebung in enger Verknüpfung mit einer Repräsentation ihrer Beliebtheit enkodiert sind. Diese Schemata mit Bewertungskomponente könnten eine Erklärung für die Relevanz des mPFC für Gedächtnis- und Bewertungsprozesse darstellen.

Diskussion

Wie wir unsere Umgebung und die Menschen in dieser Umgebung bewerten und wahrnehmen, wird durch unsere Erinnerungen an vergangene Ereignisse beeinflusst. Über individuelle Erfahrungen hinweg erlernen wir über die Zeit hinweg, wie unsere Umgebung typischerweise ist. Dieses Wissen wird in generalisierten schematischen Gedächtnisrepräsentationen enkodiert. Die Ergebnisse der drei beschriebenen Projekte legen den Schluss nahe, dass wir dieses Wissen über unsere Umgebung in adaptiver Weise nutzen können, um uns lebhaft potenzielle Ereignisse in der Zukunft vorzustellen.

Der mPFC ist sowohl eine zentrale Gehirnregion in einem neuronalen Netzwerk für Gedächtnisprozesse, als auch Teil eines neuronalen Netzwerks für Entscheidungs- und Bewertungsprozesse. Die Ergebnisse der vorliegenden Dissertation legen nahe, dass der mPFC solche verschiedenartigen Prozesse unterstützt, in dem er Schemata enkodiert in denen Wissen über unsere unmittelbare Umgebung eng mit Bewertungen verknüpft ist. Diese Repräsentationen könnten dem beschriebenen Lernmechanismus zugrunde liegen, der es uns ermöglicht von Simulationen in ähnlicher Weise zu lernen, wie wir von tatsächlichen Erlebnissen lernen.

Die vorliegende Dissertation hat aufgezeigt, wie Gedächtnis- und Bewertungsprozess im menschlichen Gehirn interagieren, wenn wir potenzielle Szenarien in der Zukunft simulieren. Die dargestellten Ergebnisse liefern dadurch Evidenz, dass der Mensch über ein konstruktives Gedächtnissystem verfügt, dessen adaptiver Wert sich in der Fähigkeit zeigt, die Zukunft zu antizipieren.

Chapter 1

1 Introduction

As humans we are not stuck in an everlasting present. Instead, we are able to mentally project ourselves back and forth in time. This ability, also known as mental time travel, allows us to remember the past and imagine the future. This fascinating human capacity enables us to mentally relive episodes that have or could have happened in our personal past and permits us to mentally project ourselves into a hypothetical future. Traditionally, research on memory has primarily focused on the past. However, this has changed over the course of the past twenty years with researchers rediscovering the hypothesis that our memories of the past primarily serve us in ways that are oriented toward the future. Moreover, this prospective perspective on episodic memory has led to the rediscovery that remembering is also a fundamentally constructive process. This line of research has not only revealed that simulations of the future are based on our memories of the past, but demonstrated that they are also supported by largely the same network of brain regions.

The ability for mental time travel is a great adaptive device that allows us to anticipate the consequences of experiences we have already made in the past. However, and more importantly, it also enables us to anticipate consequences of experiences we have never had and probably should never have. Our ability to simulate experiences allows us to know that we would prefer a donut filled with vanilla pudding over one filled with sauce hollandaise, that missing a train would be preferable to being involved in an accident, that we would rather win the lottery than being devoured by a lion, and that it is worth brushing our teeth every day to evade tooth ache and the shrieking sound of the dentist's drill.

Admittedly, only some of these examples have immediate consequences for our survival. However, a memory system that allows us to use our memories of past experiences to construct simulations of imaginary happenings, provides us with means to anticipate what the future might be like. Thus, we can close our eyes, imagine a potential scenario and pre-experience the emotional quality (*what it would feel like*) if it actually happened already in the here and now. By this, simulations provide us with emotional cues that can serve as motivators for farsighted decisions and may render our behavior congruent with the anticipated needs of our future selves.

In this thesis, I take two complementary perspectives on our ability for episodic simulation: (i) I examine the potential of simulations to serve as an imaginary parallel to actual experience. Specifically, I ask the question whether we can learn from simulated experiences much as we learn from actual past experience. (ii) I examine the neural system that supports episodic simulations. Specifically, I investigate the structure of memory representations that are activated whenever we imagine hypothetical events. Where in the human brain are these

representations encoded? What shapes the structure of these memory representations? To answer these questions, I conducted three empirical projects using different versions of an episodic simulation task in combination with fine grained behavioral assessments, functional neuroimaging, and psychophysiological measures.

In this introductory section, I will review evidence that remembering the past and simulating the future are a common process that is supported by the same network of brain regions. Moreover, I will examine evidence that this core network for episodic memory and simulation partially overlaps with a neural network that supports value-based decision making. In this regard, I will specifically highlight the contribution of the rostral and ventral medial prefrontal cortex (mPFC) that has been associated both with the encoding of memory schemas and the representation of value and affect.

Following the introductory section, the main part of the thesis reports results from a series of two independent studies that investigated whether we can learn from episodic simulations much as we learn from actual past experience. Moreover, I report evidence for a central role of the mPFC in supporting this kind of learning. In a third empirical project that employed functional neuroimaging, fine-grained behavioral assessments, and multivariate pattern analysis techniques, I examined the representational format of memory representations in the mPFC. In that paper, we argue that the mPFC encodes schematic representations within which information about individual known people and locations are inherently intertwined with a representation of their value. This kind of memory representation would be in accordance with the mPFC's involvement in both memory functions and valuation. In the final part of the thesis, I will discuss these findings in the broader context of the existing literature on episodic simulation and mPFC functioning. Ultimately, this thesis attempts to shed more light on the adaptive functions that are supported by our ability for episodic simulation: Allowing us to anticipate the affective consequences – "what it would feel like" – if events actually happened in the future. These affective cues, in turn, may motivate adaptive behavior and render our decisions more farsighted.

1.1 Memory and the medial temporal lobes

Research on human memory systems has long been based on the examination of individual patients with focal brain lesions. The probably most intensely studied individual in the history of neuroscience is patient HM (see Squire, 2009 for a comprehensive overview about HM's contribution to the neuroscience of memory). In a desperate attempt to treat his epileptic seizures, large parts of HM's medial temporal lobes including the hippocampus, the amygdala, and the entorhinal cortex were surgically removed in both hemispheres. While the surgery

relieved him of the seizures, it left him largely unable to form new memories. Moreover, he was almost entirely unable to remember anything but fragments of factual knowledge – also known as semantic knowledge – about his personal past (Milner, 1972; Squire, 1992, 2009). Apart from this striking impairment, HM still retained the majority of his cognitive functions: He was able to communicate using a large repertoire of vocabulary, he was able to recall factual knowledge about the world as well as about himself and even able to learn new skills (Milner, 1972; Squire, 2009). Given the selectivity of his impairment, HM's case contributed to our understanding that many memory functions are supported by the medial temporal lobes and the hippocampus in particular. Moreover, given the selectivity of his impairments, his case also revealed that memory is a capacity that is largely independent of other cognitive functions (Milner, 1972).

Later, the findings in HM were mirrored in descriptions of a different patient KC (see also Klein et al., 2002 for a related case). This patient lost his entire repertoire of episodic memories after sustaining a traumatic brain injury from a motorcycle accident. His injuries were more widespread, but crucially also included his hippocampus bilaterally as well as adjacent parts of his medial temporal lobes. Similar to patient HM, KC was still able to recall some factual knowledge about the world as well as about himself (Rosenbaum et al., 2005). Interestingly, above and beyond these memory impairments, the descriptions of KC also revealed a selective impairment of his ability to imagine events that could take place in his personal future (Buckner, 2010; Mullally & Maguire, 2014; Tulving, 1985). This suggests that intact medial temporal lobes are also a necessary condition for the ability to imagine events in the future. These observations in individual patients were later replicated in systematic comparisons of groups of patients with specific bilateral lesions to the hippocampus and healthy control participants (Hassabis, Kumaran, Vann, et al., 2007). The main results indicate that patients' descriptions of remembered past as well as imagined future events lacked much of the typical episodic details and spatial context that is apparent in event descriptions of healthy control participants.

In sum, these findings suggest a common neural system that supports our ability to remember past events and imagine the future that is dependent on the hippocampus. Both abilities are based on our autobiographical memories that comprise of both generalized semantic knowledge and episodic memories of unique past events. Lesions to the medial temporal lobes, specifically the hippocampus, selectively impair access to episodic memories leaving semantic knowledge relatively unharmed. Thus, these types of memory are most likely stored in separate memory systems.

1.2 Impairments of mental time travel after lesions outside the medial temporal lobes

From the evidence provided so far, one might conclude that the medial temporal lobes and the hippocampus are the only neural structures required for mental time travel. However, lesions to the mPFC can also cause profound mnemonic impairments (Bertossi, Tesini, et al., 2016). Specifically, mPFC lesions are associated with confabulations – a condition where patients produce erroneous memories in the absence of conscious awareness of their falsehood (Burgess & Shallice, 1996; Ghosh et al., 2014).

But what causes these impairments? The mPFC plays a key role in mediating remote memories (Frankland & Bontempi, 2005). Across both human (Brod & Shing, 2018; Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017; van Kesteren et al., 2012) and rodent work (Farovik et al., 2015; Tse et al., 2007), the mPFC has consistently been linked to the encoding of generalized semantic memory representations also known as *schemas*. These memory representations can be defined as "adaptable associative networks of knowledge extracted over multiple similar experiences" (Ghosh et al., 2014, p. 12057). Thus, schemas are superordinate memory representations that reduce the complexities of several related episodes into simplified and generalized representations that no longer pertain to unique instances in space and time (Ghosh & Gilboa, 2014). For example, across all events that we experienced in kitchens, we have formed a schema of what kitchens are like. Upon entering a new kitchen, we can use our *kitchen schema* to anticipate where to find the cutlery, where to expect the dishwasher, and – more generally – what activities are typically performed in kitchens. Schemas thus provide us with a scaffold or template that can be used to make sense of ongoing experience. Moreover, these templates allow us to anticipate what we should expect in a given situation – a crucial component to our ability for adaptive and flexible behavior (Bartlett, 1932; Piaget, 1952). Finally, the comparison of our ongoing experience with these templates allows us to identify relevant new information which in turn facilitates their encoding into pre-existing knowledge structures (van Kesteren et al., 2012). By this, schemas greatly influence how we retain new information and how we remember it later on (Bartlett, 1932; Gilboa & Marlatte, 2017).

Lesions to the mPFC disrupt schema facilitated memory processes and can cause confabulations. Individuals affected by this condition produce vivid but highly inaccurate recollections of events that never happened (Burgess & Shallice, 1996; Schacter & Addis, 2007). These erroneous memory productions are thought to result from the disruption of two relevant processes for veridical memory recall: (i) lesions disrupt schema-guided memory activation processes and (ii) lesions impair monitoring processes that would reveal the correctness or falsehood of the contents of their recall (Burgess & Shallice, 1996; Ghosh et al.,

2014). As a consequence, lesions to the mPFC lead to an impoverished ability to remember past events and to simulate future events (Bertossi, Aleo, et al., 2016). This impairment is particularly striking when patients attempt to remember or imagine highly stereotypical events that require the activation of schematic knowledge (Kurczek et al., 2015).

Thus, lesion studies in humans suggest that our ability to remember past events is supported by both a medial temporal lobe system with the hippocampus as the central unit and the medial prefrontal cortex. Both regions subserve complementary functions that are crucial for intact mnemonic processing (McClelland et al., 1995): The mPFC has been associated with the representation of generalized semantic memory representations (Tse et al., 2011; van Kesteren et al., 2012; van Kesteren, Fernández, et al., 2010; van Kesteren, Rijpkema, et al., 2010) and the hippocampus has been linked to the representation of episodic memories (Frankland & Bontempi, 2005; Moscovitch et al., 2016; Rosenbaum et al., 2005). It has been argued that the mPFC provides the overarching scaffold that supports context adequate memory reactivation (Gilboa & Marlatte, 2017; McCormick et al., 2020; van Kesteren, Rijpkema, et al., 2010). In this regard, the mPFC might initiate and monitor hippocampal reactivation processes that are crucial for context adequate and veridical reactivation of specific episodic memories (Burgess & Shallice, 1996; Ghosh et al., 2014; Hebscher & Gilboa, 2016). The necessary bi-directional flow of information is realized via direct monosynaptic (Gabbott et al., 2005; see also Euston et al., 2012) as well as indirect pathways between hippocampus and mPFC (Eichenbaum, 2017; Vertes et al., 2007). Thus, memory processes are jointly concerted by close interactions of both regions.

1.3 Remembering and simulating: a common process

Early observations in patients with focal brain lesions provided first insights that remembering and simulating might be a common process that is supported by the same neural mechanisms. More evidence for this hypothesis comes from studies that investigated the developmental trajectories of both abilities over the course of our lives, studies that investigated the phenomenal characteristics of remembered and imagined scenarios, and functional neuroimaging studies that compared the activation of brain regions during both remembering and simulating.

1.3.1 Remembering and imagining follow similar developmental trajectories

Studies on the emergence of the ability to remember the past and imagine the future suggest that both abilities develop rather late between three and five years of age (Schacter et al., 2007). It has largely been established that episodic memories are not formed or can at least not be

consciously remembered before the age of three (Tustin & Hayne, 2010; see Hayne, 2004 for an extended discussion). In an elegant study, Scarf and colleagues (2013) probed the emergence of both abilities in samples of three and four year old children. Children were required to form a new memory of an event, retain this memory over a period of time, and then use this memory for a future directed choice. In this critical test, only children older than four years used the previously acquired knowledge to select the choice option that would serve a future purpose (see also Suddendorf & Busby, 2005).

More evidence for joint developmental trajectories of our ability to remember the past and imagine the future comes from studies that compare event descriptions provided by young and old adults. In these studies participants either described a remembered past or an imagined future event. The event descriptions were then analyzed using the autobiographical interview – a technique that allows for the quantification of internal episodic details and external semantic details (Levine et al., 2002). The measure of internal details provides an index for episodic memory contents (i.e., what happened, where it happened, etc.) and external details yield an index for semantic contents (i.e., facts, commentary, etc.). Both for descriptions of remembered past events and of imagined future events, old adults provided less internal and more external details as compared to their younger counterparts (Lyons et al., 2014; see also Addis et al., 2008; Schacter et al., 2013). Together these findings support the notion that the ability to remember past events and the ability to make use of these memories to imagine the future emerge at roughly the same age and are similarly affected by ageing.

1.3.2 Remembering and simulating share common phenomenal characteristics

The second line of research that provides evidence for commonalities between remembering and simulating has investigated the phenomenal characteristics of both functions. D'Argembeau and Van der Linden (2004) had participants mentally re-experience past events or pre-experience potential future events that were either temporally close or distant and either of positive or negative valence. Interestingly, past and future events that were closer in time evoked more vivid experiences than the temporally more distant counterparts (see also Trope & Liberman, 2003). Similarly, both imagined and remembered events evoked more vivid experiences when they were of positive as compared to negative valence. Moreover, participants who described themselves as well able to produce visual imagery also reported more visual and other sensory details regardless of whether they remembered a past event or imagined a future event (D'Argembeau & Van der Linden, 2006). Together these findings

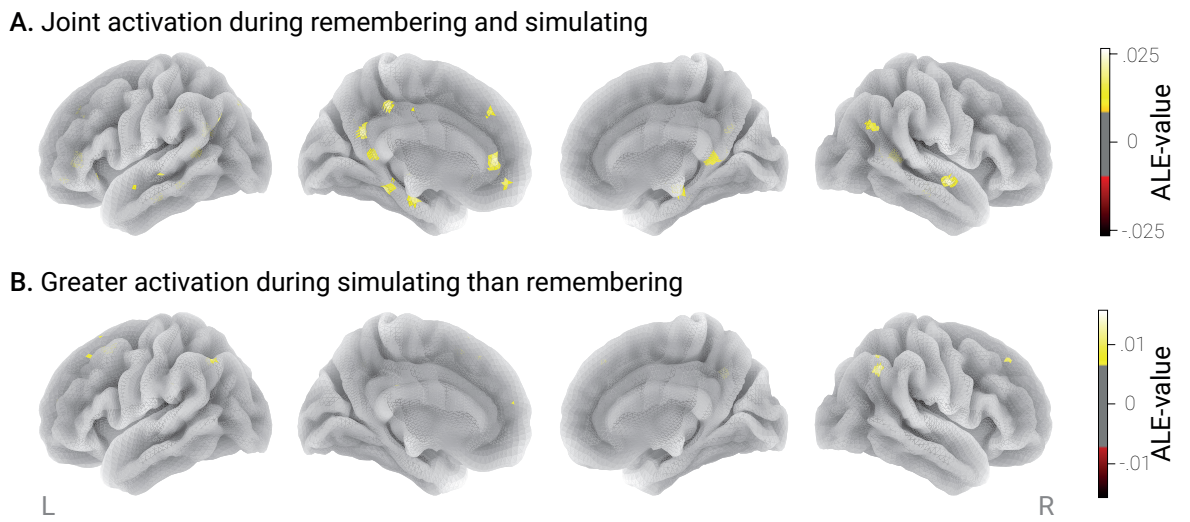


Figure 1. The core network of brain regions. A. Meta analytic map depicting regions in the brain that are jointly activated when individuals remember past events and simulate potential episodes in the feature. **B.** Meta analytic map depicting regions in the brain that are more strongly activated when individuals simulate an episode in the future as compared to when they remember past experiences. ALE = Activation Likelihood Estimate. Figure recreated from Benoit & Schacter (2015).

suggest that memories of the past and simulations of the future share joint phenomenal characteristics.

1.3.3 Remembering and imagining are supported by the same network of brain regions

More evidence for the close connection between our ability to remember the past and imagine the future comes from functional neuroimaging studies. The typical design of these studies requires participants to either remember specific past events, imagine specific events in the future, or perform some form of a control task (see Benoit & Schacter, 2015 for an overview). The results of these studies reveal striking overlap in the activation patterns when participants remember past events or simulate an event in the future (Benoit & Schacter, 2015; Schacter et al., 2007; Spreng et al., 2009; Stawarczyk & D'Argembeau, 2015). Together these studies suggest the existence of a core network for episodic memory and imagination (Buckner & Carroll, 2007; Hassabis, Kumaran, & Maguire, 2007) that largely overlaps with the default mode network that is more strongly activated when individuals lie in the scanner at rest (Buckner et al., 2008; Spreng et al., 2009).

More specifically, meta-analytic evidence indicates joint activation during episodic remembering and simulation (Benoit & Schacter, 2015). Evidence for joint activation was found in regions of the medial temporal lobes as well as the medial surface of the brain. Medial temporal regions included the parahippocampal cortex and the hippocampus. Regions on the medial surface comprised the mPFC, the posterior cingulate cortex (PCC), as well as the retrosplenial cortex. Moreover, regions in the lateral parietal and temporal cortex are jointly activated during remembering and simulation (Figure 1A). Regions that are more strongly

engaged whenever individuals simulate hypothetical events as compared to when they remember past events encompass the dorsomedial prefrontal cortex, the PCC and precuneus. There was also evidence for greater activation in parts of the medial temporal lobes including the hippocampus, the lateral temporal cortex, as well as the postcentral gyrus, and the cerebellum (Figure 1B; see also Benoit & Schacter, 2015) when participants imagine potential episodes (Benoit & Schacter, 2015; see also Spreng et al., 2009; Stawarczyk & D'Argembeau, 2015).

In sum, there is considerable evidence that remembering the past and imagining the future are abilities that follow common developmental trajectories, share many phenomenal characteristics, and are jointly supported by the same network of brain regions. Together, these findings provide evidence for the existence of a constructive memory system that supports both our ability to remember past events and to imagine the future.

1.4 Explaining the similarities between remembering and simulating

There are three main theoretical accounts that attempt to explain the striking similarities between our ability to remember past experience and imagine the future. Each of these theories highlights commonalities and differences between the two processes and puts particular emphasis on different neuroanatomical regions.

1.4.1 Scene construction theory

The *scene construction theory* suggests that the process that is common to episodic memory, episodic simulation, as well as spatial navigation is a scene construction mechanism that depends on the hippocampus (Hassabis, Kumaran, & Maguire, 2007; Hassabis & Maguire, 2009; Maguire & Mullally, 2013). The theory is reminiscent of Bartlett's initial observation that remembering is mainly a reconstructive process (Bartlett, 1932) that involves the re-activation of specific memory traces and subsequent enrichment processes that then induce the subjective experience of *remembering* (Tulving, 2002). Scene construction theory hypothesizes that the hippocampus provides a spatial context or scaffold into which these re-activated disparate event components are integrated (Mullally & Maguire, 2014). As such, scene construction theory views the similarity of remembering and simulating as a byproduct of their common dependency on scene construction. However, explaining the similarities between remembering and simulating is not really at the focus of scene construction theory. Instead, the theory mainly attempts to provide a functional account of the hippocampus (Mullally & Maguire, 2014).

Empirical support for the scene construction theory comes from patient studies as well as from functional neuroimaging studies. Hassabis, Kumaran, Vann, and Maguire (2007) had patients with focal hippocampal lesions and healthy controls describe specific imagined future events in response to cue words. Compared to event descriptions of healthy control participants, patients produced only loosely connected images that lacked the typical spatial coherence of an imagined scene. The authors reasoned that the hippocampus might add this spatial context to the imagined scenarios into which the disparate imagined event components such as objects, people, and actions can be integrated.

In a functional neuroimaging study, Hassabis, Kumaran, and Maguire (2007) had participants either remember a specific past event, remember a scenario they had imagined one week before, or simulate a specific event that might take place in their personal future. This design allowed them to differentiate regions that are commonly active during episodic memory retrieval and episodic simulation as well as regions that are selectively active during episodic recall. Moreover, they included a control task where participants were required to remember, imagine, or remember an imagined object. Using this design enabled them to identify regions that are activated in circumstances that require a scene construction mechanism. The results indicate that a core network of brain regions including the hippocampus, PCC and retrosplenial cortex, as well as regions in the dorsal mPFC were commonly activated in the episodic recall and imagination conditions. Real memories engaged the anterior mPFC, the PCC, and precuneus more strongly than remembering an imagined event. The results are thus in line with the argument that scene construction is a common mechanism that is required both for remembering as well as for simulating, but question the stipulated selective dependency of this mechanism on the hippocampus

More evidence for the scene construction theory comes from experiments that employ the boundary extension paradigm. During initial encoding trials, participants are presented with a visual scene. After a delay, participants are presented with the same image again and asked to rate whether this image is closer up or farther away. Participants tend to indicate that this second image is closer up than the original image they remembered having seen. This boundary extension effect may result from a scene construction process that extrapolates beyond the boundaries of the original physical stimulus (Maguire & Mullally, 2013). Critically, patients with hippocampal lesions exhibit strongly attenuated boundary extension and thus produce paradoxically far lower boundary extension errors than healthy controls (Mullally et al., 2012). Thus, research on the boundary extension paradigm indicates that the hippocampus supports an automatic scene construction mechanism. This mechanism might continually construct internal

representations that extrapolate beyond the boundaries our current field of view. These results are in line with the idea that the hippocampus is not only concerned with a representation of the past, but also supports our ability to anticipate the future.

Scene construction theory and the empirical work that surrounds it have greatly advanced our understanding of hippocampal contributions to episodic memory, episodic simulation, and spatial navigation. The empirical results support the notion that lesions to the hippocampus cause profound impairments to these functions. Thus, the theory receives empirical support for its claim that scene construction provides a unifying account of hippocampal function. However, the question how the hippocampus interacts with other brain regions in all of these processes and whether remembering and simulating are in fact a common process is beyond the scope of the theory.

1.4.2 Self-projection hypothesis

The *self-projection hypothesis* put forward by Buckner and Carroll (2007) attempts to explain the involvement of the same regions of the brain in episodic memory, episodic simulation, theory of mind – the ability to take another person's perspective – and spatial navigation. The theory conceives of *self-projection* as the ability to shift the perspective from the immediate present to alternative perspectives. Conceptually, this idea is related to Tulving's idea of mental time travel (Tulving, 2002) and extends it to cognitive functions beyond episodic memory. The key empirical observation that supports the self-projection hypothesis is the identification of an extended network of brain regions that is commonly engaged in all of these functions as well as when participants lie in the MRI scanner in the absence of task instructions. This network is also known as the default mode network and comprises of the mPFC, the PCC and retrosplenial cortex, the inferior parietal lobe, the lateral temporal cortex, the dorsal mPFC, and the hippocampal formation (Buckner et al., 2008). It thus largely overlaps with the core network of brain regions that is commonly activated when we remember the past and imagine the future (Benoit & Schacter, 2015).

But why would this default mode network be activated both at rest and in all of these different mental activities? Buckner and Carroll (2007) argue that individuals spontaneously engage in various forms of self-projection when they lie in the scanner in the absence of task instructions. Moreover, they argue that a common set of processes is necessary in all these functions: Individuals are required to imagine perspectives and events beyond those that emerge from the current environment and sensory inputs of the individual.

The ability to shift our perspective away from the present situation toward experiences we or others could have made in the past, might make in the present, or could make in the future is critical for both our ability to remember the past and imagine the future. The unique contribution of the self-projection hypothesis is the explicit inclusion of the simulation of other individuals' mental states and perspectives. Given the diversity of the cognitive functions that the self-projection hypothesis attempts to unify, it is no surprise that empirical work identified an extended network of brain regions that supports these functions.

1.4.3 The constructive episodic simulation hypothesis

The constructive episodic simulation hypothesis is rooted in the observation that remembering is rarely completely accurate. It tries to explain the many fallacies of our memories of the past by suggesting a highly adaptive memory system that uses memories of the past to construct simulations of potential future happenings (Bjork & Bjork, 1988; Schacter, 1999; Schacter et al., 2007). Much as the scene construction theory, the constructive episodic simulation hypothesis emphasizes the (re-)constructive nature of memory. It acknowledges that any form of remembering and simulating always requires the combination of both our semantic knowledge and memories of unique episodic details. The hypothesis is based on the descriptions of patients with focal brain lesions and their impaired ability to remember past experiences and imagine the future (Rosenbaum et al., 2005; Scoville & Milner, 2000; Squire, 2009). The key argument of the *constructive episodic simulation hypothesis* is that our memories of the past serve us in adaptive ways that are oriented toward the future.

In what is probably his most influential paper, Schacter (1999) describes the many ways in which our memories of the past are imperfect. Schacter describes typical fallacies of our memory systems that can roughly be summarized under the terms of forgetting, distortions, and intrusive recollections that are hard to forget. In his review paper, he argues that these imperfections are not malfunctions of our memory systems. Instead, he suggests that they are a byproduct of a system that needs to be highly flexible in order to adaptively support our ability to imagine a host of potential scenarios in the future. A memory system that would rely on a literal replay of memorized past experiences, akin to a video recorder, would be highly inefficient (see Bjork & Bjork, 1988; Schacter, 1999) and unfit to provide the building blocks required to simulate the multitude of imaginable future scenarios (Schacter et al., 2007; Schacter & Addis, 2007). According to this view, some degree of veridical memory recall is necessary to generate realistic simulations of potential future happenings. However, the theory suggests that the ability to flexibly recombine aspects of individual past experiences to construct

simulations of potential happenings is far more important for adaptive behavior than complete veridicality in the recall of past experiences.

Indeed, there is much reason to believe that episodic simulations provide an adaptive account of our ability to remember the past: Our ability to simulate the future supports farsighted decisions and flexible behavior. Delay discounting provides one example for future-oriented decision making where we need to reject an offered immediate small reward in order to receive a larger reward later on. A typical delay discounting paradigm would require the individual to either select a small immediate reward (e.g., 1€ now) or endure a prolonged period of waiting to receive a larger delayed reward (e.g., 10€ in 10 days). Individuals tend to discount the subjective value of each choice option as a function of the temporal delay they have to endure until they receive the reward. As a consequence, individuals are prone to reject the delayed offer and myopically choose the smaller immediate offer instead (Ainslie, 1975; Bulley & Schacter, 2020; Green & Myerson, 2004). However, when participants are asked to mentally simulate an episode in which they receive or consume this larger reward at the distant point in the future, the tendency for myopic decisions is attenuated (Benoit et al., 2011; Peters & Büchel, 2010a). In this context, episodic simulations of the future might have allowed the individuals to experience the emotional quality of actually receiving the reward ("*what it would feel like*") already in the present. This emotional experience, in turn, might have provided a motivational cue that rendered their decisions more farsighted (Boyer, 2008). Critically, patients suffering from Alzheimer's disease or hippocampal damage who are impaired in their ability to simulate future events, do not show the attenuation in delay discounting typically observed in healthy control participants (Lebreton et al., 2013; Palombo et al., 2015; but see Kwan et al., 2015).

Episodic memory and episodic simulation jointly support our ability for adaptive behavior. The ability for prospective memory requires the individual to encode, store, and carry out initially formulated intentions after a delay in the future (Kliegel et al., 2007). Several studies have shown that prospective memory can be improved when individuals mentally simulate actually carrying out that action in the future (Altgassen et al., 2015; Brewer & Marsh, 2010; Neroni et al., 2014). In this context, our ability to anticipate the future can help us to identify the appropriate circumstances to remember and execute our initially formulated plans. Thus, remembering and simulating closely interact to enable us overcome anticipated future impediments that would otherwise stop us from executing our plans (Schacter et al., 2017; see also Gollwitzer, 1999).

In sum, the *constructive episodic simulation hypothesis* explains the similarities between remembering the past and imagining the future by suggesting that our memories of the

past primarily serve us in ways that oriented toward the future. Thus, we can use our knowledge about past experiences and simulate the possible outcomes of a plethora of different events and their likely consequences. Our ability to remember these simulated scenarios and their outcomes can help us render our decisions more farsighted and allow us to overcome future impediments that would otherwise stop us from executing our plans. Thus, according to the constructive episodic simulation hypothesis our ability to remember past experiences and simulate the future are closely intertwined because they serve the same ultimate goal: Supporting adaptive and flexible behavior.

1.4.4 Summary of the theoretical accounts

In sum, there are three broad theoretical accounts that attempt to explain the commonalities between remembering and simulating both on a cognitive as well as on a neuroscientific level. The *scene construction theory* argues that remembering, simulating the future, and spatial navigation share a common necessity to evoke a mental representation of a spatial scaffold. This scaffold is used to bind the disparate remembered or simulated event components into a coherent scene. According to the theory, this scene construction mechanism is supported by the hippocampus. Empirical findings generally support the importance of the hippocampus for both our ability to remember the past and imagine the future. However, the theory falls short to account for the involvement of other brain regions in the same or similar cognitive processes. More generally, the theory is agnostic with regard to prefrontal contributions to these functions and may thus rather be regarded as a theory of hippocampal functioning.

The *self-projection hypothesis* has a stronger focus on the entire network of brain regions that are commonly activated by remembering, simulating, theory of mind, and spatial navigation. The hypothesis suggests that this network is commonly activated whenever we transcend from our immediate environment to conceive of other perspectives or events at different points in time and space. While this hypothesis allows for the accommodation of many empirical findings, it provides only few testable predictions and offers little to explain the processes involved in each of the related cognitive functions.

The *constructive episodic simulation* hypothesis is a broad framework that argues that our ability to remember past experiences serves us in ways that are oriented toward the future. Thus, we can use our memories of past experiences and flexibly recombine them to simulate a plethora of potential episodes as well as their likely consequences. This provides a great adaptive device: It allows us to anticipate what it would feel like if these simulated events actually happened and thus motivate farsighted decision making. Empirical work that surrounds

this hypothesis has highlighted the contributions of an entire network of brain regions to both abilities. The main strength of this hypothesis is that it does not focus on one specific brain region or argue for a single mechanism that accounts for the similarities between remembering and simulating. Instead, the hypothesis argues that both functions support our ability for adaptive and flexible behavior. Notwithstanding the differences between the three accounts, they converge on the idea that our memories of the past and the neural systems that encode them are highly flexible and support a wide range of adaptive human abilities.

1.5 The medial prefrontal cortex: Memory and value-based decision making

The previous sections have reviewed evidence that a core network of brain regions including the mPFC, the hippocampus, und the PCC jointly supports our ability to remember past experiences and imagine the future. One key argument of the reviewed literature is that these mnemonic functions support behavioral flexibility in a wide range of contexts. However, such adaptive behavior requires more than a representation of what might have happened in the past or could happen in the future: It also requires knowledge about the consequences of our choices (O'Doherty et al., 2017).

Consider the following example where an individual is required to choose between a small and a large amount of food. In this situation, the individual would estimate the value of the two choice alternatives and then decide for the option that is overall most valuable. However, this task is substantially more difficult when the value of choice options is not readily accessible: Would you rather work less and accept a lower salary or put in extra hours to get that promotion? Would you rather want two apples or one liter of milk? In order to make such decisions between incommensurable choice options, we require an internal representation of value that allows for evaluations and comparisons of even the most disparate elements. To achieve such a representation we require a brain mechanism that maps all choice alternatives onto a common scale and computes a domain general subjective value signal for decision making (Bartra et al., 2013; Peters & Büchel, 2010b).

Typical tasks that investigate this neural mechanism require the individual to select one out of several choice alternatives. For example, Plassmann et al. (2007) had hungry participants place real money bids in the fMRI scanner to be allowed to consume food items. The amount of money they were willing to commit served as a proxy measure for participants' subjective value of the food items. The results revealed that activation in the mPFC scaled with the magnitude of the placed bid. This finding supports the notion that the mPFC encodes the value of different choice alternatives in decision making. In a related study, Talmi et al. (2009) had participants choose between two different stimuli that would yield a high or a low probability

of a monetary reward. However, these choice options were simultaneously also associated with a high or low risk of receiving a painful shock. Activation in the mPFC covaried with the linear combination of both the value of the monetary reward and the perceived cost of the painful stimulus. Thus, the mPFC simultaneously tracked both gains and losses (see also Bartra et al., 2013).

The value signal in the mPFC is highly flexible. In a recent study, Castegnetti, Zurita, and De Martino (2021) devised a novel decision-making task where participants had to perform choices in the fMRI scanner. In this task, the monetary values of choice options (e.g., a wooden vs. a metal chair) were decoupled from their specific utility to solve a task at hand (i.e., light a fire vs. anchor a boat). The results of this study revealed both a stable representation of the objects' monetary value as well as a context-dependent value signal in the mPFC that reflected the utility of the object for the current goal. Together these findings suggest that activation in the mPFC scales with the subjective value of different choice alternatives and may thus be regarded the central brain region for the representation of subjective value (Bartra et al., 2013; Chib et al., 2009; Levy & Glimcher, 2012).

Critically, the mPFC also codes for value in situations where individuals are not explicitly required to make value judgements for decision making. Activation in the mPFC is greater during the imagination of positive scenarios as compared to negative scenarios (Benoit et al., 2014; D'Argembeau et al., 2008; Sharot et al., 2007). Regardless of whether participants rate the pleasantness or age of faces, paintings, or houses, activation in the mPFC is predictive of subsequently assessed preferences (Lebreton et al., 2009). Moreover, mPFC activation reflects the anticipated reward of an imagined scenario (Benoit et al., 2011) potentially by conveying the reward values of individual event components of the simulated scenario (Boyer, 2008).

The mPFC is not the only brain region associated with both mnemonic functions and the representation of value. The hippocampus and PCC have similarly been associated with the representation of value and value-based decision making. Neurons in rodent hippocampus have been shown to encode reward amount and delay of the reward in a delay-discounting paradigm (Masuda et al., 2020). Single cells in primate hippocampus combine information about physical locations of rewards and reward magnitude in a map of an abstract value space (Knudsen & Wallis, 2021; see also Landi & Buffalo, 2022). Human hippocampus has been demonstrated to support deliberation in value-based decision making. It has been shown that hippocampal lesions impair value-based decision making, probably by impairing deliberation processes during the decision phase (Bakkour et al., 2019). Individuals with hippocampal lesions are also

impaired in the construction and recall of preferences that are critical for value-based choices (Enkavi et al., 2017). Moreover, the PCC has been shown to code for the time-discounted reward value in delay discounting paradigms (Kable & Glimcher, 2007; Peters & Büchel, 2009).

In sum, the brain system that is involved in valuation and value based decision making at least partially overlaps with the core network of brain regions involved in episodic memory and episodic simulation (Bartra et al., 2013; Benoit & Schacter, 2015; Clithero & Rangel, 2014; Lebreton et al., 2009). Regions that are both associated with the representation of value and memory encompass the mPFC, the PCC, and the hippocampus. The apparent overlap between mnemonic and evaluative functions in the brain is well in line with the claim that our memories of the past serve us in ways that are oriented toward the future: To select the best possible course of action and flexibly adjust to the requirements of our environment, we require knowledge about past experiences. We can then use this knowledge to generalize from these individual experiences to anticipate what the future might be like. By simulating the future, we may achieve an intuition what the affective consequences of our behavior might be. To put it differently, flexible and adaptive behavior requires close interaction between mnemonic and evaluative processes.

1.6 Scope of the thesis and study overview

This thesis examines the role of the mPFC for our ability for episodic simulation. More specifically, the thesis attempts to better characterize the dual involvement of the mPFC in both mnemonic processes and the representation of value and affect.

Chapter 2 will provide an overview of the key methods employed in this thesis. The chapter will provide a broad overview of fMRI acquisition, preprocessing, and statistical analysis methods. The section will then highlight some of the employed statistical methods and describe the experimental tasks in detail.

Can we learn from simulated experiences much as we learn from actual past experiences? Chapters 3 and 4 provide insights into a learning mechanism that is based on episodic simulations. In both studies, participants provided lists of personally known people and locations. Neutral locations were then paired with liked and disliked people and participants simulated vivid episodes of interacting with the people at the locations.

Across an fMRI study and a pre-registered replication study, chapter 3 reports evidence that simulating episodes with liked people increased the liking of the paired initially neutral locations more strongly than simulating episodes with disliked people. Activation in the mPFC predicted the magnitude of this simulation-induced transfer of valence from the person to the

location. Moreover, results of a representational similarity analysis provide evidence that the mPFC encodes representations that are unique to the individual people and locations. These results might be taken to suggest that the mPFC encodes schematic representations of our environment that also entail a representation of the value of the encoded exemplars.

Chapter 4 reports results from a study that investigated the mechanisms that support the transfer of valence from the person to the location more closely. To this end, we extended the previous design in a number of ways. The results indicate that simulations cause an integration of the jointly simulated exemplars' memory representations. This merging of memory representations might then allow for a transfer of valence between the person and the location. Together the results reported in chapter 3 and 4 provide evidence that we can learn from simulated experiences much as we learn from actual past experience.

Chapter 5 reports results of a study that examined the structure of memory representations in the mPFC more closely. Specifically, the study tested whether the structure of participants' memory representations of personally known people and locations can be predicted by (i) the degree to which these people are central to participants' everyday environment, (ii) how well they know them, and, critically, (iii) how much they like them. Using representational similarity analysis, we demonstrate that representations in the mPFC can best be predicted from the principal component of (i) centrality, (ii) familiarity, and (iii) liking of the known people and locations. This was not the case in the hippocampus and PCC – two other brain regions that have been associated with both mnemonic functions and value. These results provide evidence that the mPFC encodes schematic representations within which representations of individual exemplars are closely intertwined with a representation of their value.

In the final chapter of the thesis, I summarize the results of the three studies and discuss them in the wider context of the literature on episodic simulation and valuation. Moreover, I discuss limiting factors of the presented studies and provide an outlook on research that may follow up on the presented results.

Chapter 2

2 Methods

2.1 Functional magnetic resonance imaging

2.1.1 Signal generation and acquisition

FMRI is a non-invasive neuroimaging technique and a special form of MRI. Both techniques rely on the use of a strong magnetic field that aligns the orientation of atomic nuclei (particularly hydrogen protons) with the field lines of the magnetic field. Radio frequency coils of the scanner emit a radio frequency pulse that is absorbed by the atomic nuclei. This induction of energy by the nuclei perturbs the initially achieved equilibrium and forces the atomic moments out of alignment with the magnetic field of the scanner. When the radio frequency pulse ends and, thus, the energy source is removed, the atomic nuclei return to the baseline state and release the previously acquired energy as an electromagnetic pulse. To spatially tag the electromagnetic pulses and thereby allow for a spatial localization of the source, the gradient coils of the scanner modulate the strength of the magnetic field along the X, Y, and Z axis. A receiver coil that is placed around the body part under investigation is then used to record the emitted electromagnetic pulses, i.e., the MR signal. Due to the induced modulation along the X, Y, and Z axis the electromagnetic pulses can be recorded as a function of their spatial origin. Depending on the pulse sequence used, different tissue types emit MR signals of varying intensities allowing for a visualization of different tissue types (see Huettel et al., 2008 for a comprehensive overview).

As a measure of functional activity, fMRI is based on the same principle and makes use of the following phenomenon: The magnetic properties of blood vary as a function of the oxygenation level. While oxygenated hemoglobin is diamagnetic (i.e., it does not have a magnetic moment), deoxygenated blood is paramagnetic (i.e., it has a magnetic moment) (Pauling & Coryell, 1936). Thus, whenever neuronal assemblies process information, their increased consumption of energy in the form of glucose and oxygen leads to changes in local concentrations of deoxygenated blood. Paramagnetic substances distort the surrounding magnetic field and thereby also alter the MR signal leading to a local decrease in signal intensity (Thulborn, 2012; Thulborn et al., 1982). Coupling processes between neural assemblies and the surrounding blood vessels – also referred to as neurovascular coupling – cause an increased flow of oxygen rich blood to brain regions with increased metabolic demands providing a relative surplus of oxygen rich blood (Logothetis et al., 2001; Poldrack et al., 2011). This increase in local blood flow effectively flushes the deoxygenated blood from the regions with high metabolic demands thereby yielding an overall increase in MR signal intensity. These observations provide the basis for the blood oxygenation level dependent (BOLD) contrast (Ogawa et al., 1990). However, this neurovascular coupling causes a dependency of the BOLD

signal on local blood flow. This is one of the main reasons why the BOLD signal evolves rather slowly over time with evoked responses peaking approximately six seconds after onset of the stimulation. Typically, it is thus sufficient to measure fMRI at rather low sampling rates of 0.3 to 0.5 Hz, providing one image of the entire volume of interest every 2 to 3 s. The time required to acquire one functional image is also referred to as the repetition time (TR).

In sum, the BOLD signal is an indirect measure of neural activity. BOLD reflects changes in oxygenation levels that are linearly related to the local field potential in a given region of the brain (Logothetis et al., 2001). The signal has a low *temporal* but high *spatial* resolution, providing information about *where* in the brain information is processed.

2.1.2 Preprocessing of the raw images

Before differences in functional activity between experimental conditions may be estimated, the acquired time-series of raw images need to be preprocessed. Preprocessing typically comprises several processing steps that either try to account for potential artifacts caused by the scanner (e.g., spatial unwarping), by the individual being scanned (e.g., spatial realignment) or that prepare the images for later statistical analysis (e.g., spatial smoothing). These processing steps usually comprise slice-timing correction, spatial realignment, spatial unwarping, spatial coregistration, spatial normalization, and spatial smoothing. As preprocessing pipelines vary drastically between different software packages, I will discuss the processing steps as implemented in the software package Statistical Parametric Mapping (SPM; Penny et al., 2011) that is used in all fMRI studies of this thesis.

Correction for slice acquisition times is done to account for the fact that the individual slices that make up one brain scan are not all acquired at the same time. Instead, images are acquired over a time-course of typically two to three seconds (repetition time, TR). However, the assumptions of the statistical model require that all datapoints must be sampled at the same point in time (Henson et al., 1999). To account for differences in slice acquisition times, a reference slice (typically the slice that is acquired at TR/2) is selected and the data of all other slices are interpolated linearly to ensure that all data points refer to the same point in time.

To account for slight head movements that typically occur over the course of a scanning session, the acquired time series is realigned to either the first, a representative image, or the mean image of the time series. Linear transformations applied to each image ensure that all volumes are in alignment with each other (Friston et al., 1996). The translation and rotation parameters used for this linear transformation are saved and later included in the statistical analysis of the images (Friston, Frith, et al., 1995).

The process of spatial unwarping requires the acquisition of field maps. These field maps provide estimates of the homogeneity of the magnetic field and thereby allow for a post-hoc adjustment of the images to counter-act artifacts induced by local inhomogeneities of the magnetic field (Jezzard & Balaban, 1995). Inhomogeneities of the magnetic field disproportionally affect those parts of the brain where brain tissue borders with air, such as at the sinuses and ear canals. Unwarping thus allows for the improvement of signal quality particularly in these parts the brain.

Spatial coregistration ensures that scans of different modalities (i.e., the functional and anatomical scans) are in alignment with each other. To achieve alignment of the functional scans with the anatomical scan, the anatomical image is transformed using linear transformations.

Spatial normalization accounts for individual differences in brain morphometry. Given that heads and brains differ in shape and size between participants, spatial normalization determines a non-linear transformation matrix that allows for the projection of the data of individual participants from their *native* space into a *common space* (Ashburner, 2007). In this *common space* coordinates and therefore time-series of the single voxels correspond to the same parts of the brains of all participants. The transformation matrix required to transform an image from *native* into *common space* is estimated from the anatomical image and can be applied in both directions. Thus, the transformation matrix enables projection of the data of individual participants into *common space* or projecting masks of anatomical regions taken from anatomical atlases from this *common space* into each participants' *native* space. One main advantage of performing analyses of functional data in *native space* is that only minimal (i.e., linear) transformations are required when preprocessing the raw data. This reduces the degree to which raw data must be averaged and interpolated, retaining more of the original information in the time series of images.

In a final processing step that may be omitted for multivariate analyses, a spatial filter is applied to the data that causes a smoothing of the acquired images. Smoothing removes high frequency components in the signal and thereby increases the signal-to-noise ratio using a Gaussian smoothing kernel. Moreover, when statistical maps of different participants are compared in *common space*, smoothing also reduces effects of spatial variability and potential artifacts induced by the normalization procedure between participants (see Friston, Holmes, et al., 1995; Worsley & Friston, 1995). However, smoothing also blurs information that is carried in the fine-grained activation patterns preventing the researcher from discriminating between experimental conditions usually at the focus of multivariate analysis techniques (Lewis-

Peacock & Norman, 2014). In univariate analyses, where the researcher mainly attempts to identify the approximate peak coordinate and amplitude of a neural response, smoothing improves signal to noise ratio and reduces between subject variability. For multivariate analyses that are conducted within participants the application of spatial smoothing is a matter of debate and may be omitted (Dimsdale-Zucker & Ranganath, 2018; but see also Hendriks et al., 2017). Here, we conduct our multivariate analyses on the unsmoothed images in participants' *native* space.

2.1.3 Modeling and statistical analysis

To extract parameter estimates that quantify the magnitude of the BOLD response to our experimental manipulations in individual voxels, we make use of the General Linear Model (GLM). This requires us to first formulate a prediction of what the time-series should look like if a voxel responded to our experimental manipulation. In a next step, we can then quantify how well this prediction fits to the actual data.

To predict the assumed response to a given stimulus we first need to make a number of simplifying assumptions: To this end we regard the BOLD response as the output of a linear time invariant (LTI) system (Poldrack et al., 2011). This simplification of biophysical reality assumes that the response of the system to two identical stimuli is always the same (the output depends on the input only) and that an increase in neural activation by a factor of *n* also leads to an increase in the BOLD response by the same factor *n*. The model also assumes that the overlapping responses to two stimuli presented closely in time are the sum of the responses to the individual inputs (linearity assumption). The model further assumes that if a stimulus is moved by a factor *t* in time then the response is also moved by the same factor *t* (time invariance assumption) (Poldrack et al., 2011; see also Bach & Friston, 2013). An LTI is unambiguously defined by its response function and – given some boundary conditions – the LTI properties can be regarded as met for fMRI. Note that hemodynamic responses are not the only biophysical signal that can be regarded as output of an LTI system. Other physiological variables such as the skin conductance or heart-period response can also be described as outputs of LTI systems (Bach et al., 2009, 2018; Bach & Friston, 2013; Paulus et al., 2016).

In the example of fMRI, the hemodynamic response function (HRF) specifies the prototypical shape of the BOLD response as it evolves over time. We can thus predict a time-series by combining the onsets and durations of our events of interest with the HRF thereby achieving a predicted time series for our conditions of interest (Cohen, 1997; Friston et al., 1994). This predicted time series contains an expected activation level for each time-point in

our experiment. The predicted time series for all conditions of interest are then arranged in columns of the design matrix. Columns that code for conditions of interest are then combined with movement parameters (translation and rotation in X, Y, and Z direction) and intercept columns are added as nuisance regressors. Inverting this design matrix with the data yields beta estimates that describe the fit of our prediction to each voxel of our data. In typical univariate analyses, we would then contrast the beta maps of all participants in one condition of interest with other conditions using, e.g., a paired *t*-test to determine where in the brain univariate activation levels for one condition are significantly larger than in another.

2.1.4 Representational similarity analysis

More recently developed analysis techniques aim at estimating the information contained in the multivariate or multi-voxel activation pattern instead of only estimating univariate activation levels. In brief, these techniques, that are also referred to as Multi-Voxel Pattern Analysis (MVPA), attempt to differentiate experimental conditions on the basis of a spatially distributed activation pattern. One core feature that differentiates MVPA from univariate analysis techniques is that even if a voxel in itself is not significantly activated by a given experimental condition or stimulus, it might nonetheless contribute to the combinatorial multi-voxel code that allows for the separation of experimental conditions of interest (see Norman et al., 2006).

MVPA approaches can roughly be divided into two groups: Linear classifiers and Representational Similarity Analysis (RSA). Linear classifiers try to estimate the cognitive state of a participant based on the activation pattern in multiple voxels. Linear classifiers rely on separate datasets used for training and evaluation of the classifier and will not be discussed here (but see Lewis-Peacock & Norman, 2014; Norman et al., 2006 for an overview). The second subtype is RSA (Kriegeskorte et al., 2008) and it is one of the key methods used in the present thesis. For the purpose of RSA, activation patterns are treated as coordinates in high-dimensional voxel space. Each voxel is regarded as one dimension in this space and the activation level describes the coordinate on that dimension. The distance between any two activation states within this high-dimensional space is treated as a measure of neural pattern similarity. These distances between activation states are summarized in a distance matrix that is also referred to as the representational dissimilarity matrix (RDM). Typically, the correlation distance (1 – Pearson correlation of any two patterns) serves as a distance measure that ranges from 0 (maximum similarity, identical activation pattern) to 2 (maximum dissimilarity, inverse activation pattern) (Kriegeskorte et al., 2008). The similarity structure or representational geometry, that is captured within this matrix is then taken to describe the features or dimensions

by which neural activation patterns that are elicited by our experimental conditions can be differentiated.

Once this distance matrix is computed, the observed *neural* RDM can be compared with some other *model* RDM. This other RDM could for example contain the predicted similarity structure derived from some cognitive variable (e.g., the similarity structure in visual cortex should reflect the similarity of the illuminance of displayed visual stimuli) or computational model (e.g., the similarities of outputs of a layer of a deep neural network). One key advantage of this approach is that it allows for the construction of RDMs from any type of data. Comparisons between RDMs are typically done by computing the correlation between the neural and model RDM. This approach even allows for a joint examination of RDMs derived from different neuroimaging techniques (Cichy et al., 2016) or even of RDMs obtained from the observation of different species (Kriegeskorte, 2009).

RSA can be conducted using a region of interest approach or using an unconstrained whole-brain searchlight approach. For this second kind of analysis, a spherical searchlight is moved through all grey matter voxels of a participant's brain. The searchlight is centered on each individual voxel, then an RDM is constructed for this searchlight and a comparison to a given model RDM is done. The correlation of the searchlight RDM and the model RDM is then assigned to the center voxel's position in a searchlight map. Evaluation of these searchlight maps across all participants then provides a whole-brain map that quantifies where in the brain information – as described in the *model* RDM – is encoded (Kriegeskorte et al., 2006; Nili et al., 2014).

RSA can also be used to test hypotheses about the representations of individual exemplars. Initially, multivariate analysis techniques for neuroimaging data were developed to probe hypotheses about the representation of categorical information (e.g., differentiate between brain representations of scenes vs. objects). However, such multivariate analyses can also be conducted within categories on the more subtle differences between individual exemplars of the same category (Chan et al., 2010; Kay et al., 2008; Kravitz et al., 2010). To this end, at least two repeated assessments of an exemplar are necessary to examine the replicability of activation patterns (Nili et al., 2020). The exemplar discriminability index compares the similarity of two or more patterns elicited in trials that pertain to the same exemplar (*same item similarity*) and compares it with the similarity of patterns elicited in trials that pertain to different exemplars of the same category (*different item similarity*). Regions that encode stable representations of unique exemplars are characterized by greater *same item similarity* as compared to *different item similarity* (Nili et al., 2020). This technique can be

combined with a searchlight thereby yielding information about where in the brain unique representations of individual exemplars are encoded.

2.2 Statistical modeling

2.2.1 Linear Mixed-Effects Models

Linear models are used to describe the degree to which two variables of interest are interrelated. Often, however, relationships between variables of interest are based on multiple observations of several variables in different individuals. Given that observations made from the same individual are often correlated with one another, a central assumption of many statistical tests is violated: The independence assumption (Brown, 2020). Thus, as soon as repeated observations of the same participant need to be taken into account repeated measures analyses of variance (ANOVAs) are preferable to standard linear models or ANOVAs. However, ANOVAs require the computation of grand means per condition and are somewhat inflexible when it comes to the handling of missing data (Brown, 2020).

To overcome the limitations of multiple regression and repeated measures ANOVAs the use of linear mixed-effects models (LMEMs; Baayen, 2008; Singmann & Kellen, 2019) is recommended. LMEMs provide a flexible framework within which both fixed and random effects can be explicitly modeled and accounted for simultaneously. Moreover, LMEMs handle missing data and unbalanced designs well (Brown, 2020). The term "mixed" refers to the simultaneous inclusion of both *fixed* and *random* effects in the statistical model. *Fixed effects* are constant across participants and levels of observation and are the conceptual counterpart to the correlation coefficient or beta weights in standard linear models. *Random effects* describe influences that are variable across individuals or levels of observation. Due to the inclusion of both effect types, statistical analyses using LMEMs can directly be conducted on individual observations. The inclusion of random effects ensures that statistical dependencies between individual observations are adequately accounted for (Barr et al., 2013; Singmann & Kellen, 2019). In study 2 and 3 we use LMEMs as implemented in *lme4* (Bates et al., 2015) in *R* (R Core Team, 2016).

2.2.2 Model selection

Psychological and cognitive neuroscience – just like any other empirical research area – are concerned with the collection, evaluation, and interpretation of data. One of the most central aspects in this regard is the selection of an adequate statistical model to describe the data. Ad-hoc methods that are commonly used to identify this best possible model are based on adding

and removing model terms and the application of sequential statistical tests. Unfortunately, these techniques can lead to biased model parameters or prevent the subsequent interpretation of the statistical significance of individual model terms. In these cases, the researcher would be required to collect additional data for an unbiased interpretation of model parameters. Fortunately, methods concerned with model selection formalize this process and enable the researcher to identify the most probable model from a population of alternative statistical models given the data (Burnham et al., 2002). The parameters used for this selection are independent of the statistical significance of individual model terms and are typically based on the evaluation of the fit of the model onto the data (Cherkassky & Ma, 2003; Kass & Raftery, 1995). Typical parameters that are used for model selection are Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) (Burnham, 2004; Burnham et al., 2002). Both parameters contain penalty terms for model complexity that ensure that the simpler of two equally suited models is always preferred over a more complex model (Cherkassky & Ma, 2003). Similar to classical hypothesis testing, certain community standards have been established that define when a model should be regarded as substantially better than another model to account for the observed data. In the present thesis, we adhere to the conventions put forward by Kass and Raftery (1995).

After identifying the most probable model from a population of candidate models, we can then evaluate individual model parameters and determine their statistical significance. This two-step approach ensures that we select the best possible model, while simultaneously maintaining interpretability of the subsequent statistical tests. Model selection is used in study 3 to identify the likely best model to account the structure of neural representations in different regions of interest.

2.2.3 Resampling methods: Permutation test

Parametric statistical tests are based on the comparison on the observation of statistical parameters in an empirical sample of data with some theoretically defined statistical density function (e.g., the Gaussian probability density function). In contrast, resampling methods directly estimate a null distribution by repeatedly shuffling the condition labels of the observed data and computing the statistic of interest for each resample (e.g., a mean correlation). This process yields a null distribution of the parameter of interest. The observed true statistical parameter is then compared with the empirical null distribution and statistical significance is determined. As in classical inferential statistic, the parameter of interest is regarded as statistically significant different from zero if it falls outside the 95% interval around the central

moment of the null distribution (Collingridge, 2013). One disadvantage of resampling methods is their computational inefficiency as they rely on iterative permutations of the data that typically restricts a full permutation of all possible combinations. Thus, the researcher is required to select a number of practically feasible permutations of the data. This number is typically set at values > 1,000 allowing the computation of *p*-values up to a significance level of $p < .001$ (Davidson & MacKinnon, 2000; see also Kriegeskorte et al., 2008). In brief, resampling methods are powerful alternatives to parametric tests and are used in study 3.

2.3 Experimental paradigm

In the three empirical projects, I employed different versions of an episodic simulation task alongside fine-grained assessments that quantified different aspects of the employed stimulus material. In this chapter, I will briefly describe the episodic simulation task that we used across the three studies and describe the arrangement tasks.

2.3.1 Episodic simulation task

Across the empirical projects, I employed an episodic task that uses self-generated stimulus material from participants' real life. This ensures that the knowledge that the tasks operate on is not artificially generated in simplified experimental paradigms in the lab, but really reflect participants' knowledge about their environment. To this end, the episodic simulation task always comprises a preparation session, during which participants generate the lists of known people and locations, as well as a simulation session, during which participants simulate vivid episodes of interacting with the people or being at the locations.

The preparation session is typically done on the day before the simulation session. To ensure that the self-generated exemplars (i.e., the people and locations) are sufficiently different from each other and to achieve sufficient variance in all variables under investigation, participants are instructed to come up with large numbers of 90 – 200 exemplars (see Addis et al., 2009; Benoit et al., 2014; Szpunar et al., 2012). Participants are allowed to use their contact lists (e.g., from their phone or the social media platform Facebook) to ease the process of stimulus generation. Participants are required to name only one location per building (e.g., when a participant names the elevator of the Max Planck institute, they are not allowed to also name the cafeteria) or open space (e.g., when a participant names the Musikpavillon they are not allowed to name the Sachsenbrücke). This ensures that the locations are sufficiently different from each other. Participants are allowed to use google maps to ease the process of listing the locations.

Participants return to the lab for the simulation session. Each trial of the simulation task commences with the presentation of a fixation cross (0.5 s) that is then followed with the main simulation trial (7.5 s). During this time, the name of a person and / or a location is presented on the screen. Participants are either asked to vividly imagine interacting with the person in a way that would be typical (e.g., meeting your friend hearing them talk about a recent experience they had and seeing typical gestures) or being at the location, engaging in a location specific activity (e.g., at a restaurant browsing the menu). In studies 1 and 2, participants were also required to imagine meeting the person at the location and engaging in a location specific activity (e.g., meeting your friend at that new Italian restaurant, discussing to share starters). In that case, the name of both the person and the paired location are presented on the screen simultaneously. After the simulation-trial proper, participants rate the vividness of the simulated episode (≤ 2 s) on a five-point scale (1: not at all vivid – 5: very vivid). The vividness rating is followed by a flexibly jittered inter-trial interval (≤ 2 s), during which the screen is blank.

Before engaging in the main simulation task, participants receive training on a number of training trials. To this end, we use people and locations that are not part of the main experimental task. Upon completing these training trials, the experimenter conducts a structured interview with the participants to determine whether participants simulated a specific episode that focused only on the presented person and / or location. Moreover, this interview ensures that participants really simulate a specific episodic scenario that might actually happen. Following this training, participants engage in the main simulation task, that comprises of two or more blocks of the episodic simulation task. In the MRI experiments, this part of the study is conducted within the fMRI scanner.

2.3.2 Quantifying the self-generated material: Arrangement tasks and ratings

Across the studies, we use different versions of a modified multiple arrangements task (Goldstone, 1994; Kriegeskorte & Mur, 2012). To estimate how people and locations are associated with one another, we use a two-dimensional arrangement task that is directly based on the original multiple arrangements task and the inverse multidimensional scaling technique described in Kriegeskorte and Mur (2012). Moreover, we quantify the individual expression of different variables that can best be mapped onto continuous scales, such as how familiar the individual people and locations are, or how much our participants like them. To this end, we employ one-dimensional arrangement tasks. This section will first describe in detail the

function and logic of the two-dimensional arrangement task before describing the one-dimensional arrangement tasks in detail.

The square RDM is at the core of representational similarity analysis (Figure 2A). For a set of items, this matrix describes the pairwise dissimilarities between all items. A technique that allows for the visualization of the RDM is multidimensional scaling (MDS). The method provides a projection of the distance matrix onto a set number of specified dimensions thereby estimating a coordinate for each of the individual items. Conceptually, the MDS estimates a projection of the data that minimizes the disparities between the true RDM and an RDM that is recovered from this lower-dimensional projection of the data. The coordinates that are estimated by MDS can be used to visualize the relationships embedded in the RDM in lower dimensional space (Figure 2B).

Overt estimation of representational similarities from participants' behavior. Often representational similarities can directly be estimated from the stimulus material: In an experiment that employs visual stimuli, we can quantify a given feature of the stimuli (e.g., the spatial frequency) and construct a dissimilarity matrix that describes similarities in that stimulus domain. However, sometimes the features of interest are not directly accessible. In these cases, representational similarities need to be estimated from participants' behavior. Theoretically, one might just present a participant with each individual pair of exemplars and ask the participant to estimate how similar the exemplars are with regard to that given feature. However, this approach is highly inefficient as the number of comparisons increases as a quadratic function of the number of exemplars that need to be assessed (i.e., while only 45 estimations are needed to rate all pairwise similarities of ten items, 30 items require 435 comparisons). An efficient and less time consuming alternative is provided by the inverse multi-dimensional scaling technique (Goldstone, 1994; Kriegeskorte & Mur, 2012) (Figure 2C). Here, participants simultaneously arrange all items according to their similarities on a two-dimensional circular arena. Participants use mouse drag-and-drop to arrange the individual exemplars on a computer screen. After the initial arrangement of all items, participants repeat this task for different subgroups of items. The selection of the subgroups is achieved by an adaptive algorithm that zooms into different subsets of items to provide maximal information for the estimation of the RDM on each individual trial of the task. The inverse MDS technique estimates the RDM across the individual arrangements as a weighted average of the individual trials.

Representational similarities estimated from the multiple arrangements task reflect the structure of elicited representations in human inferior temporal cortex (Charest et al., 2014; Mur

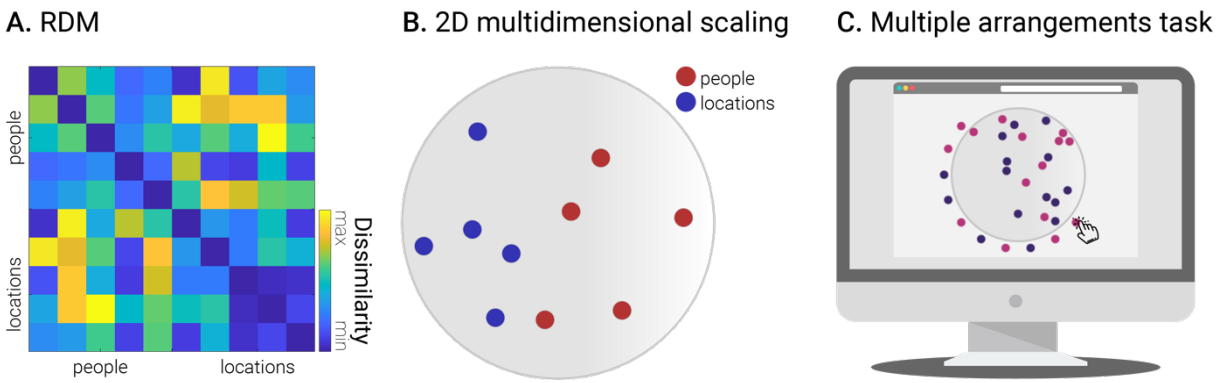


Figure 2. Multidimensional scaling and inverse multidimensional scaling. A. Exemplary Representational Dissimilarity Matrix (RDM). The RDM contains pairwise dissimilarities for all ten exemplars (here: five people and five locations, synthetic data). **B.** 2D multidimensional scaling of the RDM. Data are projected on a 2D surface such that the disparities between the true RDM and a recovered RDM from the 2D projection are minimized. **C.** Visualization of the multiple arrangements task that was used in the present thesis. Inverse MDS estimates the square RDM from the XY coordinates provided by repeated mouse drag and drop arrangements of individual tokens on the computer screen. Tokens are labeled with the names of personally known people and locations.

et al., 2013). Charest and colleagues (2014) presented participants with unfamiliar and personally known objects from their everyday life. Similarities estimated from the arrangement task were mirrored in the representational similarities in the inferior temporal cortex and were qualitatively different for personally known vs. unfamiliar material. Representational similarities may also be collected from different individuals or even across different species. Human participants estimated the similarity structure of visual stimuli using the multiple arrangements task. These arrangements predicted the structure of neural representations in both monkey inferior temporal cortex and different human participants (Mur et al., 2013). These results demonstrate the validity and versatility of the task.

Here, we use an adapted version of the multiple arrangements task where participants arrange tokens that are labeled with the names of known people and locations. Participants are instructed to arrange the exemplars according to their associations. Specifically, participants place items in close proximity if they are closely associated and as remote as possible if they are not. This task is employed in studies 2 and 3.

To quantify the individual expression in familiarity and liking, we take two different approaches. We either use ratings with nine-point Likert-scales or simplified one-dimensional versions of the arrangement task to quantify different variables of interest. In study 1 and 2, we use ratings to assess the degree to which participants like and know the individual people and locations. Ratings are done, one item at a time, using nine-point Likert-scales. In study 3, we employ the one-dimensional arrangement tasks. Here, participants place tokens labeled with the names of the known people and locations on a continuous scale in a single trial. The position on that scale serves as a metric for the variable under investigation.

Chapter 3

3 Study 1. Forming attitudes via neural activity supporting affective episodic simulations

This chapter is published as Benoit, R. G., Paulus, P. C., & Schacter, D. L. (2019). Forming attitudes via neural activity supporting affective episodic simulations. *Nature Communications*, *10*(1), 2215. https://doi.org/10.1038/s41467-019-09961-w

Abstract

Humans have the adaptive capacity for imagining hypothetical episodes. Such episodic simulation is based on a neural network that includes the ventromedial prefrontal cortex (vmPFC). This network draws on existing knowledge (e.g., of familiar people and places) to construct imaginary events (e.g., meeting with the person at that place). Here, we test the hypothesis that a simulation changes attitudes towards its constituent elements. In two experiments, we demonstrate how imagining meeting liked versus disliked people (unconditioned stimuli, US) at initially neutral places (conditioned stimuli, CS) changes the value of these places. We further provide evidence that the vmPFC codes for representations of those elements (i.e., of individual people and places). Critically, attitude changes induced by the liked US are based on a transfer of positive affective value between the representations (i.e., from the US to the CS). Thereby, we reveal how mere imaginings shape attitudes towards elements (i.e., places) from our real-life environment.

3.1 Introduction

A remarkable feat of the human mind is its ability to vividly imagine a plethora of prospective events (Schacter et al., 2017; Suddendorf & Corballis, 2007). The core brain network supporting such episodic simulations comprises parts of the medial surface including the vmPFC, lateral parts of the inferior posterior and temporal cortices, and the medial temporal lobes (Benoit & Schacter, 2015; Hassabis, Kumaran, & Maguire, 2007; Schacter et al., 2017). This network has been suggested to mediate episodic simulation by supporting the integration of elements from disparate episodic and semantic memories (e.g., of a liked person and a neutral but hitherto unrelated place) into novel events (e.g., meeting that person at that place for the first time) (Irish et al., 2012; Schacter et al., 2017; Schacter & Addis, 2007; Suddendorf & Corballis, 2007).

Simulating prospective events influences how we anticipate the future, for example by conveying the anticipated affective consequences of an imagined event (Benoit et al., 2018; Demblon & D'Argembeau, 2016). Here, we examine the hypothesis that it also changes how we value our immediate present by shaping real-life attitudes.

Episodic simulation creates an imaginary parallel to a situation of actually pairing a valenced unconditioned stimulus (US) with an initially neutral conditioned stimulus (CS). Such evaluative conditioning forms attitudes by changing the liking of the CS to align with the valence of the US (Hofmann et al., 2010; Jones et al., 2010; Wimmer & Shohamy, 2012). We hypothesize that imaginings of possible events (e.g., meeting a beloved person at a specific place) can effectively transfer affective value from one of the integrated elements (e.g., the person) to the other (e.g., the place). By this process, episodic simulations modify attitudes towards the very elements that the simulations had been based on, thus influencing how we evaluate our real-life environment.

Our hypothesis assigns a key role to the vmPFC in mediating such presumed attitude change. This region has been shown to integrate similar memories into schematic representations of the elements that are shared across those memories (Gilboa & Marlatte, 2017; Milivojevic et al., 2015; Richter et al., 2016; Schlichting et al., 2015). The concurrent reactivation of disparate representations, in turn, can support simulations of even novel experiences (Barron et al., 2013; Benoit et al., 2014).

Critically, the vmPFC does not only represent 'cool' models of the world (Metcalfe & Mischel, 1999). Activation in this region also generally varies with subjective value (Bartra et al., 2013; Peters & Büchel, 2010b), and it specifically scales with the affective quality of simulated experiences (Barron et al., 2013; Benoit et al., 2011, 2014; Lin et al., 2015). The

vmPFC has thus been associated with both schematic knowledge and the representation of affective value. During episodic simulation, these two functions are supported by overlapping parts of the vmPFC (Barron et al., 2013; Benoit et al., 2014; Lin et al., 2016; see also Shenhav et al., 2013), consistent with the hypothesis that this region codes for schematic representations that also entail associated affect, i.e., for 'hot' models of the world (Metcalfe & Mischel, 1999; Roy et al., 2012).

The vmPFC may thus code for affective representations of elements from our environment that can be flexibly integrated to support affective episodic simulations. Here, we hypothesize that such simulation-based integration induces experience-dependent plasticity in the neuronal coding of the individual elements (Barron et al., 2013). This plasticity could then enable the transfer of affective value from one element (i.e., the US) of the episode to the other (i.e., the CS).

To test this hypothesis, we combined a novel experimental procedure with functional MRI (fMRI) and representational-similarity analysis (Kriegeskorte et al., 2008) (Figure 3A). Before the fMRI session, participants provided names of places and people that they personally knew. For the latter, we specified that participants should name both people that they much liked, as well as those that they much disliked. They then rated the familiarity and liking (as an index of value) of each place and person (*pre-rating*). Based on these ratings, we selected places that the participants felt neutral towards (i.e., the CS) and paired these with either the most liked or much disliked people (i.e., the positive and negative US).

These elements and their pairings then featured during the three phases of the fMRI session (Figure 3A). During *phase I*, we presented each person and place, one at a time (i.e., the items were not presented as pairs during this phase), and participants vividly imagined interacting with the given person or acting in a way that would be typical for the location. During *phase II*, they encountered each person/place pairing repeatedly, and their task was to imagine interacting with the person in a location-specific manner. *Phase III* repeated *phase I* with a different presentation order. Finally, outside the scanner, participants rated the liking of each person and place again (*post-rating*) before they indicated the plausibility of meeting a given person at its paired location as well as the anticipated pleasantness of such an event.

This procedure allowed us to test the predicted impact of affective simulations on real-life attitudes and to examine key predictions of the neural basis of this effect. These predictions are based on the premise that the vmPFC does code for affective representations of elements from everyday life and our proposal that the attitude change is mediated by a transfer of affective value between such representations.

3.2 Methods

3.2.1 Participants

All participants reported no history of psychiatric or neurological disorders and gave informed consent as approved by the Harvard University Institutional Review Board (fMRI study) and the ethics committee of the University of Leipzig (replication study). All thirty participants of the fMRI study were right-handed, native English speakers, who all had normal (or corrected to normal) vision. Twelve participants had to be excluded either because of falling asleep in the scanner (two) or excessive head movements (ten) (defined as maximal absolute motion > 3 or more than 5 individual movements > 0.5 mm in any functional run). We thus included data from 18 participants (3 male; mean age: 21.33 years; range: 18 – 27). The replication study included thirty native German speakers (17 male; mean age: 23.97 years; range: 20 – 32) (as pre-registered to provide 80% power to detect an effect size of approximately 2/3 the original; https://aspredicted.org/blind.php?x=th9zv6).

3.2.2 Tasks and procedure

fMRI study

The procedure, adapted from Benoit et al. (2014), comprised a *preparation* and a *simulation* session (Figure 3A). During the *preparation* session, participants provided 100 places and 150 people that they were personally familiar with. Of the people, 100 had to be ones that they much liked, 30 people that they felt neutral towards, and 20 people that they much disliked. Participants then rated on 9-point scales (i) how familiar they were with each person and place (1: unfamiliar; 5: intermediate; 9: very familiar), indicating the degree of knowledge, and (2) how much they liked that given item (1: much dislike; 5: neutral; 9: much like), indicating the affective value.

We then selected 28 neutral places (i.e., with a rating of 5 and, if necessary, additional places with the next smaller and greater ratings), the 14 most liked people, and 14 of the least liked people. Piloting indicated that, overall, disliked people tended to be less familiar than liked people. To minimize this gap, we selected the disliked people (of all those that had received a likableness rating smaller than 4, or, if necessary, with the next greater ratings) that were most familiar. (However, the mean *Pearson* correlation between liking and familiarity across liked and disliked people remained at $r = 0.58$, $SD = 0.31$). Finally, we randomly combined each neutral place (i.e., the CS) with either a liked or disliked person (i.e., the US), thus creating 14 pairings in each condition.

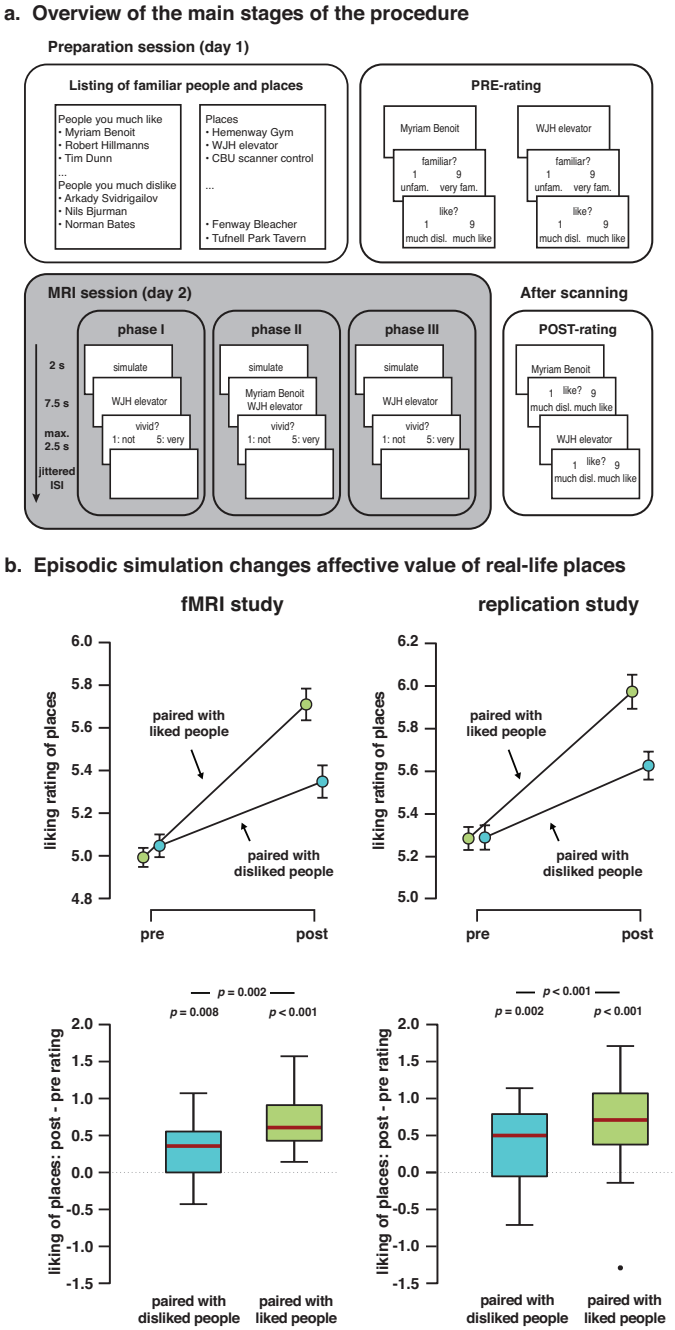


Figure 3. Main stages of the procedure and behavioral results. A. In an initial session, participants provided names of both liked and disliked people as well as of specific places from their everyday environment. They then rated their liking (indexing value) of and familiarity with each. Based on the ratings, we selected neutral places and combined each of those with either a liked or a disliked person. In a second session, participants were scanned with fMRI during three phases: In phase I, they imagined interacting with each person and place in isolation. In phase II, they were shown the critical pairings and imagined interacting with the respective person in a way that would be specific to that place. Phase III was identical to phase I except for a different presentation order. Finally, outside the scanner, participants re-rated their liking of each person and place. Moreover, they indicated, for each person-place pairing, the plausibility of such a meeting and the anticipated pleasantness. **B.** Consistent across a fMRI study and a preregistered replication study, we observed that places were deemed more positive following the integrative simulations. Critically, this pattern was stronger for places that had been the imaginary locations of meetings with liked than with disliked people. Episodic simulation thus induced a change in attitude of the US that was contingent on the valence of the CS. Error bars in the pre- vs. post panels indicate the respective standard error of the mean. Boxplots indicate the median, central quartiles, and +/- 2.7 SD. The dot denotes an outlier beyond that range.

At the beginning of the *simulation* session, participants received training on the tasks for the different phases (with items that were not part of the critical pairings). On any trial of phases I and III, they were presented with a fixation cross for 500 ms, followed by a person or

a place from the critical pairings for 7.5 s. During this time, participants imagined an episode of interacting with the person or place. They were instructed to imagine the episode as vividly as possible, ensuring that they have a clear mental picture of the respective item. They then rated the vividness of their imagination on a 5-point scale within a maximum of 3 s. The remainder of the maximal response time, if any, was added to the subsequent ITI, which lasted for at least 3 s plus an additional jittered period (0 to 8 s in 2 s intervals). The screen during the ITI was blank. In *phase II*, participants were presented with both the person and place of a given pairing, and then imagined an interaction with the person that would be specific to the given place as in Benoit et al. (2014).

The MRI session began with a resting state scan (not reported), before participants entered *phase I*. Here, they imagined each person and place across two functional runs. The person and place of a given pairing appeared in the same run in a pseudorandom order, with the constraint that one item appeared in the first and the other in the second half. During *phase II*, participants encountered each pairing three times in as many functional runs (pairs were presented in a different random order for each run). Participants were instructed to keep imagining the same episode for a given pairing, adding in more details and attempting to make it as vivid as possible. We had chosen three repetitions, because piloting indicated (i) that this was not too strenuous for the participants and (ii) that it was sufficient to yield the behavioral effect. *Phase III* repeated *phase I*, though with a newly pseudo-randomized presentation order and the additional constraint that the items that had been presented in the first run of *phase I* were also presented in the first run of *phase III*. Following this phase, participants performed a localizer task as in Benoit (2014) (results not reported).

Outside the scanner, we assessed the main behavioral dependent measure: Participants were shown each person and place in a random order and indicated their respective liking on the same scale as in the preparation session. Finally, participants were shown each pairing in a random order and rated the plausibility of such a meeting and its anticipated pleasantness (both on 9-point scales).

Replication study

The overall procedure of the pre-registered replication (https://aspredicted.org/blind.php?x=th9zv6) was identical to the fMRI study except for the omission of phases I and II. Moreover, to match the liked and disliked people in terms of familiarity, we used an alternative selection approach (Supplement A.1).

3.2.3 fMRI acquisition

Using a 3 Tesla Siemens Magnetom TimTrio MRI scanner with a 32-channel head coil, we acquired anatomical images with a T1-weighted magnetization-prepared rapid gradient multi-echo sequence (MEMPRAGE, 176 sagittal slices, TR = 2530ms, TEs = 1.64, 3.50, 5.36, and 7.22ms, flip angle = 7°, 1mm^3 voxels, FoV = 256mm). During each of seven functional runs, we acquired 220 volumes of blood-oxygen-level-dependent (BOLD) data with a T2*-weighted echo-planar imaging (EPI) pulse sequence that employed multiband RF pulses and Simultaneous Multi-Slice (SMS) acquisition (Moeller et al., 2010; Setsompop et al., 2012) with the following parameters: 69 interleaved axial-oblique slices (angled 17° towards coronal from ACPC), TR = 2000ms, TE = 27ms, flip angle = 80°, 2mm^3 nominal voxels, 6/8 partial fourier, FoV = 216mm, SMS = 3. The first five volumes of each run were discarded to allow for T1 equilibration effects.

3.2.4 fMRI analysis

Data were analyzed using SPM12 (www.fil.ion.ucl.ac.uk/spm). The functional images were realigned, corrected for slice acquisition times, and coregistered with the structural image. This was spatially normalized and the resulting parameters served to normalize the functional images by fourth-degree B-spline interpolation (preserving the functional voxel resolution of 2mm^3 isotropic) to the Montreal Neurological Institute reference brain. The images were then smoothed by an isotropic 8mm full-width half-maximum Gaussian kernel for the general linear models (GLM) assessing parametric modulations. The GLM that provided the input for RSA was based on unsmoothed data.

The GLMs analyzed regional activity by decomposing the variance in the BOLD time-series, separately for each functional run (Friston, Holmes, et al., 1995). Each model included six regressors representing residual movement artifacts and the mean over scans. A further regressor coded for the onsets and durations of trials for which participants did not provide a rating in time, if applicable. The additional regressors in a given GLM coded for the respective effects-of-interest by analyzing the remaining trials.

A first GLM assessed brain activation associated with the affective value of simulated items during phases 1 and 3. We therefore entered a regressor coding for the duration of all simulation trials plus an additional parametric regressor coding for the liking of the respective simulated item. Given that the paired simulations in *phase II* changed attitudes, we used the pre-ratings for *phase I* and post-ratings for *phase III*.

A second GLM assessed brain activation associated with the transfer of affective value during the integrative simulations in *phase II*. We entered (i) a regressor coding for the duration of all simulations, (ii) a first parametric modulator coding for the affective value of the US (i.e., liking of the person, averaged across pre- and post-rating), and (iii) a second parametric modulator coding for the change in value of the CS (i.e., post- minus pre-rating liking of the place). The first parametric regressor reveals regions where activation is modulated by the value of the US, whereas the second regressor indicates where the residual activation is greater in case of a more positive change in liking of the CS. Additional GLMs corroborated effects of affect-transfer without controlling for the effect of the US and controlling for the plausibility of the pairing. In two additional analyses, we further established this effect by controlling for the familiarity of the US – either by including it as a first parametric regressor or by computing the analysis based on the residuals of the change scores after regressing out possible effects of familiarity.

A final GLM estimated activity patterns separately for each simulation during phases 1 and 3 (thus including 112 regressors, one for each of the two simulations of the 28 places and 28 people). The ensuing parameters were used for the RSA (Kriegeskorte et al., 2008; Nili et al., 2014) that tested for individual representations in vmPFC.

All trial regressors were convolved with the canonical hemodynamic response function. A 1/128-Hz high-pass filter was applied to the data and the respective model, and parameter estimates for each regressor were calculated from the least-mean-squares fit of the model to the data.

Following Liu, Grady, & Moscovitch (2017), an anatomical mask of our region-of-interest, the vmPFC, was created by merging the gyrus rectus and the medio-orbital section of the frontal gyrus of the AAL template (Tzourio-Mazoyer et al., 2002) using the WFU-Pickatlas toolbox (Maldjian et al., 2003) (Figure 4A). For univariate effects, we extracted parameter estimates, for each participant, from this *a priori* ROI. For complementary and exploratory whole-brain analyses, the respective contrast estimates were entered into a second-level analysis, where we used cluster-level inference at $p < 0.05$ (FWE-corrected) with a cluster forming threshold of $p < 0.001$ and at least 15 contiguous voxels. These analyses also employed the vmPFC mask for targeted small-volume-correction. In addition, for an exploratory analysis of the hippocampus, we also used the AAL template (Tzourio-Mazoyer et al., 2002) to create a bilateral mask.

The RSA analyses were conducted using the toolbox by Nili et al. (2014). We only included trials on which participants had provided a response within the allotted time. Analyses

were based on the *t*-values of the estimated parameter estimates from each voxel within our ROI. *Within-item similarity* was assessed, for each person and place, by computing the *Pearson* correlation of these values between phases 1 and 3. *Between-item similarity* was only based on the correlations between elements of the same material and valence (e.g., only between liked people), to ensure that the results are not driven by category differences in neural coding. Moreover, due to temporal autocorrelations of noise, the activity patterns of proximal events tend to be more similar than of those events that are more distant in time (Alink et al., 2015). To quantify *between-item similarity*, we therefore only included similarity values of the same functional run as for the corresponding *within-item* comparison (i.e., correlating events from the 1st and 6th as well as from the 2nd and 7th functional runs only). Inferential statistics were based on *Fisher-z*-transformed correlation values.

3.3 Results

In the following, we first establish whether simulated episodes, similar to actual encounters (Hofmann et al., 2010; Jones et al., 2010), can shape real-life attitudes by reporting the behavioral results of an fMRI study (n = 18) and a preregistered replication (n = 30). We then examine the complementary hypothesis regarding the involvement of the vmPFC.

3.3.1 Episodic simulation changes attitudes towards real-life places

Based on the pre-ratings, we paired each neutral place with either a liked or disliked person (difference in liking: $t_{17} = 47.2$, $p < 0.001$, $d = 11.13$). The liked people, however, were also more familiar than the disliked people (difference in familiarity rating: mean = 2.23, standard error = 0.35; $t_{17} = 6.47$, $p < 0.001$, $d = 1.53$), while the places in the two conditions were well matched on this dimension ($t_{17} = -0.46$, $p = 0.65$, $d = -0.11$) (Supplement A.2).

Critically, in *phase II*, participants then repeatedly imagined interactions with each person at their respective paired place. Participants experienced episodes featuring liked people as more plausible ($t_{17} = 3.19$, $p = 0.005$, $d = 0.75$) and, importantly, also as more pleasant ($t_{17} = 15.88$, $p < 0.001$, $d = 3.74$). Mentally integrating elements into a common episode thus elicited an affective experience that was aligned with the valence of the US.

We predicted that the affective experience, in turn, would change attitudes towards the episodes' (initially neutral) locations. In particular, there should be a greater increase in liking for places that had been the stage for imaginary meetings with liked than with disliked people. We quantified the change in liking by computing the difference scores of the post- and pre-rating. Though these scores indicated that both kinds of places were deemed more positive following any simulation (paired with liked people: $t_{17} = 7.9$, $p < 0.001$, $d = 1.86$; paired with

disliked people: $t_{17} = 3.043$, $p = 0.008$, $d = 0.71$), this shift in attitude was indeed greater for places that had been imagined with liked people ($t_{17} = 3.68$, $p = 0.002$, $d = 0.87$) (Figure 3B; for concomitant changes in the attitudes towards the people, see Supplement A.3).

Given that the liked and disliked people also differed in familiarity, we examined the change in liking while controlling for this possible confound. In particular, for each difference score, we first regressed out the effect of familiarity. We then examined the residual scores, which were indeed still larger for the places imagined with liked than with disliked people (due to a deviation from normality, as indicated by Shapiro-Wilk, $W_{17} = 0.831$, $p = 0.004$, tested with a Wilcoxon test: $W_{17} = 143$, $p = 0.005$, matched rank biserial correlation $r = 0.67$). This result thus indicates that the attitude change towards the places was indeed based on the valence of the paired people rather than their familiarity.

3.3.2 The simulation-induced attitude change is a replicable phenomenon

Due to the novelty of this behavioral effect, we sought to determine its replicability by running a pre-registered study (https://aspredicted.org/blind.php?x=th9zv6). The procedure was identical to the fMRI study except for the omission of phases I and III. We also employed a modified algorithm that successfully matched the selected liked and disliked people on familiarity (Supplementary Methods A.1 and Supplementary Table A.2). Critically, this study yielded the identical pattern of a more positive change in liking for places that had been imagined with liked rather than with disliked people ($t_{29} = 3.77$, $p < 0.001$, $d = 0.69$) (Supplement A.4 and Figure 3B).

The mere act of imagining interactions can thus change real-life attitudes. In the following, we examine the hypothesis that such changes are mediated by a transfer of affective value between neural representations in the vmPFC. We tested three key predictions: First, a premise of the hypothesis is that neurons in the vmPFC code for representations of elements from our environment. Second, it posits that these representations entail information about the elements' affective value. Finally, the hypothesis proposes that the vmPFC mediates attitude changes by transferring affective value from the US (i.e., the person) to the CS (i.e., the place).

3.3.3 vmPFC activity patterns reflect the identity of imagined people and places

First, if the vmPFC codes for individual representations, then activation in the vmPFC should carry information about the identity of specific people and places. We tested this prediction by examining the replicability of simulation-induced activity patterns from *phase I* to *phase III* using representational-similarity analysis (RSA) (Kriegeskorte et al., 2008).

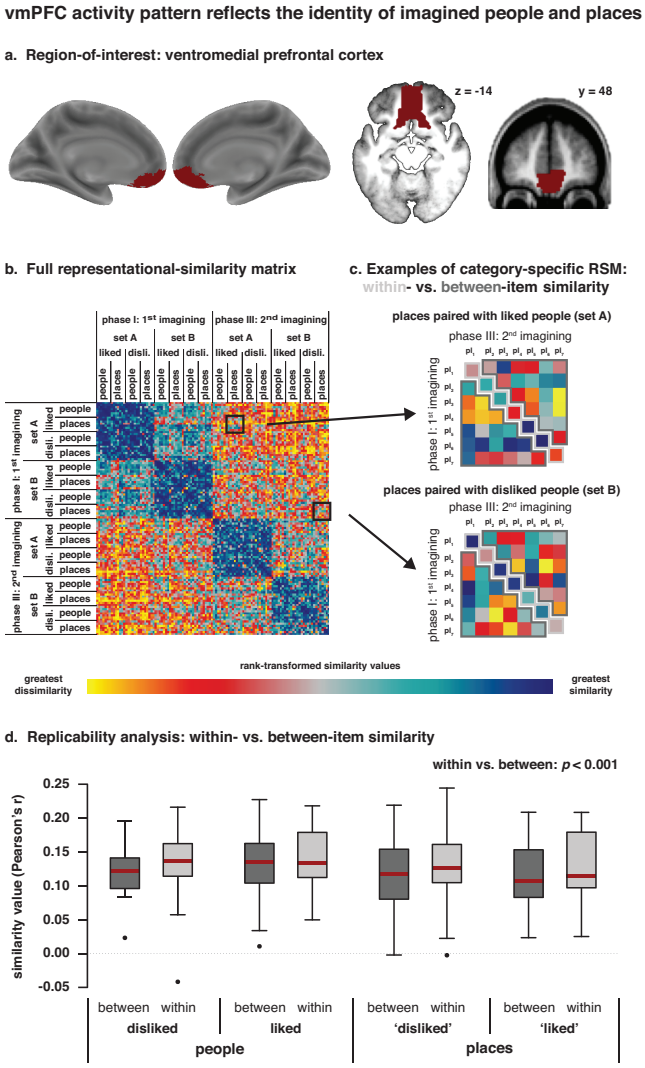


Figure 4. Representational similarity analysis (RSA) yields environmental representations in the vmPFC. A. The vmPFC region-of-interest used for all fMRI analyses. **B.** Full representational-similarity matrix indexing the similarity (expressed as Pearson correlation coefficients) between the activity patterns of any two simulated episodes. **C.** To test whether the vmPFC carried information about individual people and places, we examined the replicability of the associated activity patterns across phase I and III. If the vmPFC codes for information about particular people and places, we expect greater similarity for the repetition of the same element (within-item similarity) than for the comparison of an element with a different element of the same category (e.g., other places paired with liked people) (between-item similarity). **D.** Consistent with this prediction, we observed greater within- than between-item similarity across the different categories. The activity pattern in the vmPFC thus carries information about individual, personally-known people and places. Boxplots indicate the median, central quartiles, and +/- 2.7 SD. Dots denote outliers beyond that range.

Neuronal representations are assumed to be reflected in distributed activity patterns that can be assessed with fMRI (Charest et al., 2014; Kriegeskorte et al., 2008). Thus, to the degree that a specific representation is engaged whenever one imagines a particular person (or place), a similar activity pattern should get re-instated whenever one simulates an episode featuring the same person (or the same place). Accordingly, activity patterns should be more similar for the comparison of a given element with its repetition (*within-item* similarity) than with a different element at the time of the repetition (*between-item* similarity) (Charest et al., 2014). Moreover, if the activity pattern truly reflects the neural representation of a given element (e.g., a particular liked person) – rather than just category membership (e.g., all people) or valence

(e.g., all liked elements) - we expect the *within-item* similarity to be greater even when restricting the *between-item* similarity to elements of the same category (e.g., only other liked people).

We assessed the specific activity pattern associated with each simulation by modeling the fMRI time-series with a separate regressor for every episode. For each regressor, we then calculated the *t*-values of the parameter estimates (Kriegeskorte et al., 2008) and extracted a vector of all *t*-values from the voxels within an anatomical mask of the vmPFC (Figure 4A). A vector thus characterizes the activity pattern for a specific episodic simulation. Next, we quantified the neural similarity of any two simulations by computing the *Pearson* correlation of their activity patterns, yielding values ranging from 1 (i.e., greatest similarity) to -1 (i.e., greatest dissimilarity). We then analyzed the *Fisher-z*-transformed similarity values with a repeated-measures ANOVA that included the factors *comparison* (within, between), *material* (people, places), and *valence* (liked, disliked) (Figure 4). In addition to a significant effect of *material* ($F_{1,17} = 4.94$, $p = 0.04$, $\eta^2 = 0.23$), reflecting overall greater similarity for people than places, we also obtained the critical effect of *comparison* ($F_{1,17} = 18.86$, $p < 0.001$, $\eta^2 = 0.53$; see also Supplement A.5 for a control analysis corroborating that this effect does not merely reflect greater variation in the activity patterns due to greater variability in value for the *between-* than for the *within-item similarity* measure).

Episodic simulations are thus associated with replicable activity patterns in vmPFC that are more similar for repeated simulations of the identical element than for any two simulations of different elements. The results support the premise that this region encodes representations that can be re-instated during episodic simulation (Barron et al., 2013; Benoit et al., 2014). In the next section, we examine whether activation in the same region-of-interest (ROI) also contains information about associated affect.

3.3.4 vmPFC activation reflects the affective value of the simulated element

If vmPFC representations entail the affective value of specific elements, this should be reflected in the activation profile of this region (Bartra et al., 2013). We performed a parametric modulation analysis of the BOLD time series using the liking scores as an index of the elements' respective affective value. Given that the integrative simulation of people and places changed their affective value, we used the liking ratings of the pre-rating for *phase I* and of the post-rating for *phase III*. Note that the people contribute more variance in value to the analysis. People included liked and disliked exemplars, whereas the places were selected to be neutral.

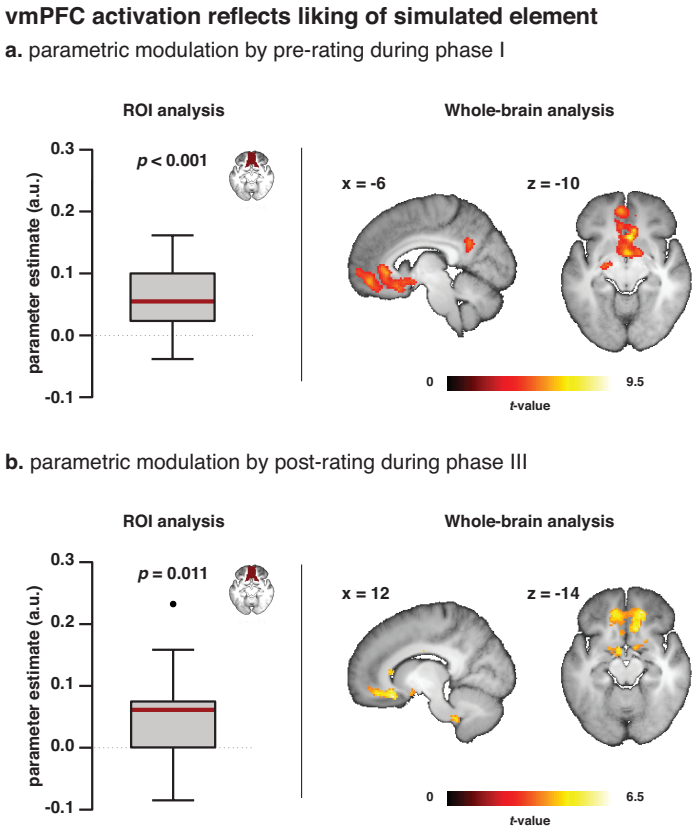


Figure 5. vmPFC activity reflects the value of the imagined element. Consistent across phases I and III, BOLD signal in the vmPFC was modulated by liking (as an index of value) of the people and places. Representations encoded in this region thus seem to carry information about the elements' affective value. Note that the greatest source of variance in value stems from the inclusion of liked and disliked people, though the model also predicts that BOLD signal for the neutral places should fall in between those extremes. Boxplots indicate the median, central quartiles, and +/- 2.7 *SD*. The dot indicates an outlier beyond that range. For display purposes, exploratory whole-brain maps are thresholded at $p < 0.001$, uncorrected with a cluster extend of at least 15 voxels.

However, the model also entails the prediction that activation for the neutral people should fall in between activation for the liked and disliked people.

Mirroring the ROI of the RSA analysis, we averaged across all parameter estimates within the anatomical mask of the vmPFC. Importantly, this analysis demonstrated that, for both time periods, activation in this region was modulated by the value of the simulated event (*phase I*: $t_{17} = 4.23$, $p < 0.001$, $d = 1$; *phase III*: $t_{17} = 2.85$, $p = 0.011$, $d = 0.67$) (Figure 5). The results were further corroborated by complementary whole-brain analyses. These revealed consistent modulation of brain activation in a cluster that included parts of the vmPFC (Figure 5 and Supplement A.6).

The foregoing analyses support the hypothesis that the vmPFC codes for affective representations of our environment that are engaged during episodic simulations. In the following, we examine the proposal that a transfer of affective value between such representations mediates the observed changes in attitude.

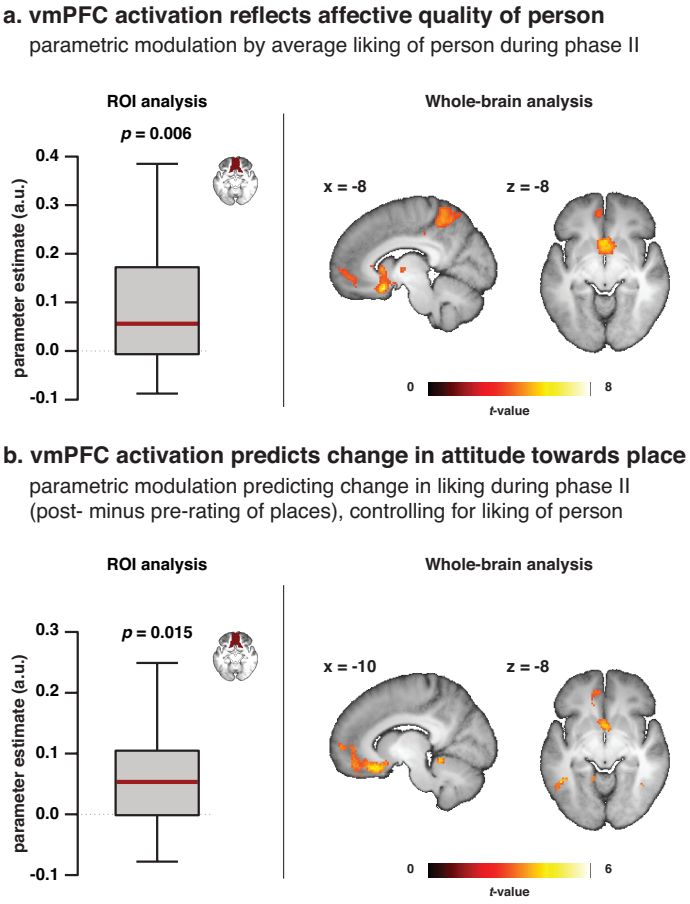


Figure 6. Transfer of affective value from the US to the CS during integrative simulations. A. BOLD signal in the vmPFC was modulated by the liking of the US (i.e., the person), reflecting the contribution of its affective value to the simulation. For display purposed, exploratory whole-brain maps are thresholded at $p < 0.001$, uncorrected with a cluster extend of at least 15 voxels. **B.** Even controlling for (a), BOLD signal in the vmPFC further predicted the ensuing change in liking of the CS, thus indicating a transfer of affective value. For display purposes, exploratory whole-brain maps are thresholded at $p < 0.005$, uncorrected with a cluster extend of at least 15 voxels. Boxplots indicate the median, central quartiles, and +/- 2.7 SD.

3.3.5 vmPFC activation during integrative simulations predicts attitude shifts

Our hypothesis posits that simulations change attitudes by transferring affective value from the US (i.e., the person) to the CS (i.e., the place). On one hand, during integrative simulations, activation in the vmPFC should thus be modulated by the liking of the US, reflecting its affective value. On the other hand, the activation should be predictive of the ensuing change in attitude towards the CS, indicating the transfer of affective value.

To test these two predictions, we performed a parametric-modulation analysis of the fMRI time-series obtained during *phase II*. As a first modulator, we included the liking of the person (averaged across the pre- and post-rating). This regressor thus yields regions where activation varies with the affective value of the US. As a second modulator, we included the change in liking for the respective place (i.e., post- minus pre-rating). This regressor thus identifies regions where greater activation during the joint simulations predicts a more positive shift in attitude for the CS (even when controlling for linear effects of US value). Both

regressors yielded the predicted modulations of vmPFC activation in our ROI (liking of US: t_{17} = 3.16, p = 0.006, d = 0.64; change in liking of CS: t_{17} = 2.7, p = 0.015, d = 0.75) (Figure 6). Again, this pattern was also evident in exploratory whole-brain analyses (Figure 6 and Supplement A.7).

The predictive signal in the vmPFC was also reliable when we analyzed the ROI data without controlling for the affective value of the US. A Shapiro-Wilk test suggested a deviation from normality (W_{17} = 0.83; p = 0.004), thus using a Wilcoxon test (W_{17} = 141, p = 0.014, matched rank biserial correlation r = 0.65). It was moreover significant when we controlled for the plausibility of the pairing (t_{17} = 2.48, p = 0.024, d = 0.58) (Supplement A.8). Critically, given the difference in familiarity for liked versus disliked people, it is important to note that we also obtained this effect when controlling for familiarity of the paired person - either by including it as a first parametric regressor (t_{17} = 2.65, p = 0.017, d = 0.62) or by first regressing out the contribution of familiarity from the individual change scores and then performing the parametric modulation analysis based on the residuals (using a Wilcoxon test: W_{17} = 136, p = 0.027, matched rank biserial correlation r = 0.59 due to a significant Shapiro Wilk test: W_{17} = 0.88; p = 0.029) (Supplement A.8). When we tested this effect in a control region, we did not observe a concomitant effect for an anatomical mask of the hippocampus (t_{17} = 0.04, p = 0.97, d = 0.009) a region that has previously been associated with the transfer of value between arbitrary and novel stimuli (Gilboa et al., 2014; Wimmer & Shohamy, 2012).

In summary, activation in the vmPFC was modulated by the affective value of the US and predicted the subsequent change in liking of the CS. The results therefore provide support for the hypothesized transfer of affective value. At the same time, they provide more general insights into the functions supported by the vmPFC.

3.4 Discussion

It is a long-standing view that the PFC supports control processes that operate on representations stored in posterior brain regions (Miller & Cohen, 2001). Somewhat contrary to this ostensible dichotomy, it has been suggested that the medial PFC also creates (schematic) representations of the environment, presumably by extracting commonalities across different episodes (Milivojevic et al., 2015; Richter et al., 2016; Schlichting et al., 2015). However, though activity patterns within this region have been shown to carry information about, for example, individual people (Szpunar et al., 2014; Thornton & Mitchell, 2017), locations (Robin et al., 2018), or the degree of connectedness within a social network (Parkinson et al., 2017), there is scarce evidence that the vmPFC codes more generally for representations of our

environment. The current data provide such evidence: In the vmPFC, replicable activity patterns emerged not only for particular known people but also for specific familiar places.

Though the current data indicate that the vmPFC represents information about individual entities (i.e., of individual people and places), this observation does not preclude the possibility that the representations are organized in a higher-level structure. Indeed, we further observed an overall greater pattern similarity for people than places, indicating that this region also codes for categorical information. More generally, neuroimaging evidence indicates that the medial PFC acts as a hub that integrates diverse information that is distributed across the brain (Benoit et al., 2014; Gilboa & Marlatte, 2017; Hassabis et al., 2014; Lim et al., 2013; van Kesteren et al., 2012). The integration may take the form of a dimension reduction that only codes for the information that is currently most relevant (Lim et al., 2013; Mack et al., 2020). Accordingly, the vmPFC may represent information along (hidden) dimensions rather than coding for distinct entities *per se*.

A dimensional coding is also consistent with previous observations that the relative activation in the vmPFC, when thinking about oneself versus other people, scales with the perceived similarity of the other person to oneself (Benoit et al., 2010; Mitchell et al., 2004). This suggests that the vmPFC does not code for individual people *per se* but for individual features along continuous dimensions that, in turn, differentiate individual people (Hassabis et al., 2014).

Importantly, in phases I and III, we observed that activation in the vmPFC also reflects the affective value of the constituting elements of an episode with an increase in activation from disliked via neutral to liked elements. Generally, the data are thus consistent with the hypothesis that the vmPFC codes for a continuum of value ranging from negative to positive rather than for other features such as salience (Bartra et al., 2013; Litt et al., 2011).

More specifically, the data support the proposal that representations in the vmPFC integrate conceptual information with associated affect and thus code for 'hot' models of the world (Benoit et al., 2014; Metcalfe & Mischel, 1999; Roy et al., 2012). In *phase II* of this study, the affective value of an episode was likely determined by the valence of the US (i.e., the person), given that the paired CS (i.e., the place) had been selected to be initially neutral. This point was corroborated by the finding that vmPFC activation during the integrative simulations was modulated by the liking of the respective featured person. Behaviorally, it was also reflected in a greater experienced pleasantness for episodes featuring liked than disliked people.

Importantly, the value signal in the vmPFC during episodic simulation has previously been shown to go beyond the value of the individual elements of the episode. That is, even when controlling for the nominal value of the constituting elements, this region signals the anticipated emergent value of the imagined scenario (Benoit et al., 2014). Episodic simulation may contribute to the processing of such emergent value by emphasizing the elements' features that are particularly salient in the imagined event (Lin et al., 2015, 2016). It thus affords an estimate of context-specific value that can deviate from the value that is more commonly attached to a given entity. Similarly, in the present experiment, we observed that vmPFC activation did not just reflect the value of the US. It moreover predicted the ensuing shift in attitude towards the CS. We suggest that this vmPFC signal reflects a prediction error regarding the CS that indicates the degree to which the experienced affect deviates from the expectation (e.g., more pleasant than expected) (Garrison et al., 2013; see also Lin et al., 2015). This signal may then drive plasticity in the representation of the CS and lead to the updating of its value (Garrison et al., 2013). Given that the vmPFC signal codes for a continuum from negative to positive value (Bartra et al., 2013; Litt et al., 2011), this mechanism may support both downward and upward value-updating.

We had hypothesized a particular involvement of the vmPFC in mediating simulation-induced attitude changes due to the region's dual contribution to the representation of schemas (Gilboa & Marlatte, 2017; Milivojevic et al., 2015; Richter et al., 2016; Schlichting et al., 2015) and the processing of value (Bartra et al., 2013). However, we do not suggest that this region performs this function in isolation. Specifically, the striatum has long been associated with the transfer of value from a US to a CS (Shohamy, 2011). The current work indicates that striatal activity also tracks the value of the imagined US during the simulation of new events (see also Benoit et al., 2014). This information may then be conveyed to the vmPFC and interact with the existing schematic representation of the CS to process an updated value estimate.

It is noteworthy that we did not observe concomitant effects in the hippocampus. This region, and its interactions with the striatum, has previously been implicated in the transfer of value (Gilboa et al., 2014; Wimmer & Shohamy, 2012). We think it is critical to note that these studies examined such transfer between arbitrary combinations of novel stimuli. The hippocampus may be particularly important in such situations, because they require the rapid encoding of both the individual items and their relations (Kumaran et al., 2016, 2016; see also Wimmer et al., 2012). By contrast, in the current study, we examined changes in attitude towards well-established elements from participants' real-life environment. As such, the integrative simulations could be based on the co-activation of established knowledge structures

that are already represented in the vmPFC (Barron et al., 2013; Benoit et al., 2014; Gilboa & Marlatte, 2017; Milivojevic et al., 2015; Richter et al., 2016; Schlichting et al., 2015). Therefore, simulation-induced attitude changes may be less reliant on hippocampal processes than more episodic forms of value transfer (as in Wimmer et al., 2012). It will be an important avenue for future studies to systematically delineate the contributions of the striatum, hippocampus, and vmPFC as well as their interactions (Gerraty et al., 2014; Shohamy & Daw, 2015).

Episodic simulation has previously been shown to have powerful influences on how we perceive and plan for the future. It increases the perceived plausibility of a prospective scenario (Gregory et al., 1982; Szpunar & Schacter, 2013) and conveys its anticipated affective experience (Benoit et al., 2011, 2014; Demblon & D'Argembeau, 2016). This experience, in turn, can foster more farsighted decisions by increasing the salience of future rewards (Benoit et al., 2018). Similarly, simulations of even unlikely future threats can help avoiding grave danger (Bulley et al., 2017; Miloyan & Suddendorf, 2015). However, such simulations may also contribute to the development of depression and anxiety (Holmes et al., 2011; Miloyan et al., 2014). The current data show that imaginings can further have a fundamental impact on how we evaluate our environment in the present.

In the fMRI study, we observed that merely imagining meeting a known person at a familiar place can boost the value that we attach to that location. Importantly, we obtained the same effect, with a similar effect size, in a preregistered study with a larger sample size, thus demonstrating the replicability of the simulation-induced attitude change. The extent of this effect was associated with vmPFC activation during the integrative simulations. This observation indicates that the attitude change was induced at that stage rather than as a process of deliberative revaluation during the post-test.

Somewhat surprisingly, we also consistently observed a positive change in liking for places imagined with disliked people. We caution that our design did not include a baseline condition (such as neutral places imagined with neutral people) that would have allowed us to infer simple effects, such as mere exposure (Zajonc, 2001), that could potentially account for a general positive shift in liking. Such an effect may boost the value of even those places that had been imagined with disliked people, thus possibly masking any downward impact of simulations. Critically, however, the change in attitude was more positive for places that had been the location for meetings with liked than with disliked people, indicating a critical influence of the US's valence. The results thus demonstrate how mere imaginings can have a similar impact on our attitudes as real happenings (Hofmann et al., 2010; Jones et al., 2010).

The observation that episodic simulations can change our attitudes towards the very basic elements that the simulations had been based on has potentially wide-ranging implications. Exaggerated simulations of prospective rewards and threats can provide adaptive benefits by inducing biases that motivate farsighted decisions (Bulley et al., 2017; Miloyan & Suddendorf, 2015). Critically, however, the current data suggest that exaggerated simulations can also produce a distorted model of the environment that becomes decoupled from actual experiences. This mechanism going awry may thus contribute to the development and maintenance of mental health problems such as depression, bipolar disorder, and anxiety that are often characterized by pronounced prospective thoughts (Benoit et al., 2016; Holmes et al., 2011; Miloyan et al., 2014). More generally, the findings highlight the powerful function of simulations not just in guiding future-oriented decisions but also in creating our models of the world.

Acknowledgements

This work was supported by the National Institute of Mental Health (MH060941 to Daniel L. Schacter), a National Institutes of Health Shared Instrumentation Grant (S10OD020039 to Harvard University Center for Brain Sciences) and a Max Planck Research Group (awarded to Roland G. Benoit). We thank Roxanne Eisenbeis, Janine Held, Leonie Kanne, Sarah-Lena Schaefer, and Sadie Zacharek for assistance in data collection, Himanshu Bhat and Thomas Benner of Siemens Healthcare for the SMS-EPI sequence, and Steven Cauley of Massachusetts General Hospital for modifications that enabled implementation of our protocols in a routine session.

Author contributions

Roland G. Benoit and Daniel L. Schacter designed the study, Roland G. Benoit collected the data for the fMRI study and Philipp C. Paulus for the replication study, Roland G. Benoit and Philipp C. Paulus analyzed the data, and all three wrote the manuscript.

Chapter 4

4 Study 2. Simulation-based learning: how imaginings shape real-life attitudes

This chapter is currently under review at Cognition and available as a preprint: Paulus, P. C., Dabas, A., Felber, A., & Benoit, R. G. (2021). Simulation-based learning influences real-life attitudes [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/k8nxm

Abstract

Humans can vividly simulate hypothetical experiences. This ability draws on our memories (e.g., of familiar people and locations) to construct imaginings that resemble real-life events (e.g., of meeting a person at a location). Here, we examine the hypothesis that we also learn from such simulated episodes much like from actual experiences. Specifically, we show that the mere simulation of meeting a familiar person (unconditioned stimulus; US) at a known location (conditioned stimulus; CS) changes how people value the location. We provide key evidence that this simulation-based learning strengthens pre-existing CS-US associations and that it leads to a transfer of valence from the US to the CS. The data thus highlight a mechanism by which we learn from simulated experiences.

4.1 Introduction

Humans possess the remarkable ability to vividly imagine hypothetical episodes that could have happened in the past or may take place in the future (de Brigard & Parikh, 2019; Schacter et al., 2007; Suddendorf & Corballis, 2007). Such episodic simulation shares many similarities with episodic memory (Schacter et al., 2017). For example, it is largely supported by the same neural network (Benoit & Schacter, 2015) and exhibits similar phenomenological qualities (D'Argembeau & Van der Linden, 2004). These commonalities suggest that episodic simulations are based on the recollection of memories (e.g., of known people and locations) that get flexibly recombined into fictitious episodes (e.g., of meeting a person at a location) (Benoit et al., 2014; Schacter et al., 2007).

Given this strong relationship between simulating and remembering, can we also *learn* from imagined events much as we learn from actual past experiences? Evidence from motor learning supports this idea: Imagining to execute a specific action can boost performance akin to actually carrying out that action (Driskell et al., 1994). Moreover, fear conditioning can arise not only from the actual experience of an unconditioned stimulus (US) but also from merely imagining an aversive event (e.g., stepping onto a thumbtack) (Mueller et al., 2019).

Simulated experiences also seem to shape our attitudes in a similar fashion as real experiences. We have recently described that merely imagining a meeting with a known person (serving as US) at an initially neutral location (serving as conditioned stimulus; CS) changes how much we like the location (Benoit et al., 2019). These simulations induced a positive shift in attitude that was more pronounced following imaginary meetings with liked than with disliked people. This effect resembles the phenomenon of evaluative conditioning, where the actual co-occurrence of a positive or negative US leads to a transfer of its respective valence to a paired CS (Hofmann et al., 2010; Walther et al., 2018; see also Wimmer & Shohamy, 2012).

Here, we seek to further establish the mechanism that supports simulation-based learning and to gauge its similarity with experience-based learning. First, evaluative conditioning has been argued to involve an associative integration of the US and CS (Forester et al., 2020; Madan & Kensinger, 2021; Palombo et al., 2021; Stahl & Aust, 2018). If simulation-based learning is based on a similar integrative mechanism, we expect simulations to strengthen the associations between the paired US (here: the imagined person) and CS (here: the imagined location) (see also Benoit et al., 2019). By this, simulations would reshape the very semantic space that they operate on.

Second, we further scrutinize whether simulation-based changes in attitude indeed reflect a transfer of valence from the US to the CS. Compared to a neutral baseline condition,

simulations featuring liked people should cause an upward shift and simulations with disliked people a downward shift in the liking of the locations. Third, we hypothesized that the integrative simulations induce an affective experience consistent with the nature of the US (Benoit et al., 2019). That is, we predict greater skin conductance levels for simulations featuring liked and disliked compared to neutral people, and expect that the experienced pleasantness is contingent on the valence of the US. Finally, we predict that this affective experience accounts for the transfer of valence.

4.2 Methods

4.2.1 Participants

We recruited 48 healthy adults from the database of the Max Planck Institute for Human Cognitive and Brain Sciences, all of which had normal or corrected-to-normal vision, provided written informed consent, and received monetary compensation for participating. The experimental protocol was approved by the Ethics Committee at the Medical Faculty, Leipzig University, Germany; reference number 403/16-ek). Two participants had to be excluded due to missing data and three for an insufficient number of provided disliked, neutral, or liked people (see below). We thus included data from 43 participants (22 females, 21 males; *mean* age = 24.2 years, *SD* = 3.1, *range*: 19 to 35). The included sample yields approximately 80% power to detect an effect half the size as in Benoit et al. (2019). Due to recording errors, five further participants were excluded from analyses of the psychophysiological data (leaving 38 participants; 20 females, 18 males, *mean* age = 24.4 years, *SD* = 3.2, *range*: 19 to 35).

4.2.2 Task and procedures

The procedure, adapted from Benoit et al. (2019), consisted of a preparation and a simulation session (Figure 7). During the preparation session (*mean* duration = 4.5h, *SD* = 1.2, *range:* 2.8 to 7.2), participants provided 150 familiar locations with the help of Google Maps and 150 personally known people using their Facebook contact lists as aid. To ensure sufficient variance in liking and familiarity to form all experimental conditions, participants were instructed to name the 100 people they liked most, 30 they felt neutral toward, and 20 they disliked. Participants rated all people and locations (i) according to their familiarity (1: not at all, 5: intermediate, 9: very much) and (ii) according to their liking (1: not at all, 5: neutral, 9: very much) as an index of valence.

We selected 30 neutral locations (*mean* = 5.06, *SD* = 0.28) as well as ten disliked (*mean* = 1.96, *SD* = 0.55), ten neutral (*mean* = 5.13, *SD* = 0.21), and ten liked people (*mean* = 7.8,

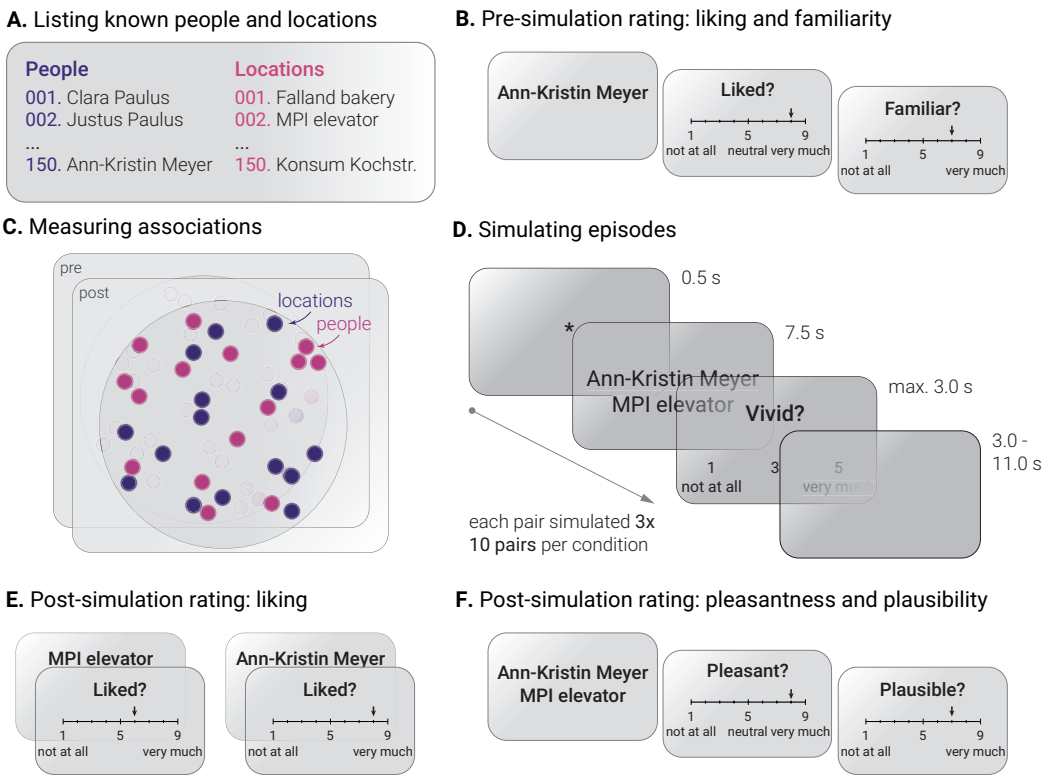


Figure 7. Overview of the experimental procedure. A. Participants provided lists of 150 known people and locations. **B.** Participants rated how much they liked and how well they knew the people and locations. We used these ratings to select 10 disliked, 10 neutral, 10 liked people, and 30 neutral locations. Each selected person was then randomly paired with one of the locations. **C.** Both before and after the main part of the experiments, participants arranged the selected people and locations according to their associations. Participants placed two exemplars closely together if they associated them strongly and far apart if they did not. **D.** Participants repeatedly simulated imaginary encounters with the known people at their paired location. After the simulation, participants rated the vividness of the simulated episode. **E.** After the simulation task, participants rated again how much they liked the selected people and locations. **F.** Participants also indicated how pleasant the simulated episode had been and how plausible it would be to actually meet the person at the paired location.

SD = 0.47) (see Supplement B.1 for details of the selection procedure). We then randomly assigned each neutral location (i.e., the CS) uniquely to either a liked, neutral, or disliked person (i.e., the US).

We quantified the pre-experimental associations among the selected people and locations using the multiple-arrangements task (Kriegeskorte & Mur, 2012): participants arranged the 60 names on a circular arena so that strongly related exemplars were placed closely together and unrelated exemplars at opposing sides of the arena (Figure 7C). As a measure of associatedness, we computed the difference score between the proximity of a given person to its paired location (*within pair*) and the average proximity of that person to all other locations (*outside pair*) (Nili et al., 2020; Paulus et al., 2020).

Participants returned for the simulation session (*mean* delay = 1.1 day, *SD* = 0.4, *range*: 1 to 2). Each trial commenced with a 0.5 s fixation period, followed by the simulation period during which a person-location pairing was presented for 7.5 s. During this time, participants vividly simulated interacting with the person in a manner that would be specific for the location (e.g., at a restaurant, discussing appetizers). Participants rated the vividness of their simulation

(1: not at all, 5: very much) within 3 s. The remainder of this time, if any, was added to the inter trial interval that lasted for 3 s plus a jitter (0-8 s in 2 s intervals). After participants had been trained on the task, we randomly presented all pairs once in each of three runs. On the second and third encounter of the pairs, participants were instructed to simulate the same episodes as in the original encounter and to use the additional time to make their simulations even more vivid or longer (Figure 7D).

Afterwards, participants rated their liking of each person and location as well as the plausibility (1: very implausible, 5: somewhat, 9: very plausible) and pleasantness of the imagined meetings (1: very unpleasant, 5: neutral, 9: very pleasant). They then again indicated the person-location associations using the multiple-arrangements task. Upon completing the main task, participants completed the Big Five Inventory (Rammstedt & John, 2005) as well as questionnaires on mind wandering and their ability to imagine visual scenes (not reported). A subsample was also invited to pilot a task similar to an implicit association test currently in development (not reported).

4.2.3 Psychophysiological recordings and analysis

Skin conductance responses (SCRs) were recorded using *Ag/AgCl* electrodes attached to the thenar and hypothenar of the non-dominant hand. Signals were recorded at 1,000 Hz using a Biopac (BIOPAC Systems Inc., Santa Barbara, CA) MP 150 data acquisition system running AcqKnowledge (4.3). For exploratory purposes, we also monitored heartbeats (not reported).

We quantified SCRs during the simulation trials following the recommendations of the Society for Psychophysiological Research (Boucsein et al., 2012) implemented in the Psychophysiological Modeling toolbox (4.0.2, http://bachlab.org/pspm) in Matlab 2017b (9.3, Mathworks, Natick, MA).

Statistical analysis

Statistical analyses were done in *R* (3.6.2, www.r-project.org) using repeated measures ANOVAs including the factors *Valence* of the US (*disliked*, *neutral*, *liked*) and, where appropriate, *Time* (*pre-*, *post*-simulation). We accounted for violations of sphericity using the Greenhouse-Geisser method and, in planned follow-up tests, for multiple comparisons using Holm's method. Whenever a Shapiro-Wilk test indicated a deviation from normality, we used Wilcoxon rather than Student's *t*-tests. We tested our directed hypotheses with one-tailed tests.

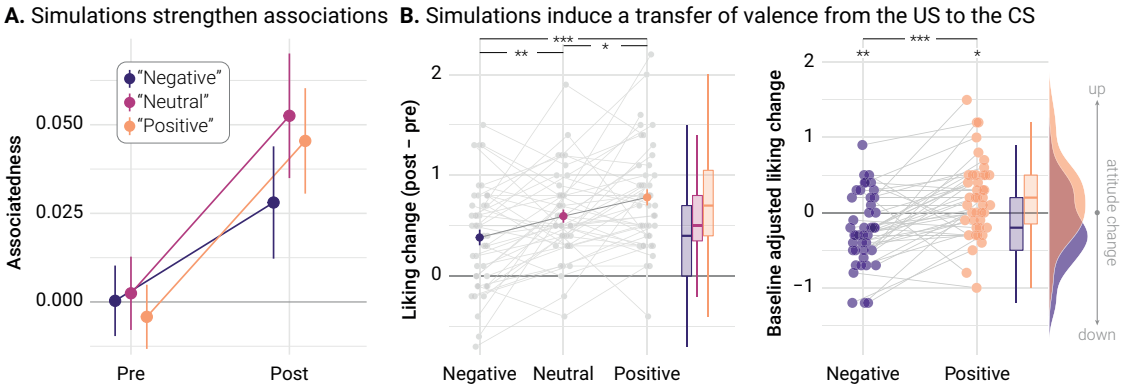


Figure 8. Simulation-based learning of real-life attitudes. A. Episodic simulations strengthen already existing associations between paired people and locations, irrespective of the person's valence. **B.** Left: Episodic simulations induce a transfer of valence from the person (US) to their paired location (CS). Right: Simulated episodes with disliked people yielded a significant transfer of negative valence and episodes with liked people a significant transfer of positive valence as compared to the neutral baseline condition. Dots and whiskers in A, B left: *mean* ± *SEM*. Dots in B right: condition mean individual participants, Box-plots: center line, median; box-limits, first and third quartile; whiskers, 1.5x interquartile range. *** - $p < .001$, ** - $p < .01$, * - $p < .05$.

4.3 Results

4.3.1 Simulations strengthen associations between paired people and locations

We first tested whether episodic simulations strengthened the associations between people and their paired locations by comparing their associatedness on the pre- and post-test. The rANOVA revealed a significant effect of *Time* only ($F(1,42) = 10.59, p = .002, \eta^2_G = .06$; *Valence*: $F(1.96,82.3) = 0.61, p = .542, \eta^2_G = .004$; *Time* × *Valence:* $F(1.96,82.28) = 1.37, p = .259, \eta^2_G = .004$). Thus, regardless of valence, episodic simulations strengthened the associations between the people and their paired locations (Figure 8A). We next examine whether this integration of memory representations was accompanied by the predicted transfer of valence.

4.3.2 Episodic simulations change the liking of the simulated locations

We predicted a transfer of positive valence from the liked, and a transfer of negative valence from the disliked people. To test this prediction, we assessed how the liking of the paired locations changed from the pre- to the post-rating. We observed a significant effect of *Time* ($F(1,42) = 99.98, p < .001, \eta^2_G = .32$), no effect of *Valence* ($F(1.88, 78.90) = 1.95, p = .152, \eta^2_G = .01$), but, crucially, the significant interaction ($F(1.91, 80.25) = 12.39, p < .001, \eta^2_G = .04$; Figure 8B left).

The *Time* effect reflected a positive shift in liking for all locations. This shift may be due to generic effects such as mere exposure (Benoit et al., 2019; Zajonc, 2001) or increased availability (Tversky & Kahneman, 1974). To control for this generic upward shift, we used the paired locations from the neutral condition as a baseline. We thus could unmask possible

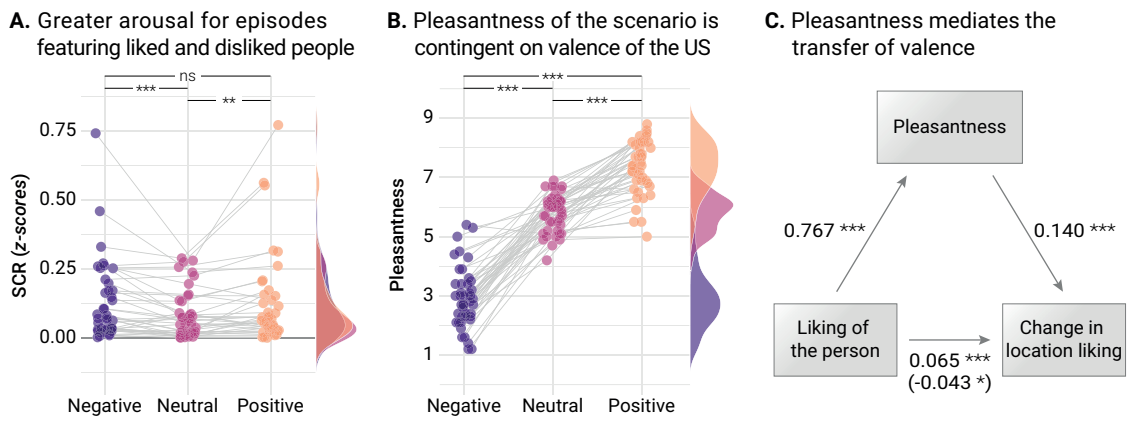


Figure 9. The transfer of valence is mediated by the affective experience of the simulated episode. A. Evoked skin conductance responses (SCRs) are significantly larger for simulations featuring both disliked and liked people, as compared to the neutral baseline condition. Due to a deviation from normality, tested with a non-parametric test. **B.** Compared to the neutral baseline condition, episodes featuring liked people were experienced as more pleasant and episodes featuring disliked people as less pleasant. **C.** The effect of the liking of the person on the change in liking of the location is mediated by the pleasantness of the simulated episode. *** - $p < .001$, ** - $p < .01$, * - $p < .05$.

valence-specific effects. Specifically, if the interaction truly reflects a transfer of valence, we expected *opposite* shifts in liking for locations imagined with liked versus disliked people compared to those imagined with neutral people.

Indeed, relative to the neutral baseline condition simulations with liked people yielded the significant relative upward shift ($t(42) = 2.39, p = .011, d = .36$), whereas simulations with disliked people yielded the significant relative downward shift ($t(42) = 2.88, p = .003, d = .44$; Figure 8B right). The analyses thus provide evidence for opposite effects of imagining positive versus negative US. We obtained the same effect when statistically controlling for the familiarity of the people (see Supplement B.2 and B.3).

4.3.3 Stronger affective responses in simulations featuring liked and disliked people

We next tested whether simulations induce an affective experience that is contingent on the valence of the US. We therefore examined two facets of participants' affective responses: (i) Their arousal as indexed by skin conductance and (ii) the valence of their experience as indexed by the pleasantness ratings.

The analysis of the SCRs revealed a significant effect of *Valence* ($F(1.66,61.5) = 5.59, p = .009, \eta^2_G = .02$). Simulations featuring liked ($W = 176, p = .002, r = 0.46$; Shapiro-Wilk: $W = 0.69, p < .001$) and disliked people elicited stronger SCRs ($W = 604, p < .001, r = 0.55$; Shapiro-Wilk: $W = 0.81, p < .001$) than simulations featuring neutral people (with no difference for disliked versus liked people; $W = 365, p = .943$, *two-tailed*, $r = 0.01$; Shapiro-Wilk: $W = 0.91, p = .005$). The SCR data thus corroborate a stronger arousal during simulations including liked and disliked people (Figure 9A).

The analysis of the episodes' pleasantness yielded a significant effect of *Valence* of the person ($F(1.51,63.56) = 292.21, p < .001, \eta^2_G = .82$). As expected, planned comparisons with the neutral condition revealed significantly lower pleasantness for simulations featuring disliked ($t(42) = 16.06, p < .001, d = 2.45$) and greater pleasantness for simulations featuring liked people ($t(42) = 11.23, p < .001, d = 1.71$) (Figure 9B).

4.3.4 The transfer of valence is mediated by the experienced pleasantness

We had hypothesized that the experienced pleasantness would account for the transfer of valence from the US to the CS. Indeed, a causal mediation analysis revealed a mediation of the transfer of valence from the person to its paired location by the pleasantness of the simulated episode. The indirect effect was 0.108 ($p < .001$) and the average direct effect was -0.043 ($p = .048$) (Figure 9C).

Thus, episodic simulations induce an affective experience aligned with the valence of the US. This experience, in turn, accounts for the ensuing changes in valence of the CS.

4.3.5 Exploratory: Neuroticism and simulation-based learning

People high in neuroticism are at a greater risk of developing mood disorders (Lahey, 2009) and exhibit reduced spontaneous positive future thought (Gamble et al., 2019; MacLeod & Byrne, 1996). We thus explored whether neuroticism is associated with weaker learning from simulated positive experiences. Indeed, individuals higher in neuroticism exhibited a weaker upward shift in liking (i.e., a less positive change from the pre- to the post-test) (Spearman's $\varrho = -0.38, p = .012$; Supplement B.4).

4.4 Discussion

Remembering the past and imagining the future share many similarities (Schacter et al., 2007, 2017). Here, we corroborate that we also learn from simulated experiences much as we learn from actual past experiences (Benoit et al., 2019; Driskell et al., 1994; Mueller et al., 2019). Specifically, we build on our previous observation that simulations can change attitudes towards our real-life environment. Changing existing attitudes tends to be more difficult than forming new ones (Jones et al., 2010). It is therefore particularly remarkable that simulations affected evaluations of locations that the participants were already personally familiar with. Critically, this study furthers our understanding of the mechanism underlying such simulation-based learning.

We demonstrate that simulations strengthen pre-existing associations between jointly simulated people (US) and places (CS). This process may induce experience-dependent

plasticity that allows for the transfer of valence from the US to the CS (Barron et al., 2013). Indeed, the episodic integration of the CS and US has been highlighted as a prerequisite for real experiences to induce evaluative conditioning (Palombo et al., 2021; see also Madan & Kensinger, 2021; Forester et al., 2020).

The observed strengthening of the CS-US associations indicates that mere simulations can fulfil the same prerequisite. Notably, we show that simulations not only establish novel associations (e.g., Martin et al., 2011). Like real experiences, they can also affect the associative strength of semantic representations, such as of personally familiar people and locations (Renoult et al., 2012). Episodic simulations thus seem to modify the overall configuration of the semantic space that they operate on.

How do the integrative simulations mediate the transfer of valence? It has been proposed that evaluative conditioning requires the binding of US and CS features into a common representation (Walther et al., 2018). Our analysis indicates, more specifically, that this transfer hinges on the experienced affect. On the one hand, this affective experience may be encoded as an episodic memory. Subsequent deliberations on the CS could then lead to the retrieval of this experience and thus influence its evaluation (Kensinger & Ford, 2020). On the other hand, the experienced affect may directly become integrated into the existing representation of the CS (Madan & Kensinger, 2021).

Our data moreover demonstrate that simulation-based learning is indeed valence-specific. We were previously unable to quantify unspecific changes arising from general effects such as mere exposure (Zajonc, 2001) or availability (Tversky & Kahneman, 1974). We could thus not determine whether simulations can also cause a downward shift in attitude. By including a neutral condition as a baseline for such generic effects, we have now established that the simulation-based changes are contingent on the valence of the US. In its directionality, the transfer of valence thus resembles the extant literature on experienced-based evaluative conditioning (Hofmann et al., 2010).

The bidirectional transfer of valence has possible clinical implications. Our analysis indicated that people high in neuroticism learn less from positive simulations. They may thus build a model of their environment that is primarily based on negative simulations. This effect may contribute to the development and maintenance of affective disorders (see also Bulley et al., 2017; Renner et al., 2017). It may accordingly sometimes be more beneficial for one's well-being to stop imaginings of hypothetical events (Benoit et al., 2016).

There are several limitations to this work. Notably, the temporal stability of this effect warrants further investigation. The literature on evaluative conditioning indicates that

experience-based attitude changes last for extended periods (Jones et al., 2010). The similarities to simulation-based learning suggest that also mere imaginings can have an extended influence.

Moreover, it would be desirable to investigate the question of a dose-response relationship. Does the number of repeated simulations determine the strength of the learning effect as is the case in experience-based learning (Madan & Kensinger, 2021) and is it also governed by similar computational mechanisms? We suggest that our experimental approach provides an avenue for exploring these questions.

To conclude, episodic simulation does not only influence our outlook towards the future (Benoit et al., 2018; Rösch et al., 2021). It also constitutes a learning device that influences how we evaluate our everyday environment. Indeed, simulations may contribute to our view of the world in a similar fashion as our actual experiences.

Author contributions

Philipp C. Paulus, Aroma Dabas, and Roland G. Benoit designed the study. Philipp C. Paulus and Annalena Felber collected the data. Philipp C. Paulus wrote the analysis scripts and analyzed the data together with Roland G. Benoit. Philipp C. Paulus and Roland G. Benoit wrote the original draft of the manuscript and all authors contributed to review and editing of the manuscript.

Acknowledgements

This work was supported by a Max Planck Research Group (awarded to Roland G. Benoit). We thank Roxanne Eisenbeis, Felicia Heilgendorff, and Vera Seyffert for assistance in data collection as well as Heidrun Schultz and Angharad Williams for comments on a draft of this manuscript.

Data and code availability

All data, custom code, and the experiment scripts (except for the multiple arrangements task) are publicly available at the Open Science Framework:
https://osf.io/9yt7s/?view_only=c532cb3e669749a9ac503b75a601b643

Chapter 5

5 Study 3. Value shapes the structure of schematic representations in the mPFC

This chapter is available as a preprint: Paulus, P. C., Charest, I., & Benoit, R. G. (2020). Value shapes the structure of schematic representations in the medial prefrontal cortex [Preprint]. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.08.21.260950v3

Abstract

Adaptive cognition is fostered by knowledge about the structure and value of our environment. Here, we hypothesize that these two kinds of information are inherently intertwined as value-weighted schemas in the medial prefrontal cortex (mPFC). Schemas (e.g., of a social network) emerge by extracting commonalities across experiences and can be understood as graphs comprising nodes (e.g., people) and edges (e.g., their relationships). We sampled information about unique real-life environments (i.e., about personally familiar people and places) and probed the neural representations of their schemas with fMRI. Using model-based representational-similarity analysis, we show that the mPFC encodes indeed both, the nodes and edges of the schemas. Critically, as hypothesized, the strength of the edges is not only determined by experience and centrality of a node but also by value. We thus account for the involvement of the mPFC in disparate functions and suggest that valuation emerges naturally from encoded memory representations.

5.1 Introduction

Our rich knowledge of the past allows us to readily make sense of the present. It also facilitates adaptive planning for the future, for example by supporting simulations of prospective events (Barron et al., 2013; Hassabis & Maguire, 2007; Irish et al., 2012; Schacter et al., 2017; Suddendorf & Corballis, 2007). Critically, these capacities are not exclusively dependent on individual memories of unique past experiences. Instead, they are also based on generalized knowledge about our environment that is derived from multiple experiences (e.g., knowledge about relationships between familiar people) (Addis, 2020; Irish et al., 2012).

A type of such generalized knowledge structures are memory schemas (Ghosh & Gilboa, 2014; Rumelhart & Ortony, 1976). These representations of our environment can be understood as graphs comprising information about nodes (e.g., individual people) and their edges (e.g., their relationships) (C. Chen et al., 2021; Ghosh & Gilboa, 2014; Parkinson et al., 2017; Rumelhart & Ortony, 1976). Schemas are formed by extracting commonalities across related events (Ghosh & Gilboa, 2014; Moscovitch et al., 2016). They thereby reduce the complexity of our experience into simplified models of the world (e.g., about the people we know or about the locations we frequently visit) (Ghosh & Gilboa, 2014; Mack et al., 2020). Such models, in turn, foster planning and facilitate adaptive decisions (Behrens et al., 2018; Morton et al., 2017).

However, beyond a representation of the environment's structure, adaptive cognition also requires a representation of what's valuable within that environment (O'Doherty et al., 2017). Here, we test the hypothesis that these two kinds of information are inherently intertwined in the rostral and ventral medial prefrontal cortex (mPFC) (Benoit et al., 2019; Farovik et al., 2015; Roy et al., 2012; Zhou et al., 2019). As detailed below, this proposal accounts for the involvement of this region in two seemingly disparate functions: representing *memory schemas* and *value*.

Evidence from humans (Ghosh et al., 2014; Gilboa & Marlatte, 2017) and rodents (Farovik et al., 2015; Tse et al., 2011) indicates a critical role for the mPFC in mediating memory schemas (Gilboa & Marlatte, 2017; van Kesteren et al., 2012). Activity patterns in this region have been shown to code for individual nodes of the environment, such as for familiar people (Benoit et al., 2019; Thornton & Mitchell, 2017) and places (Benoit et al., 2019; Robin et al., 2018). However, it remains unclear whether the mPFC encodes representations of the nodes in isolation or whether these representations also entail information about their edges (i.e., their relationships to other nodes).

A largely independent line of research has associated the mPFC with the representation of affect and value (Bartra et al., 2013; Roy et al., 2012; Winecoff et al., 2013). Activity in this region tracks the value of objects, people, or places that we currently perceive or imagine (Bartra et al., 2013; Benoit et al., 2014, 2019; Clithero & Rangel, 2014; Lin et al., 2015). Moreover, in humans, focal lesions disrupt value judgements (Camille et al., 2011; Fellows, 2019). The mPFC has thus been argued to represent value in a common currency that allows for flexible decision making in a wide range of contexts (Bartra et al., 2013; Lim et al., 2013).

Notably, evidence from human neuroimaging (Barron et al., 2013; Benoit et al., 2014, 2019; Lin et al., 2015; Shenhav et al., 2013) and rodent single cell-recordings (Farovik et al., 2015; Zhou et al., 2019) has shown that representations of memories and of value are supported by *overlapping* parts of the mPFC. We thus reconcile the common attribution of these functions by hypothesizing that the schemas encoded by this region are shaped by value.

Specifically, we propose that the mPFC encodes representations of individual nodes (e.g., individual familiar people) and that the representations also entail information about their edges (e.g., the overall associations between the people). Critically, we suggest that nodes that are more important for a person exhibit stronger edges.

We hypothesize that the importance of a given node is jointly determined by three features: Given that schemas build up with experience (Ghosh & Gilboa, 2014), we first expect that more *familiar* nodes should be more prominently embedded in the overall graph (Benoit et al., 2014). Secondly, for the same reason, we expect stronger embedding of nodes that are more *central* to the respective environment (Parkinson et al., 2017). Finally, given the role of the mPFC in affect and valuation (Bartra et al., 2013; Roy et al., 2012), we propose that the edges are also weighted by the nodes' *value* (Benoit et al., 2019; Farovik et al., 2015). The encoded schemas would thus emphasize connections of behaviorally relevant elements of the environment, reminiscent of the hippocampal weighting of rewarded locations (Kumaran et al., 2016; Schafer & Schiller, 2018).

Here, we test this hypothesis by probing the neural representations of two distinct and individually unique schemas: about people's social networks and about places from their everyday environment. This allows us to examine whether the suggested coding principles generalize across these individual schemas. Participants provided names of people and places they personally know and arranged these names in circular arenas according to their associations (Kriegeskorte & Mur, 2012). This allowed us to quantify the centrality of each exemplar (e.g., a person) to its respective schema (e.g., the social network). Participants also indicated their familiarity with each person and place (as an index of experience) and their

liking of each of these exemplars (as an index of affective value). In a subsequent session, we measured their brain activity using functional magnetic resonance imaging (fMRI) while they imagined interacting with each person and being at each place. We took the ensuing activity patterns to assess the neural representations of the individual nodes and their edges using representational similarity analysis (RSA) (Kriegeskorte et al., 2008).

First, we hypothesized that the mPFC encodes unique representations of the nodes that get reinstated during mental simulation (Benoit et al., 2019; Robin et al., 2018; Thornton & Mitchell, 2017). We thus predicted similar activity patterns to emerge in the mPFC whenever participants imagine the same person or place. Second, we hypothesized that the structure of neural similarity *across* nodes reflects the structure of their edges. That is, we reasoned that pairs of nodes that are more strongly connected (i.e., that exhibit stronger edges) are encoded by more overlapping neuronal populations (Barron et al., 2013; Garvert et al., 2017; Josselyn & Frankland, 2018; Sawamura et al., 2006). This, in turn, should be reflected in overall greater neural similarity for nodes with particularly strong edges. As a consequence, if more important nodes have stronger edges, they should also exhibit overall greater neural similarity. In addition, we further gauge the regional specificity of such value-weighted schemas to the mPFC. Therefore, we also examine the posterior cingulate cortex and the hippocampus, two regions that have similarly been associated with memory (Andrews-Hanna et al., 2010; Baldassano et al., 2018; Benoit & Schacter, 2015) and valuation (Bartra et al., 2013; Clithero & Rangel, 2014; Grueschow et al., 2015).

5.2 Methods

This study examines the nature of individually unique real-life schemas and their representations in the mPFC. In the following, we describe the experimental procedure designed to assess these representations and our analysis approach.

5.2.1 Experimental design and participants

We recruited 39 right-handed healthy unmedicated adults [sex, 23 females; age, 25.4 ± 2.6 years (mean ± SD)] from the study database of the Max Planck Institute for Human Cognitive and Brain Sciences. All participants had normal or corrected to normal vision, provided written informed consent and received monetary compensation for their participation. The experimental protocol was approved by the local ethics committee (Ethical Committee at the Medical Faculty, Leipzig University, Leipzig, Germany; Proposal number: 310/16-ek). Three participants had to be excluded from analysis either because of a recording error (n = 1), or excessive movement (n = 2). Excessive movement was defined as absolute movement ≥ 3 mm

within either run or a total of ≥ 5 episodes of movement ≥ 0.5 mm. We thus included 36 participants [sex, 22 females; age, 25.2 ± 2.5 years (mean ± SD)] in the analyses.

5.2.2 Tasks and procedures

The procedure, adapted from Benoit et al. (2019), comprised two sessions. During the first session, participants provided names of personally familiar people and of such places. Participants tend to start by listing people and places that they are most familiar with and that they like the most. We therefore asked them to provide us with 90 people and 90 places and then randomly sampled 30 of each to ensure a greater variability in these variables of interest.

Arrangement tasks: Assessing the schema

To quantify the centrality of each node to its schema, participants arranged the names of the people and places on separate two-dimensional circular arenas using the multiple arrangements task (Kriegeskorte & Mur, 2012) (Figure 11A). We instructed participants to position names closer to each other that they also associate more strongly. The inverse of the distance thus serves as a measure of associatedness between any two nodes. We quantified the centrality of each person and place to its schema by computing their centrality, i.e., the sum of their associatedness values.

We then assessed how much experience participants had with each person and place. The participants therefore placed the names on continuous familiarity scales ranging from "not at all familiar" to "very much familiar" (Figure 11A). Finally, participants provided a measure of affective value for each person and place by arranging their names on continuous liking scales ranging from "not at all liked" to "very much liked". All arrangements were done separately for people and places.

Simulation task: Assessing neural representations

The participants returned for a separate session (median delay: 1 day; range: 1–4 days) to complete the episodic simulation task in the fMRI scanner. Each trial of the simulation task began with a fixation period of 0.5 s followed by the name of a person or a place for 7.5 s. During this time, participants imagined interacting with the person in a typical manner or being at the place engaging in a location specific activity. Participants were instructed to imagine the episode as vividly as possible, so that they have a clear mental picture of the respective person or place. Participants then rated the vividness of their imagination on a five-point scale within a maximum of 3 s. Trials for which participants failed to press a button within that time period were later removed from analysis. If there was time left from the response window, it was added

to the subsequent inter-trial interval. This lasted for at least 3 s plus an additional jittered period (0 to 8 s in 2 s intervals). The screen during the inter-trial interval was blank. Each person and place was presented once in a random order in each of the two functional runs. Before entering the scanner, participants practiced the simulation task with people and places that they had previously provided but that did not feature in the simulation task proper.

After the simulation task, participants were presented with people-places and faces-places localizers. Outside the scanner, they provided further information regarding the associations and identities of the individual people and places, including their addresses and locations. They also completed a number of standard questionnaires. These data were not analyzed for the current study.

5.2.3 fMRI data acquisition

Participants were scanned with a 3 Tesla Siemens Magnetom PRISMA MRI scanner with a 32-channel head coil. We acquired anatomical images with a T1-weighted magnetization-prepared rapid gradient-echo sequence (MPRAGE, 256 sagittal slices, TR = 2,300 ms, TE = 2.98 ms, flip angle = 9°, 1 x 1 x 1 mm^3 voxels, FoV = 240 mm by 176 mm, GRAPPA factor = 2). For each of the two functional runs of the simulation task, we acquired 469 volumes of blood-oxygen-level-dependent (BOLD) data with a T2*-weighted echo-planar imaging (EPI) pulse sequence (Feinberg et al., 2010; Moeller et al., 2010). This sequence employed multiband RF pulses with the following parameters: 72 interleaved axial-oblique slices (angled 15° towards coronal from AC-PC), TR = 2,000 ms, TE = 25 ms, flip angle = 90°, 2 x 2 x 2 mm^3 voxels, 6/8 partial Fourier, FoV = 192 mm by 192 mm, MF = 3). The first five volumes of each run were discarded to allow for T1 equilibration effects.

Preprocessing

Data were analyzed using *SPM12* (Penny et al., 2011) (www.fil.ion.ucl.ac.uk/spm) in Matlab (version 9.3). The functional images were corrected for slice acquisition times, realigned, corrected for field distortions, and co-registered with the anatomical scan. We also estimated forward and inverse normalization parameters using DARTEL within *SPM12*. Correction for field distortions was achieved using *FSL topup* (Smith et al., 2004) as implemented in *FSL 5.0* (https://fsl.fmrib.ox.ac.uk/).

General linear model

We then decomposed the variance in the BOLD time-series using a general linear model (GLM) in *SPM12*. Each model included six regressors representing residual movement artifacts, plus

regressors modeling the intercepts of block and session. The additional regressors in the GLM coded for the effects of interest.

Specifically, we modeled each trial as a separate condition yielding a total of 120 regressors – one for each of the two simulations of the 30 people and 30 places. The trial regressors were convolved with the canonical hemodynamic response function. A 1/128-Hz high-pass filter was applied to the data and the model. We computed *t*-maps for the estimated parameters of interest (i.e., for each simulation) against the implicit baseline. The ensuing parameters were used for representational similarity analysis (RSA) (Kriegeskorte et al., 2008; Nili et al., 2014).

Whole brain searchlight analysis: Node coding

To identify brain regions that encode representations of individual people and places (i.e., the nodes of the schemas), we employed an RSA searchlight analysis (spheres with a radius of 8 mm, 4 voxels) across all gray matter voxels. This analysis was based on the RSA toolbox (2014) and compared activity patterns across functional runs (Nili et al., 2020). It identified regions where two simulations of the same person or place yielded more similar activity patterns (*same-item similarity*) than any two simulations of different people or places (*different-item similarity*). Specifically, we assessed *same-item similarity* as the Pearson correlation between the activity pattern of the initial simulation of any given node in the first and its repeated simulation in the second run. *Different-item similarity* was computed as the average correlation of the initial simulation of a node in the first run with all other nodes of the same category (people or places) in the second run. By constraining the *different-item similarity* to items of the same category, we ensure that it is not affected by general differences in the neural representation of people versus places. Finally, we determined the magnitude of the node coding as the difference score between *same-* and *different-item similarity* (Benoit et al., 2019; Nili et al., 2020).

This searchlight analysis yielded a node-coding map for each individual participant. For second level analyses, we *Fisher-z*-transformed these maps, normalized them into MNI space using the DARTEL estimated deformation fields, and smoothed them with a Gaussian Kernel of 8 mm radius at full-width-half-maximum. We then masked the smoothed map with the normalized gray matter masks and tested the significance of the node-coding effect using a simple *t*-contrast at each voxel. We used voxel-level inference at $P < 0.05$ (family-wise-error-corrected) and regarded only clusters that comprised at least 30 contiguous voxels.

ROI-based analyses: Examining the edges

The second RSA examined whether regions that code for the nodes of the schema also code for the predicted relationships between the nodes (i.e., their edges). This analysis thus examined data from regions-of-interest based on the thresholded node-coding map. Note that the two sets of analyses are based on different parts of the neural RSM and on comparisons of model RSMs that are independent from the node-coding model.

For the mPFC, we joined the two rostral and ventral clusters. For the PCC, we took the conjunction of a broad cluster that included this region and an anatomical PCC mask from the Brainnetome atlas (Fan et al., 2016) (regions 175, 176, 181, 182). For the hippocampal ROI, we merged its rostral and caudal parts of the same atlas (regions 215 – 218). Voxels were included if they had at least 50% probability of being part of the mask and gray matter.

As complementary analyses, we also examined the data solely based on an anatomical mask of the ventral mPFC used by Benoit et al. (2019), a more spatially extended mask including the rostral mPFC (comprising Brainnetome regions 13, 14, 41, 42, 47 – 50, 187, 188), and said anatomical mask of the PCC. All masks were inverse normalized into subject space using the DARTEL estimated deformation fields and constrained using the implicit mask estimated from the first level GLMs.

5.2.4 Statistical analysis

Statistical analyses of the data from the ROIs were carried out in R (www.r-project.org). For all *t*-tests reported in the main text and the supplement, we applied an α-level of .05 (two-tailed) and adjusted for multiple comparisons using Holm's method. The linear mixed effects models were set up using LME4. The principal component analysis was computed in Matlab as described below.

Extraction of the importance weights

We had hypothesized that centrality, experience, and affective value would jointly contribute to the importance of a node and expected that they would share a common latent factor. We thus applied principal component analysis (PCA) to the three features and computed the latent factor that explained the most variance. The PCAs were conducted separately for people and places and were based on values of each variable that had been *z*-scored for each participant. This approach ascertained that neither between-category variance nor between-participant variance would bias the factor solution. We then extracted, across all participants, the respective first principal component for people and places. These principal components were positively

correlated not only with centrality and experience but also with affective value, consistent with our proposal that all three contributing features jointly quantify the importance of a given node. We thus refer to these principal components as importance factors.

Predicting the structure of the schemas

We used the importance values to predict the structure of schematic representations in the mPFC. We had hypothesized that more important nodes should, overall, exhibit greater neural similarity with the other nodes. We thus predicted the similarity for any pair of nodes by the product of their respective importance values. We scaled the vectors to the interval of zero (lowest importance) and one (highest importance) prior to multiplication. We then arranged the combined importance values in square matrices for each category (people, places). Note that all analyses are only based on the lower triangular vector of the representational similarity matrices.

Model comparisons using Linear Mixed Models

We set up a series of linear mixed effects models to test which of several alternative predictors accounted best for the structure of representations in each ROI. These models accounted for the neural similarity data as a function of the fixed effects of *category* (people, places), *predictor of interest* (i.e., centrality, experience, affective value, or the principal component), and their interaction. We further accounted for between participant variance by including random effects: one random intercept for participant and run as well as random slopes for *category* and *predictor of interest*. We estimated the models separately for each ROI and subsequently performed model comparisons based on the relative Log Evidence Ratios (LER) derived from Akaike's Information Criterion (Snipes & Taylor, 2014). The best model assumes, by definition, a relative LER of zero, and we regard relative LER differences greater than two as *decisive* evidence for the better model (Kass & Raftery, 1995).

We further examined whether the winning models in each ROI are also substantially superior to models based on random Gaussian noise. We thus created null models by randomly sampling 30 values from a standard normal distribution for both people and places. We then rescaled these values to the interval from zero to one. Subsequently, we constructed a noise null model by computing the product of every combination of two values, just as we had done for our predictors of interest. We also created a second null model by first sorting the same random noise values in descending order prior to multiplication. This was done to account for the inherent order of the original lists of people and places provided by the participants that tended

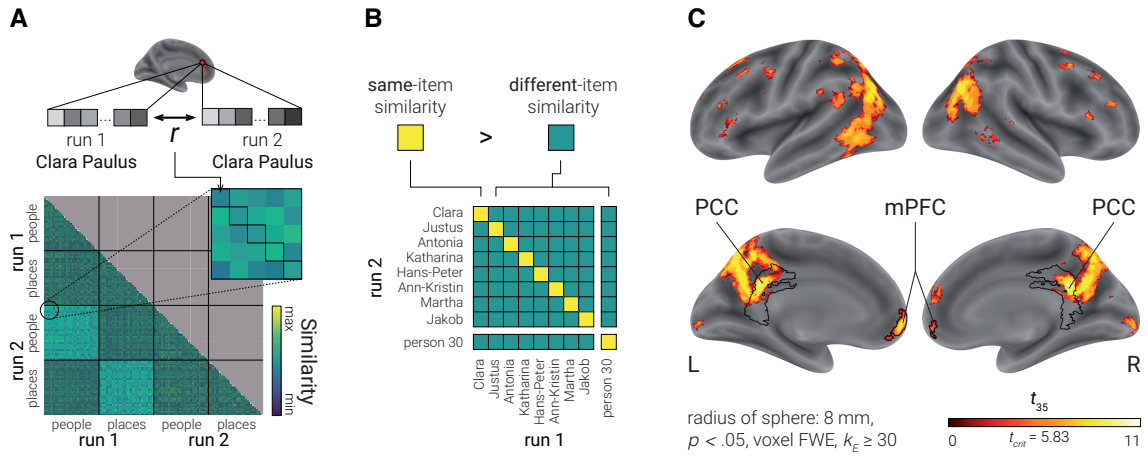


Figure 10. Representations in the mPFC and PCC code for the nodes of individually unique real-life schemas. A. We examined whether the mPFC encodes representations of the nodes by testing for the replicability of activity patterns for the same people and places across the two functional runs. Each row and column of the representational similarity matrix corresponds to a single simulation trial. **B.** Regions coding for the nodes should show more similar activity patterns for the repeated simulations of the same person or place (*same-item similarity*) than for simulations entailing different nodes of the same category (*different-item similarity*). **C.** The searchlight analysis identified regions coding for the nodes of real-life schemas. These entailed the mPFC and PCC. mPFC = medial prefrontal cortex, PCC = posterior cingulate cortex.

to start with more familiar and liked exemplars. Thus, people and places that were named first always received larger random numbers than those named later.

We then fit linear mixed effect models for these two noise null models, and performed a model comparison with the winning model(s) from the respective ROI. We repeated this estimation process 1,000 times to compute average model performance. Critically, if the winning model(s) in each ROI constitute(s) a good approximation of the structure of neural representations, they should consistently outperform both the random noise and the sorted noise models.

5.3 Results

5.3.1 The medial prefrontal cortex encodes the nodes of real-life schemas

We first examined the hypothesis that the medial prefrontal cortex encodes representations of personally familiar people and places, i.e., the nodes of the respective schemas. Whenever we simulate an event involving a particular node, its representation should get reinstated in the mPFC. We thus took the ensuing fMRI activity patterns as proxies of their respective neural representations (Benoit et al., 2019; Charest et al., 2014) and examined their replicability using an RSA searchlight approach (radius = 8 mm, 4 voxels) (Kriegeskorte et al., 2008).

In regions that encode the nodes of the schema, we predicted overall greater pattern similarity for simulations featuring the same node (*same-item similarity*) than for simulations featuring different nodes (*different-item similarity*) (Nili et al., 2020). Note that the *different-item* measure was only based on the similarity of activity patterns for nodes of the same kind

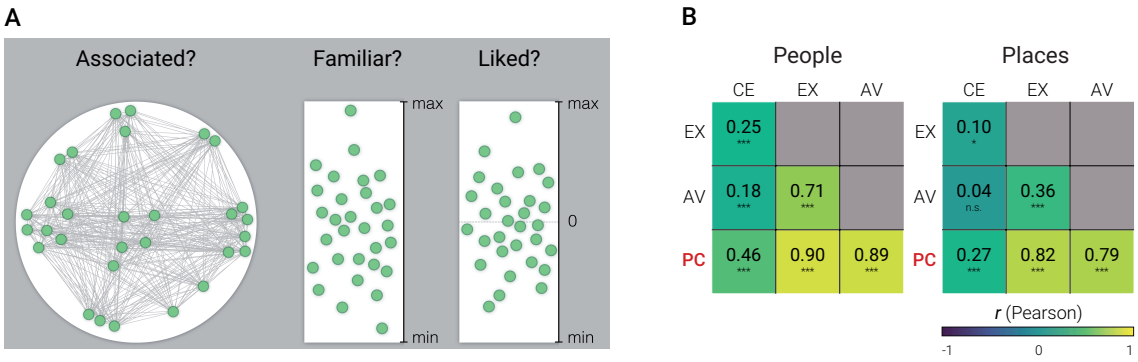


Figure 11. Centrality, experience, and affective value load on a common principal component that quantifies importance. A. Participants arranged the familiar people and places on circular arenas according to their associations, thus providing a measure of centrality. Participants also provided measures of experience and affective value by indicating their familiarity with the people and places as well as their liking. **B.** Centrality, experience, and affective value load on a common principal component as indicated by significant positive correlations (*** - $P_{Holm} < .001$, * - $P_{Holm} < .05$; $df = 35$). This component thus summarizes the importance of the nodes (i.e., the people and places) to the schema. CE = centrality, EX = experience, AV = affective value, PC = principal component.

(i.e., either people *or* places). This ensured that the results are not influenced by potential categorical differences in the representation of people versus places (Figure 10A and B) (Benoit et al., 2019; Charest et al., 2014).

Corroborating our previous finding (Benoit et al., 2019), we obtained this effect in the mPFC. This region thus yielded replicable activity patterns that were specific to individual exemplars (Figure 10C and Supplement C.1). Moreover, we also observed evidence for such replicable pattern reinstatement in a number of other brain regions that are typically engaged during the recollection of past memories and the simulation of prospective events (Benoit & Schacter, 2015; Ritchey et al., 2015; Rugg & Vilberg, 2013). These regions included the posterior cingulate cortex (PCC), the precuneus, and parts of the lateral parietal and temporal cortices. Notably, there was no evidence for pattern reinstatement in the hippocampus.

The data thus support our hypothesis that the mPFC encodes unique representations of individual nodes. In the following, we further examine the edges *between* nodes in the mPFC and PCC regions of interest (ROI) identified by this analysis. We also test for these edges in the hippocampus, even though this region showed no significant evidence of node coding.

Note that the subsequent analyses of the edges are based on different parts of the neural representational similarity matrix (RSM) than the ones used to determine node coding. Further, they are based on comparisons of model RSMs that are also independent of the node coding model. In the supplement, we provide complementary and consistent results based on anatomically defined masks (see Supplements C.2-C.6).

5.3.2 A joint importance factor for centrality, experience, and value

We had hypothesized that the importance of a node is jointly determined by its centrality, experience, and also by its affective value. These three features may thus share a common latent factor. First, to assess *centrality*, participants positioned the names of the people and places in circular arenas (Figure 11A). They were instructed to arrange nodes close together if they associate them strongly with each other and far apart if they do not (Kriegeskorte & Mur, 2012). We calculated the centrality of each node as the sum of its inverse distances to all other nodes. Participants then arranged the people and places on continuous scales providing estimates of their familiarity with each node (as an index of *experience*) and of their liking (as an index of *affective value*). All three features were assessed separately for people and places.

To test whether centrality, experience, and affective value load on a common latent factor, we *z*-scored each vector of values separately for each category (people, places) and within each participant. This approach prevents between-participant variance from influencing the factor solution. We then performed principal component analyses, separately for the people and places. The respective first principal component explained, across all participants, 61% of variance for people and 46% for places.

Critically, as predicted, both of these principal components exhibited significant positive correlations not only with experience and centrality but also with affective value (Figure 11B). We thus take them to quantify the importance of each individual node to its respective schema. In the next step, we used the individual importance values of the respective principal component to predict the structure of the schemas' edges.

5.3.3 The medial prefrontal cortex encodes the edges of value-weighted schemas

We had hypothesized that more important nodes – as indicated by the principal component – exhibit stronger edges. We had further reasoned that the strength of edges is reflected in the neural similarity of the connected nodes. That is, we assumed that more strongly connected nodes are also encoded by more overlapping neuronal populations (Barron et al., 2013; Farovik et al., 2015; Milivojevic et al., 2015). As a consequence, we had predicted that more important nodes should exhibit overall greater neural similarity.

We tested this prediction by constructing models of the expected structure of representations in the mPFC. The models were based on the importance values derived from the respective principal component. Specifically, we predicted the similarity between any two nodes by the product of their respective principal component scores (i.e., importance values).

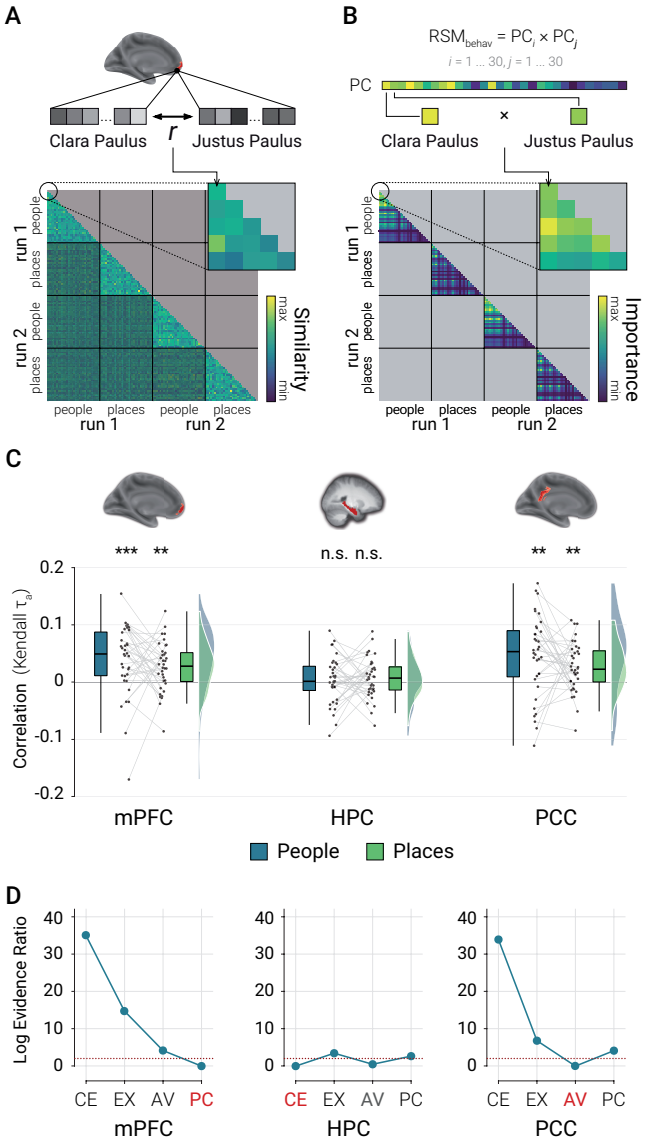


Figure 12. Only the structure of representations in the mPFC is best accounted for by the principal component model that reflects importance. A. Construction of the neural RSM. Each row and column of the matrix corresponds to a single simulation trial. In this analysis, we examine the similarity of activity patterns elicited by simulations of different people or places. **B.** Construction of the model predictions. We predicted more similar representations for people and places with overall higher principal component scores, given that these more important exemplars should be more strongly embedded in their overall schema. To this end, we computed the combined importance of any two people or places from the product of their principal component scores. **C.** Correlation of neural RSM and model prediction. Asterisks denote significant positive correlations as tested in a *t*-test on the *Fisher-z* transformed correlation coefficients (*** - $P_{Holm} < .001$, ** - $P_{Holm} < .01$; $df = 35$). Box-plots: center line, median; box limits, first and third quartile; whiskers, 1.5x interquartile range. **D.** Comparisons of linear mixed models further support the hypothesis: only the structure of representations in the mPFC is best explained by the principal component. The figure displays Log Evidence Ratios (LER). Smaller values indicate better fit. By definition, the best model assumes a value of zero. The dotted red line demarks a relative LER difference of two, regarded as decisive. mPFC = medial prefrontal cortex, HPC = hippocampus, PCC = posterior cingulate cortex, CE = centrality, EX = experience, AV = affective value, PC = principal component.

Thus, we expected more important nodes to yield overall greater pattern similarity (Figure 12B).

We then determined the neural similarity structure in the mPFC, PCC, and in the hippocampus (Figure 12A). We constrained the broader cluster containing the PCC using an anatomical mask of this region (Fan et al., 2016). We used an anatomical mask from the same

atlas to examine the representational structure for the hippocampus. All analyses were conducted in subject space.

Finally, we tested for the correspondence between our prediction and the actual structure of neural representations by computing the correlation of the respective parts of the lower triangular vectors of both matrices (Figure 12C). This was done separately for people and places to examine whether the effect is present for either category. Using Kendall's τ_a as a conservative estimate (Nili et al., 2014), we indeed observed significant correlations in the mPFC for both people (mean $\tau_a = 0.039$, tested with a Wilcoxon test, $W = 562$, $P_{Holm} < .001$, $d = 0.63$, due to a deviation from normality indicated by a Shapiro-Wilk test, $W = 0.92$, $P = 0.01$) and places (mean $\tau_a = 0.026$, $t_{35} = 3.72$, $P_{Holm} = .001$, $d = 0.62$) - with no significant differences between the two (mean difference = 0.013, $t_{35} = 1.04$, $P_{Holm} = .307$, $d = 0.17$).

Similarly, the correlations were also significant in the PCC for people (mean $\tau_a = 0.046$, $t_{35} = 3.91$, $P_{Holm} = .001$, $d = 0.65$) and places (mean $\tau_a = 0.026$, $t_{35} = 3.61$, $P_{Holm} = .002$, $d = 0.60$), again with no significant differences between the two (mean difference = 0.020, $t_{35} = 1.42$, $P_{Holm} = .165$, $d = 0.24$). However, the same analyses of the hippocampal data did not yield evidence for a match between the predicted and actual structure of representations (people: mean $\tau_a = 0.003$, $t_{35} = 0.41$, $P_{Holm} = 1$, $d = 0.07$; places: mean $\tau_a = 0.007$, $t_{35} = 1.15$, $P_{Holm} = .778$, $d = 0.19$; people vs. places: mean difference = -0.004, $t_{35} = -0.41$, $P_{Holm} = 1$, $d = -0.07$).

We obtained qualitatively identical results in analyses based on purely anatomically defined ROIs (see Supplements C.2, C.3, and C.5). The results are also in accordance with a whole-brain searchlight analysis (radius = 8 mm, 4 voxels) (see Supplement C.7). We thus show that representations in the mPFC generally align with the predicted structure of value-weighted schematic representations. However, it remains to be determined whether importance is indeed the best model to account for the structure of representations in any of our ROIs.

5.3.4 The importance model accounts best for the structure of mPFC representations

Does the structure of representations predicted from the principal component account best for the data or would any of the individual contributing features provide at least a comparable fit? If the mPFC does encode value-weighted schemas, we would expect the model based on the conglomerate index of importance to outperform models only based on centrality, experience, or affective value. Furthermore, we would expect some degree of regional specificity, i.e., that only representations in the mPFC, but not in the control regions, are best accounted for by importance.

We formally tested these predictions by setting up alternative models that were solely based on either centrality, experience, or affective value. We then compared these models with the importance model that was based on the principal component. In brief, we set up linear mixed effects models to account for the structure of representations as a function of each of these individual features.

In each of these models, we included a factor of category (people, places) and the maximum possible random effect structure that would converge across all models and regions of interest. We thus accounted for between-participant variance by including a random intercept per participant and run, as well as random slopes for our fixed effects of category and the respective predictor (e.g., the principal component scores). We then performed model comparisons within each ROI to determine the model that best fits the neural similarity structure. The comparisons were based on Log Evidence Ratios (LER) derived from Akaike's Information Criterion (Snipes & Taylor, 2014). We regarded LER differences greater than two as decisive evidence for the better model (Kass & Raftery, 1995).

Consistent with our hypothesis, in the mPFC, the principal component model accounted best for the data. It performed decisively better than affective value (LER = 4.19), experience (LER = 14.8) and centrality (LER = 35.16) (Figure 12D). The model parameters of this winning model entailed a significant main effect of category, reflecting overall higher neural pattern similarity for people than places ($\beta_{Category\_place}$ = -0.026, SE = 0.008, χ^2 = 11.61, $P < .001$). Critically, they also included a significant positive parameter estimate for the principal component, indicating overall greater neural pattern similarity for nodes of greater importance ($\beta_{PrincipalComponent}$ = 0.048, SE = 0.012, χ^2 = 17.12, $P < .001$). Moreover, the main effect of the principal component did not interact with category ($\beta_{Category\_place:PrincipalComponent}$ = -0.005, SE = 0.008, χ^2= 0.36, P = .546).

By contrast, in both control regions, other models were better suited to account for the structure of representations. For the hippocampus, the model comparison yielded the best fit for centrality, though there was only a minimal advantage for this model over affective value (*LER* = 0.49). Notably, both models performed decisively better than the ones based on either the principal component (LER = 2.68) or experience (LER = 3.46). However, of the model parameters, only the main effect of category was significant, indicating overall higher pattern similarity for places than for people ($\beta_{Category\_place}$ = 0.02, SE = 0.004, χ^2 = 31.49, $P < .001$). There was neither a significant main effect of centrality ($\beta_{Centrality}$ = -0.007, SE = 0.004, χ^2 = 0.9, P = .342) nor a significant interaction of category with centrality ($\beta_{Category\_place:Centrality}$ = 0.007, SE

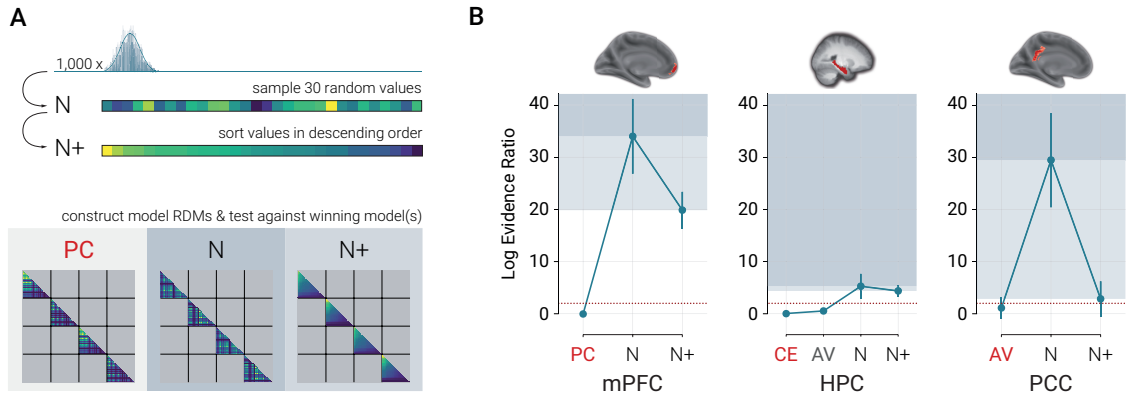


Figure 13. The importance model outperforms noise models in the mPFC only. A. Construction of the noise models. On 1,000 iterations, we sampled 30 values for each category type from a standard normal distribution to create a vector of noise values. We then used these vectors to construct a random noise model (N) by performing the same processing steps as for the other models. This approach also allowed us to create a sorted noise model (N+). Here, we first sorted the vector of noise values in descending order. This model mirrors the order in which participants tend to list people and places, i.e., by starting with ones that are more familiar and liked. **B.** Comparisons with noise null models. Points depict the mean model performance across comparisons with 1,000 random noise models (N) and sorted noise models (N+), whiskers indicate the standard deviation. Smaller values indicate better fit. The dotted red line demarks a relative LER difference of two, regarded as decisive. mPFC = medial prefrontal cortex, HPC = hippocampus, PCC = posterior cingulate cortex, CE = centrality, EX = experience, AV = affective value, PC = principal component, N = random noise, N+ = sorted noise.

= 0.004, χ^2 = 3.04, *P* = .081). The same pattern (i.e., only a main effect of category) also emerged for the model based on affective value (see Supplement C.4).

For the PCC, the model based on affective value performed decisively better than any other model: principal component (LER = 4.14), experience (LER = 6.82), and centrality (LER = 33.99) (see Figure 12D). The main effect of affective value was significant, indicating overall greater neural pattern similarity for nodes of higher affective value ($\beta_{AffectiveValue}$ = 0.026, SE = 0.007, χ^2 = 10.71, *P* = .001). There was no main effect of category ($\beta_{Category\_place}$ = 0.014, SE = 0.009, χ^2 = 1.51, *P* = .219), but an interaction of affective value with category, reflecting a stronger effect of affective value for people than places ($\beta_{Category\_place:AffectiveValue}$ = -0.01, SE = 0.005, χ^2 = 3.92, *P* = .048) (see Supplement C.4 for all model parameters).

In summary, the model based on the principal component was the clear winner in the mPFC, whereas it was outperformed by alternative models in the control regions. This pattern thus suggests some regional specificity of value-weighted schemas. Note that we obtained consistent results when examining the structure of representations in the purely anatomically defined ROIs (see Supplement C.4 and C.6).

Finally, we sought to ensure that the winning models in each ROI perform better than null models based on noise. To this end, for each familiar person and place, we randomly sampled a value from a standard normal distribution. We then used these values to construct a noise model by performing the identical processing steps as for the other predictors.

Moreover, this allowed us to derive a second noise model by first sorting the vector of noise values in descending order prior to constructing the model. This model mirrors the order in which participants tend to list people and places, i.e., by starting with people and places they like and know better. As a consequence, nodes that are listed earlier tend to also have higher values on the principal component. By sorting the noise vectors in descending order, we imposed a similar dependence between the noise values and their serial positions (Figure 13A).

Separately for each ROI, we then compared model performance of random noise and sorted noise against the winning model. We repeated this process 1,000 times to obtain an estimate of the expected performance of the noise models and the winning model. As expected, in the mPFC, the principal component remained the best model (mean LER = 0, SD = 0.04), performing decisively better than sorted noise (mean LER = 15.56, SD = 2.51) and random noise (mean LER = 25.81, SD = 6.03).

For the hippocampus, the initial model comparisons had provided only minimal evidence for centrality over affective value. We therefore compared both of these models with the noise models. Again, there was minimal evidence for superiority of centrality (mean LER = 0.09, SD = 0.53) over affective value (mean LER = 0.58, SD = 0.53). Both models performed decisively better than sorted noise (mean LER = 4.4, SD = 1.09) and random noise (mean LER = 5.32, SD = 2.38). The model comparisons in the PCC revealed strong, though not decisive, evidence for a superiority of affective value (mean LER = 1.15, SD = 2.1) over sorted noise (mean LER = 2.89, SD = 3.37). However, both did perform decisively better than random noise (mean LER = 29.46, SD = 9.01) (see Figure 13B).

The results thus provide further evidence for the hypothesis that the mPFC encodes both the nodes and the edges of value-weighted schematic representations. The model comparison moreover supports this account with some regional specificity.

5.4 Discussion

Human adaptive cognition is fostered by representations of the structure of our environment (Behrens et al., 2018; Morton et al., 2017). Such structured representations act as templates that allow us to facilitate recollections of the past, to make sense of the present, and to flexibly anticipate the future (Addis, 2020; Benoit et al., 2014; Ghosh & Gilboa, 2014; Moscovitch et al., 2016). Structured representations have been described in the mPFC for various domains, ranging from spatial and conceptual to abstract state spaces (Constantinescu et al., 2016; Doeller et al., 2010; Schapiro et al., 2013; Schuck et al., 2016; Zhou et al., 2019).

Our results support the hypothesis that the mPFC supports a specific form of such structured representations: value-weighted schemas of our environment. Generally, the mPFC

has long been argued to mediate memory schemas (Gilboa & Marlatte, 2017; van Kesteren et al., 2012), yet the exact contribution of this region has remained unclear. It has been suggested that the mPFC serves to detect congruency of incoming information with schematic knowledge that is represented in posterior areas (van Kesteren et al., 2012). This region would thus not necessarily represent any kind of schematic knowledge by itself. Our data indicate that the contribution of the mPFC goes beyond congruency detection: It directly encodes schematic representations of the environment (see also Benoit et al., 2019; Parkinson et al., 2017).

These representations could act as pointer functions that guide the reinstatement of relevant distributed information (Ciaramelli et al., 2019; Gilboa & Marlatte, 2017; Gilboa & Moscovitch, 2017; van Kesteren et al., 2012). This suggestion fits with broader accounts that situate the mPFC on top of a cortical hierarchy as a convergence zone (Andrews-Hanna et al., 2010; Margulies et al., 2016) that integrates information from diverse brain networks (Benoit et al., 2014; Ritchey et al., 2015).

Critically, our results support the hypothesis that schematic representations in the mPFC (e.g., of one's social network) inherently entail the value of the encoded nodes (e.g., how much we like individual people). That is, the structure of the edges could best be accounted for by a model based on a latent factor that quantifies the importance of the nodes. As predicted, this factor was not only influenced by the nodes' centrality (Parkinson et al., 2017) and familiarity (Benoit et al., 2014, 2019; Robin et al., 2018), but also by their value (Benoit et al., 2019; Farovik et al., 2015; Roy et al., 2012). We obtained this pattern across schemas for personally familiar people and places. The convergent results thus demonstrate that this coding scheme in the mPFC generalizes to different kinds of environmental representations.

The model comparison also suggested some degree of regional specificity for value-weighted schemas. The importance model was neither the best fit to the data obtained from the PCC nor from the hippocampus. Whereas even the best model did not decisively outperform a noise model in the PCC, there was some evidence that the structure of the edges in the hippocampus could best be accounted for by either centrality or affective value. These results are consistent with evidence showing that the hippocampus encodes map-like representations of relational abstract (Garvert et al., 2017) and social (Tavares et al., 2015) knowledge and that it is involved in value learning (Stachenfeld et al., 2017; Wimmer & Shohamy, 2012).

More broadly, a functional dissociation between the hippocampus and mPFC is also consistent with the suggested involvement of these regions in two complementary learning systems. Whereas the hippocampus is critical for the retention of individual episodes, the mPFC may extract commonalities across similar events and bind these into consolidated

representations (Kumaran et al., 2016; Morton et al., 2017; Moscovitch et al., 2016; cf. Schapiro et al., 2017). The mPFC would thus reduce the complexity of our experience into schematic summary representations.

Indeed, a recent study provided convergent evidence for such dimension reduction in this region. It demonstrated that the mPFC compresses rich perceptual input to only those features that are currently task-relevant – akin to a principal component analysis (Mack et al., 2020). While such dimension reduction entails the loss of specific details, it also affords generalizability and cognitive flexibility (Bowman & Zeithamova, 2018; Ghosh & Gilboa, 2014). These representations can thus augment planning (Addis, 2020; Momennejad, 2020) and also be flexibly used for the construction and valuation of novel events (Barron et al., 2013; Benoit et al., 2014).

The emergence of schemas in the mPFC could be fostered by hippocampal replay of past events (Carr et al., 2011; Michelmann et al., 2019). Such replay, conveyed by monosynaptic efferent projections into the mPFC (Eichenbaum, 2017), can potentially provide a teaching signal that facilitates neocortical consolidation (Kumaran et al., 2016). Moreover, to the degree that replay is biased towards valuable information, it may lead to a stronger weighting of those experiences that are of particular importance (Kumaran et al., 2016; Schafer & Schiller, 2018). However, the mPFC likewise receives direct projections from areas such as the amygdala and the striatum (Price & Drevets, 2010) that could also contribute to a shaping of the schematic representations by value (see also Roy et al., 2012).

Importantly, the highlighted structure of representations in the mPFC provides a common account for the involvement of this region in both memory schemas and valuation. That is, when we think about an individual element from our environment (e.g., a known person), its representation in the mPFC is activated. This activation then spreads throughout the network of connected nodes. Critically, we suggest that there is a wider spread from nodes that are more valuable and that are thus more strongly embedded in their overarching schema. This wider spread, in turn, may manifest as greater regional univariate activity. According to this account, the valuation signal that has been attributed to the mPFC (Bartra et al., 2013; Clithero & Rangel, 2014) thus constitutes an emergent property of the structure of its encoded representations.

This interpretation similarly accounts for the stronger engagement of the mPFC when individuals think about themselves as compared to others (Overwalle, 2009). The self can be considered a super-ordinate schema that entails abstracted representations of all our personal experiences (Gilboa & Moscovitch, 2017). Instantiating this schema would thus presumably

lead to wide spread activity, whereas thinking about specific other people would only co-activate neural representations of more restricted nodes. Moreover, the net activity would be lower for other people that we feel less connected to and that we have less experience with (Benoit et al., 2010, 2014; Mitchell et al., 2006; Rameson et al., 2010).

To conclude, this study provides evidence that the medial prefrontal cortex represents the structure of our environment in the form of value-weighted schemas. These schemas reflect our experience with individual nodes as well as their centrality. Critically, they also inherently encode information about their affective value. These schematic representations thus prioritize information that is critical for adaptive planning and that ultimately promotes our well-being and survival.

Acknowledgments

We thank Roxanne Eisenbeis for assistance in data collection, Ruud Berkers for help in setting up preprocessing, Davide Stramaccia for discussing the linear mixed effects models, and Angharad Williams for comments on an earlier draft of this manuscript.

Funding

This work was supported by a Max Planck Research Group (awarded to Roland G. Benoit).

Author contributions

Conceptualization: PCP, RGB; *Methodology:* PCP, IC, RGB; *Investigation:* PCP; *Formal Analysis:* PCP, IC, RGB; *Visualization:* PCP; *Funding Acquisition:* RGB; *Supervision:* RGB; *Writing – Original Draft:* PCP, RGB; *Writing – Review & Editing:* PCP, IC, RGB

[PCP = Philipp C. Paulus, IC = Ian Charest, RGB = Roland G. Benoit]

Competing interests

The authors declare that they have no competing interests.

Data and materials availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Because participants did not give consent for their MRI data to be released publicly within the General Data Protection Regulation 2016/679 of the EU, we can only share data with individual researchers upon reasonable request. The second level *t*-map for the node coding searchlight analysis (Figure 10C) and the functional masks of the mPFC and PCC ROI are available at neurovault: https://identifiers.org/neurovault.collection:8129

A full R Markdown of the analyses and custom code is publicly available via the Open Science Framework: https://osf.io/6h58e/?view_only=b21fa36180a845e69671f222a110bac8

Chapter 6

6 General discussion

The human ability for mental time travel allows us to mentally project ourselves back and forth in time (Suddendorf & Corballis, 2007). Episodic simulation enables us to simulate episodes that could have taken place in the past, may take place in the present, or could happen in the future (Schacter et al., 2007). This ability has great adaptive value and allows us to pre-experience the affective consequences of potential experiences without having to actually make them (Gilbert & Wilson, 2007). Remembering the past and simulating the future share many phenomenological similarities (D'Argembeau & Van der Linden, 2004, 2006) and are supported by the same network of brain regions including the mPFC, the PCC, and the hippocampus (Benoit & Schacter, 2015; Schacter et al., 2017; Stawarczyk & D'Argembeau, 2015). Interestingly, this network for episodic memory and simulation partially overlaps with a neural network for value-based decision making (Bartra et al., 2013; Clithero & Rangel, 2014).

Using episodic simulation as one instantiation of our ability for mental time travel, I examined how mnemonic and evaluative processes interact to support adaptive behavior. To this end, I have taken two complementary perspectives: (i) Across two related projects, I have investigated a mechanism that allows us to learn from merely imagined experiences much as we learn from actual past experience. (ii) I have investigated the structure of neural memory representations that are activated whenever we imagine hypothetical events that could take place in the future. More specifically, I have provided evidence that these memory representations are not only shaped by the structure of our environment and experience, but critically also reflect value. After a summary of the results of each individual experiment, this final section will discuss the broader implications of the reported findings and highlight potential avenues for future work.

Study 1 examined a neural mechanism by which merely imagined events can shape real-life attitudes. Participants imagined encountering personally familiar people at locations they know from their everyday life. These people were either liked or disliked and all locations were initially rated as neutral. Following the simulations participants rated all locations as more liked than before. Critically, in two analogous fMRI and behavioral experiments the increase in liking was significantly larger for those locations that had set the stage for encounters with liked people. The results thus provide evidence for a transfer of positive affective valence from the person toward the location.

Moreover, the neuroimaging results further demonstrate a key involvement of the mPFC in mediating this simulation-induced attitude change: Univariate activation in the mPFC reflected the liking of the simulated person and predicted the subsequent change in attitude

toward the location. Multivariate analyses using RSA revealed that the mPFC encodes unique representations of known people and locations. Thus, whenever participants simulated the same person or location a more similar activation pattern emerged in the mPFC as compared to representations that emerged in simulations of different people or locations. These findings suggest that the mPFC encodes schematic representations of our environment and simultaneously encodes a value signal that scales with the liking of the simulated people. Critically, this value signal encoded in the mPFC is predictive of the subsequently observed attitude change toward the locations.

Study 2 is a behavioral study that that aimed at clarifying the mechanisms that support simulation-induced attitude changes. The previous study had provided evidence that simulations can induce a transfer of *positive* affective valence. Here, we also included a neutral baseline condition to examine whether simulations can also induce a *true* transfer of affective valence, i.e., whether simulations can induce both a transfer of *positive* as well as of *negative* affective valence. Additionally, participants repeatedly arranged the people and locations on two-dimensional surfaces to indicate how they associate them. The results of these arrangements demonstrate that simulations cause an integration of the memory representations of the jointly simulated people and locations. The behavioral results on the attitude changes toward the locations revealed a true transfer of affective valence: Compared to a neutral baseline condition, simulations featuring disliked people induced a transfer of *negative* valence and simulations with liked people induced a transfer of *positive* valence to their paired locations. Analyses of participants' emotional responses revealed stronger emotional arousal during simulations with disliked and liked people compared to the neutral baseline condition. Moreover, participants rated the valence of scenarios with disliked people as unpleasant, scenarios with neutral people as neutral, and scenarios with liked people as pleasant. A causal mediation analysis revealed that the transfer of valence from the person to the paired location was mediated via this perceived pleasantness of the scenario. Thus, contingent on the valence of the person, episodic simulations elicited an affective experience that induced a transfer of affective valence. In sum, merely imagined experiences induce affective states that shape real-life attitudes much like actual experiences can.

In the fMRI study 3, we examined the structure of neural representations in the mPFC more closely. The study was based on the observation that the mPFC is involved both in the representation of memory schemas and in the computation of a domain general value signal. In this study, we hypothesized that the mPFC might simultaneously subserve these seemingly disparate functions by encoding a specific form of memory representation: value-weighted

schemas. Participants imagined typical scenarios of interacting with known people or being at personally familiar locations. They thereby reinstated the neural memory representation of the individual people and locations they simulated. The ensuing multi-voxel activation pattern measured via fMRI served as a proxy measure for this neural memory representation. Based on fine-grained behavioral assessments, we demonstrated that the structure of memory representations in the mPFC reflects a combination of how *central* a given exemplar is to the respective environment (e.g., our social network), the amount of *experience* we have with it, and, critically, how much we *like* it. Thus, people that are central to our social network, that we know well, and like a lot were overall also more strongly embedded in the neural memory representation in the mPFC. Critically, only the structure of neural memory representations in the mPFC matched a structure predicted from the combination of centrality, experience, and value for both known people and locations. In sum, the findings of this study suggest that the mPFC encodes generalized knowledge about our environment in schematic representations. These schemas are also shaped by the value of the encoded exemplars. Such value-weighted schematic representations may provide an account for the overlapping involvement of the mPFC in both mnemonic functions and valuation.

Across the neuroimaging projects of this thesis, the results demonstrate a central role of the mPFC for episodic simulation. These findings extend previously provided evidence that the mPFC simultaneously supports mnemonic processes and valuation (Benoit et al., 2014; Lin et al., 2015, 2016). Specifically, the results suggest that the mPFC encodes schematic memory representations within which representations of individual exemplars (i.e., known people or locations) are closely intertwined with a representation of their value. Thus, these representations not only encode the typical structure of our environment, they also reflect the values of encoded exemplars. These value-weighted schematic representations, in turn, might support a simulation-based learning mechanism that shapes real-life attitudes and may account for the critical contribution of the mPFC to flexible, adaptive behavior more broadly.

In the following sections, I will first make a case that simulations can serve as an imaginary parallel to actual experience. In a second part, I will then discuss the neural findings with a particular focus on one type of structured representation of our environment: value-weighted schematic representations. Specifically, I will highlight how our results support the view that our memories of the past serve us in adaptive ways that are oriented toward the future. Finally, I will indicate the broader implications of the reported findings and argue that more naturalistic and life-like research protocols are required to investigate the brain mechanisms that support the human ability for adaptive and flexible behavior.

6.1 Simulation-based learning of real-life attitudes

How we perceive our environment and the people that live in it is shaped by our memories of past events. Across unique individual experiences we gradually learn what our environment is typically like (Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017). Locations where we have previously had a good time are probably worth visiting again and people that we enjoyed meeting should be the ones we want to spend more time with in the future (Montague et al., 2006; Rangel et al., 2008). The results of this thesis have demonstrated that it is not always necessary to actually experience these events: We can also learn and form preferences from purely imagined experience.

This simulation-based learning mechanism is well in line with the related literature on attitude change resulting from actual experiences. Preferences for initially neutral stimuli can be acquired from associatively paired rewarded stimuli (Wimmer & Shohamy, 2012). Similarly, evaluative conditioning describes a mechanism by which neutral stimuli acquire affective valence from a simultaneously presented liked or disliked stimulus (Hofmann et al., 2010; Jones et al., 2010). The results reported in this thesis extend previous research by demonstrating that simulations can not only induce attitudes for arbitrary neutral stimuli. Instead, the results demonstrate that simulations can also shape pre-existing attitudes toward personally known locations from participants real-life environment. This is particularly remarkable as changing existing attitudes has been shown to be more difficult than forming them in the first place (Jones et al., 2010).

Simulations can even serve as a replacement for actual experiences (Kappes & Morewedge, 2016). In classical conditioning, a simulated aversive experience (stepping onto a thumbtack) can serve as a replacement for the actual physical presentation of unpleasant stimuli and can cause de-novo fear conditioning (Mueller et al., 2019). Simulated success can impede actual achievements. When individuals imagine mastering a complex task easily, they subsequently perform worse than individuals who imagine a soothing scenario or overcoming difficulties in the future (Kappes & Oettingen, 2011; Spencer & Norem, 1996). Relatedly, repeated mental imagery of motor movements yields performance benefits on a motor task that mirror the effects of actual training (Driskell et al., 1994). Imaginary consumption of food items to satiety reduces subsequent consumption of those food items, whereas imagining to taste only some of the food increases subsequent consumption (Morewedge et al., 2010). The results of the present thesis extend this body of research by demonstrating that simulations can also shape attitudes toward the very elements that those simulations had been based on. By this,

simulations can provide an imaginary parallel to actual experience and in some circumstances might even substitute actual experience.

However, for simulations to serve as actual replacements of experience, they would have to have a reliable and lasting effect. Across the studies reported in this thesis that investigated simulation-based learning the results demonstrate that simulations can shape real-life attitudes. As it stands, we can thus assert with some confidence that these effects are present immediately after participants simulated the episodes. But how long do these effects last for? A number of related findings suggest that the effects of simulation-based learning may be long-lasting. In the related literature on evaluative conditioning, a neutral stimulus is presented together with a valenced stimulus. The typical result pattern shows a change in attitude toward the neutral stimulus in the direction of the valenced stimulus. The subsequent changes in the evaluation of the neutral stimulus have been shown to be highly stable over time and only partly subject to extinction (Baeyens et al., 1988, 2005; Hofmann et al., 2010; but see Lipp et al., 2003). The described simulation-based learning mechanism may also be viewed as a special case of episodic associative learning. Compared to neutral material, emotional memories of highly arousing material have been demonstrated to be highly stable over time (LaBar & Cabeza, 2006). This might indicate that simulation-induced learning based on affective experiences might be particularly stable over time. However, there are also some indications that the temporal stability might be mediated by the direction of the transfer of valence: It has been demonstrated that details of unpleasant episodic simulations are more difficult to remember than neutral and pleasant simulations after a delay of one day (Szpunar et al., 2012). In sum, it is at least plausible that the observed effects might be stable over time, notwithstanding this, a formal investigation is required.

An investigation of the temporal stability of the observed simulation-based learning effects might shed more light on the precise mechanisms that cause the transfer of valence between the memory representations. The results of the fMRI study on simulation-based learning yielded no evidence for a reconfiguration of individual exemplars' *neural* memory representations following the joint episodic simulations: There was no effect of valence on the replicability of the activity patterns in the mPFC (see Benoit et al., 2019). On the one hand, this absence of evidence can be accounted for by the fact that three joint simulations are unlikely to cause drastic changes in pre-existing, consolidated, and generalized schematic memory representations of known people and locations. Moreover, fMRI might not provide sufficient spatial and temporal resolution to detect such subtle changes. On the other hand, it might also be the case that such changes would only be apparent at a later point in time. The simulated

scenario might itself initially be encoded as an episodic memory in the hippocampus (Frankland & Bontempi, 2005; Szpunar et al., 2012). It has been shown that episodic memories are consolidated both during subsequent wakefulness (Carr et al., 2011) and sleep (Inostroza & Born, 2013; Lewis & Durrant, 2011) via hippocampal replay (Buzsáki, 1996; Michelmann et al., 2019), i.e., additional offline practice. Such replay sequences induce plasticity in cortical memory repositories and ultimately lead to long-term consolidation (Frankland & Bontempi, 2005). Thus, reactivation and stabilization of initially hippocampally-dependent memory traces might be required to foster changes in the cortical representations. Changes in schematic representations that are encoded in the mPFC would then only be visible at a later point in time.

Formal theories of associative learning assume that learning is an iterative process and that repeated exposures with the same contingencies induce stronger learning (Gershman, 2015; Rescorla & Wagner, 1972). In the reported studies, participants repeatedly simulated interacting with the known people at the locations. These simulations strengthened associations between the simulated people and locations and induced a transfer of affective valence (see Benoit et al., 2019; Paulus et al., 2021). However, the number of simulations was kept constant across the studies. Thus, to further establish a causal link between the simulated scenarios and the observed changes a formal investigation of the effect of repetition is required. If simulations shape attitudes like real experiences (Madan & Kensinger, 2021), the number of simulations and the magnitude of the observed attitude change should exhibit a dose-response relationship with more simulations inducing stronger learning. Testing for such a relationship could be achieved by having participants simulate scenarios more or less often. Alternatively, individual locations might be paired with more than one liked or disliked person. The observed change should then be modulated by the number of simulations or paired people.

The previous sections have discussed the adaptive benefits of a learning mechanism that is based on episodic simulations. However, when going awry, this mechanism can also be maladaptive. Extensive negative future-directed cognitions such as worrying and rumination are key mediating variables in both development and maintenance of psychological disorders (Beck et al., 1987; Holmes et al., 2011; Miloyan et al., 2014). Worrying and rumination are hallmark symptoms of depression and anxiety disorders (Clark & Wells, 1995; Ehlers & Clark, 2000; Miloyan et al., 2014). The results reported in this thesis indicate that simulations can not only induce positive shifts in attitude, but also yield transfers of negative affective valence (see Paulus et al., 2021). In circumstances where due to a generally negative outlook toward the future, individuals may be prone to produce largely unpleasant future imaginations with

undesirable outcomes, it is therefore possible that simulation-based learning might contribute to the development and maintenance of psychological disorders.

At the same time simulation-based learning might be a critical mechanism in imaginal exposure therapy which has been proven to be an effective treatment for affective psychological disorders. Imaginal exposure techniques are therapeutic interventions that confront patients with feared and otherwise avoided stimuli or situations before their mind's eye (Teismann & Margraf, 2018). Like in vivo exposure, imaginal exposure elicits the associated psychological symptoms, bodily responses, as well as the typical pattern of habituation to the feared stimulus (Birbaumer, 1977; Foa & Hearst-Ikeda, 1996). Imaginal exposure is the recommended standard technique in circumstances where in vivo exposure is not possible or patients need to confront feared thoughts or memories (Bryant et al., 2003). Recent years have seen a growing interest of extending these techniques with methods derived from research on episodic simulation (Holmes et al., 2011; Holmes & Mathews, 2010; Renner et al., 2014; Simplicio et al., 2016). However, while episodic simulation and imagination-based interventions are readily applied in psychotherapy and their efficacy is well established, the precise mechanisms by which they exert effects in behavioral change are still evasive. The described neural, evaluative, and mnemonic processes of simulation-based learning provide novel insights into the mechanisms that might be at the root of the efficacy of these interventions.

In sum, the results of the studies on simulation-based learning have provided evidence for the existence of a learning mechanism by which purely imagined experiences can shape attitudes toward the very elements that these simulations had been based on. By this, the results provide support for the central argument of the constructive episodic simulation hypothesis: We can adaptively use our memories of the past to construct mental simulations of potential happenings. Evaluative processes enable us to pre-experience the affective consequences of these happenings already in the here and now. This affective experience, in turn, can shape attitudes both in the *positive* as well as in the *negative* direction. Thus, in some circumstances simulations may serve as a replacement for actual experience.

6.2 Schema, valuation, and the mPFC

Throughout the thesis, I have used the term schema to describe structured and generalized knowledge that is flexibly used whenever individuals simulate potential episodes. But what is a memory schema? This debate has been intertwined with the term schema ever since it was coined. As early as 1932 Bartlett admitted that the term 'schema' "is at once too definite and too sketchy", but in lack of a better word "continue[s] to use the term 'schema' when it seems best to do so" (Bartlett, 1932, p. 201). Based on this tradition, more recent research has used

the term ‘schema’ to describe disparate cognitive functions and neural mechanisms: Pairs of items that can be used for associative inference (Zeithamova et al., 2012), learned spatial layouts that enable rodents to quickly learn new odor-location mappings (Tse et al., 2007, 2011), and pairs of spatially arranged stimuli that are predictive of events (Kumaran et al., 2009) have all been referred to as schemas. In other work, schemas have even been defined as any form of prior knowledge that supports the encoding of new information and allows for new inferences (Preston & Eichenbaum, 2013).

Across the reported projects, I have examined two exemplary manifestations of participants’ real-life memory representations: Knowledge about their social network and about locations from their immediate environment. Ghosh and Gilboa (2014) refer to schemas as “adaptable associative networks of knowledge extracted over multiple similar experiences” (Ghosh et al., 2014, p. 12057). Within this framework, representations of participants’ social network and their immediate environment conform with the necessary features of schemas (see Ghosh & Gilboa, 2014 for an extended discussion): both have an associative network structure, are based on multiple episodes, are characterized by abstraction as well as loss of unit detail, and are adaptable.

But how do value-weighted schemas differ from related representations for structured knowledge such as the successor representation (Garvert et al., 2017; Momennejad et al., 2017) or mental maps (Behrens et al., 2018; Bellmund et al., 2018; Schuck et al., 2016)? Formalizations based on the mental map hypothesis (Tolman, 1948), including the successor representation, suggest a compositional coding format where information about the structure of the environment is encoded separately from representations of the individual objects or elements within that environment (Whittington et al., 2019, 2022). The main advantage of such representations is their flexibility: it is not necessary to re-learn the spatial layout as soon as the elements that populate that environment are changed. Similarly, representations of the objects do not have to be re-learned in a new environment. These formal ideas account for many empirical findings in the domain of spatial navigation as well as phenomena observed in tasks that construe conceptual knowledge and inference as a form of spatial navigation (Bao et al., 2019; Constantinescu et al., 2016; Theves et al., 2019; Viganò et al., 2021). Mental maps are structured representations of the actual physical environment that enable remapping as well as various forms of inferences (Behrens et al., 2018; Whittington et al., 2019). Mental maps are thus ideally suited to support fast learning in a memory system that can quickly adapt to changing environmental features (McClelland et al., 1995). It is therefore not surprising that many theoretical considerations about cognitive maps (e.g., Behrens et al., 2018; Bellmund et

al., 2018) as well as empirical findings in that domain (e.g., Bellmund et al., 2016; Garvert et al., 2017) have a common focus on the hippocampus and adjacent entorhinal cortex – brain regions that are also implicated in the encoding of episodic memories and part of a fast neural learning system (McClelland et al., 1995). Mental maps are flexible representations of our environment and ideally suited in situations where we need to make inferences about individual objects in specific environments. However, as soon as decisions abstract away from such a frame of reference (e.g., *Should you try and get that promotion?*) we also require stable representations that inform us what our environment is typically like in a more generalized sense (Gilboa & Marlatte, 2017; Mack et al., 2020). The results reported in this thesis provide evidence for the existence of a different kind of structured neural memory representation that might provide such stable context-independent representations of key features of our environment: value-weighted schematic representations that are encoded in the mPFC.

Compared to a compositional mental map representation, what would be the main advantage of an intertwined representation of identity and value in such schemas? Value-weighted schematic representations might be formed by Hebbian learning such that elements with overlapping features are encoded by overlapping neural populations. This coding principle would ensure that elements that are experienced more often and are followed by similar affective consequences would form overall stronger associations regardless of the specific context in which they appear. Upon activating such a representation, all properties that are most relevant for adaptive behavior would immediately be available to the individual. These representations would allow for an important kind of generalization and support another type of inference: simulations of the future. One prediction that follows is that these representations would only gradually change and exhibit weak remapping because they depict contingencies that are mostly stable across space and time.

The reported results provide evidence that the mPFC encodes schematic representations of our environment where representations of individual exemplars from our everyday environment are closely intertwined with a representation of their value. These memory representations are rather stable over time and support our ability to simulate potential episodes in the future. Crucially, these representations immediately reveal the affective quality of simulated future episodes (*"what it would feel like"*) and thereby allow us to infer the likely future consequences of our decisions. As we have argued elsewhere, this represents one key adaptive function of our ability for mental time travel: Motivating farsighted decisions (Benoit et al., 2019). By this, the present thesis provides crucial evidence that our memories of the past serve us in ways that are oriented toward the future.

6.3 Memory out of the box: Complex experimental paradigms for naturalistic research

Research on memory has long focused on the investigation of the human abilities to encode, remember, and retrieve relationships between arbitrary cue- and associatively paired target-stimuli. To better understand the stages and neural mechanisms of mnemonic processing, researchers have attempted, as well as they could, to ensure that "prior knowledge" does not confound their results. In doing so, these lines of research have provided us with a rich understanding of the human ability to recognize or recall individual items as well as the neural structures that support these abilities: the hippocampus and adjacent structures in the medial temporal lobes (Eichenbaum et al., 2007; Gold et al., 2006; Rugg et al., 2012). However, in our everyday life, cue and target stimuli do not readily present themselves to us. Instead, they are embedded in a rich sensory context and need to be extracted from a continuous flow of sensory information. Thus, to understand how individuals achieve this and examine memory representations of complex life-like events, we need to investigate memory functions using experimental setups that match the complexities of the real world.

Naturalistic cognitive neuroscience investigates cognitive functions using complex and life-like material and provides a promising avenue for future research (Aliko et al., 2020; Hasson et al., 2004; Hasson & Honey, 2012; Hebart et al., 2020; Zaki & Ochsner, 2009). Insights derived from such naturalistic research are not only valid in the lab, but may also generalize to contexts outside the MRI scanner. By this, they allow us to better understand how different cognitive functions and mental representations support cognition in our actual real-life (Nastase et al., 2020). In a recent study, participants were exposed to a 50-minute-long naturalistic movie and subsequently asked to recall as many episodic details as they remembered while being scanned with fMRI. The results revealed that participants encoded memory representations that generalizable, and could be readily identified both while encoding and retrieving the information (J. Chen et al., 2017). More recently, naturalistic designs have also been employed to study generalized memory representations such as schemas of event sequences. Schematic memory representations were evoked using film clips about highly stereotypical events, such as restaurant visits and scenes at airports. Representations of these events were evoked in the same regions of the brain, regardless of whether these stories were presented in an audiovisual or plainly audio modality. Critically, the mPFC emerged as the only brain region that represented both the context where these movies were set as well as whether the order of the events matched our cognitive event schemas. (Baldassano et al., 2017, 2018). Gagnepain et al. (2020) demonstrated that individual recollections of a visit to a World War II museum reflected shared collective representations of historical events. The researchers

quantified collective knowledge from thirty years of TV footage on World War II using state of the art naturalistic language processing techniques. Representations reflecting this collective knowledge were encoded as schematic representations in the mPFC. In a different study, researchers examined the mental representation of the social network structure of an entire cohort of first-year college students. The participating students' brain activation was assessed with fMRI while they passively viewed images of their fellow students. This task automatically reinstated representations of the individuals and their position within the social network. The results revealed a critical role of the mPFC in encoding key features of the social network structure (Parkinson et al., 2017). Together, these empirical studies demonstrate that naturalistic designs can evoke complex neural representations that can accurately be quantified using state of the art computational and statistical methods (Baldassano et al., 2017; J. Chen et al., 2017). Naturalistic experimental setups also enable examinations of higher order memory representations such as schemas (Baldassano et al., 2018; Gagnepain et al., 2020) and representations of our social network (Parkinson et al., 2017). These studies have thus allowed first insights into how our brains extract and encode episodic memories from a continuous flow of complex sensory information. Moreover, they have provided empirical support for the hypothesis that generalized memory representations support our ability to make sense of the complex structure of the events we encounter in our daily lives. Higher order brain regions that encode multi modal information, the mPFC in particular, appear to mediate such complex and generalized representations.

The findings of the present thesis are based on a similar approach: My work has investigated participants' *pre-existing* schematic memory representations of their actual social networks as well as of locations they know from their real life. We have demonstrated, how aspects of these memory representations can be shaped by a learning mechanism that is based on episodic simulations. This approach has allowed for the investigation of representations that were formed over long periods in our participants' actual life and were not artificially created in the lab. In line with results of other studies using naturalistic designs, we found evidence for a key involvement of the mPFC in the representation of generalized semantic memory representations. While our approach allowed us to investigate how the brain stores existing schematic representations of personally relevant material, it has prevented us from examining how these representations are formed.

According to our formalization of value-weighted schematic representations, value-weighted schemas should emerge as a function of overlapping experiences (Ghosh & Gilboa, 2014; Milivojevic et al., 2015; Reagh & Ranganath, 2021) and should concurrently be shaped

by the value of the encoded exemplars (Farovik et al., 2015). To test these predictions, a naturalistic approach appears particularly promising. Instead of using a single movie as in Chen et al. (2017), participants may be exposed to a complex and new environment using an entire season of a TV show over an extended period of time. By this, participants would immerse themselves in an extended, continuous narrative that is ideally suited to induce schematic representations that are similar to those acquired in real-life. The typical setup of TV shows dictates that individual characters vary with regard to their centrality to the overall story. The relationships between characters are often complex and tend to change over the course of the narrative. Moreover, characters differ with regard to how often they are visible on screen across the series. Finally, these characters commonly also differ with regard to their likability.

Beyond the benefits of exposing participants to highly standardized yet life-like scenarios, this material thus has the additional advantage that many variables of interest can directly be quantified from the video material (McNamara et al., 2017). The total duration that a character is visible on screen may serve as a proxy measure of experience. The structure of the social network of characters may be quantified as the ratio of any two characters' joint appearances over their total time on screen. Moreover, such *objective* measures may be complemented with repeated *subjective* assessments provided by the participants. These arrangements might then provide measures of the likability of the individual characters as well as subjective assessments of centrality and experience. Together with repeated assessments of participants' *neural* memory representations, these *objective* and *subjective* measures might then allow for a detailed investigation of how schematic representations emerge over time and how value shapes their structure.

Employing a naturalistic study design might also shed more light on the functions of value-weighted schemas. One key prediction derived from the work presented in this thesis is that value-weighted schemas should support adaptive behavior. This is well in line with Piaget's original idea that schemas serve as general purpose templates that help us interpret new situations and decide for the best possible course of action (Piaget, 1952). Thus, using these templates to evaluate ongoing experiences allows us to identify aspects of the present situation that may be particularly relevant to remember later on. This is well in line with the empirical finding that both schema *congruent* as well as schema *incongruent* experiences are subsequently better remembered (van Kesteren et al., 2012). Whenever congruency is high, new information may directly be embedded into pre-existing schematic representations. This process has been referred to as assimilation (Piaget, 1952). Whenever congruency is low, the schema may have to be updated or a new schema may have to be formed to accommodate this

new information. This process has been referred to as accommodation (Piaget, 1952; see also Gilboa & Marlatte, 2017). The degree to which information is congruent with the individual's expectation thus mediates subsequent memory. This has similarly been demonstrated using naturalistic video clips from basketball games. A recent study showed that we continuously form predictions about the outcomes of naturalistic events. Whenever these predictions are violated, memories are preferentially encoded and, as a consequence, participants are subsequently better able to remember these events (Antony et al., 2021).

A special instance of our expectations are reward prediction errors that are triggered by evaluative processes: It is well established that rewards can improve memory for past experiences (Adcock et al., 2006; Murty & Adcock, 2014; Wittmann et al., 2008). Rewards have been argued to exert this effect via prediction-error signals that indicate a higher or lower level of reward than expected. This mismatch signal has been shown to support the segmentation of ongoing experiences into discrete sub-events and may thereby support later retention (Rouhani et al., 2020; Shohamy & Adcock, 2010). Interestingly, reward does not only exert this beneficial effect on memory when it is delivered at the time of encoding: Rewards can also retroactively strengthen memories when they are provided after initial encoding (Braun et al., 2018; Elward et al., 2015; Patil et al., 2017).

It is thus highly conceivable that an adaptive memory system would closely intertwine the representations of individual exemplars with a representation of their value. Value-weighted schematic representations encode generalized knowledge about our environment (e.g., individual people) in an associative network of knowledge. Within this network, the strength of connections is not only determined by the degree of overlapping experiences, but also by value. Upon activation of an individual node within such a representation, activation would spread to closely connected nodes leading to automatic co-activation. Such a co-activation of associated memory representations might then facilitate the online formulation of expectations what should happen next. The mismatch between these predictions and the actual unfolding event would then yield a prediction error signal that drives memory encoding and improves later retention (Shohamy & Adcock, 2010). Critically, this prediction error signal would both be based on the typical structure of the environment (i.e., the schema component) *as well as* the affective consequences (i.e., the evaluative component). Thus, the structure of value-weighted schematic representations would account both for schema congruency dependent retention benefits (van Kesteren et al., 2012) as well as for benefits that relate to reward prediction errors (Shohamy & Adcock, 2010). In a naturalistic experiment, as proposed above, these predictions

may be formally tested, which could ultimately shed more light on the adaptive functions of value-weighted schematic representations.

In sum, recent years have seen a surge of interest in deploying more life-like and naturalistic research designs to increase the ecological validity of empirical findings. This surge has been fueled by the development of new sophisticated analysis methods that allow for accurate quantifications of participants neural representations in naturalistic settings. These lines of research have greatly contributed to our understanding of mnemonic functions in more ecologically valid settings. In the years to come, I expect that naturalistic paradigms will aid to unravel how generalized schematic memory representations emerge and clarify their behavioral relevance. This will extend our understanding of how semantic memory representations guide our behavior in the real world.

6.4 Concluding remarks

As humans we are not stuck in an everlasting present but can mentally project ourselves back and forth in time. The present thesis has provided evidence that our memories of the past serve us in adaptive ways that are oriented toward the future: Simulations allow us to *pre*-experience the affective consequences of hypothetical events (*what it would feel like*) already in the here and now. The thesis has demonstrated how this affective experience, in turn, can shape how we perceive of our immediate environment. This simulation-based learning mechanism may provide one instance where simulations can act as a replacement for actual experiences.

The present thesis has also provided a novel perspective on the role of the mPFC for mnemonic and evaluative processes. This part of the human brain is central to both our ability to remember the past and simulate the future. Moreover, the mPFC supports evaluative processes and mediates value-based decisions. The mPFC might support these seemingly disparate functions by encoding schematic representations where knowledge about individual exemplars is closely intertwined with a representation of their value. These value-weighted schematic representations may support a learning mechanism that is purely based on simulated experience.

To conclude, the present thesis has investigated the overlapping contributions of the mPFC to mnemonic and evaluative processes. By this, the thesis has provided new insights into the adaptive functions of our ability to imagine the future: It is not always necessary to actually experience something in order to be able to learn from it.

Bibliography

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. E. (2006). Reward-motivated learning: Mesolimbic activation precedes memory formation. *Neuron*, *50*(3), 507–517. https://doi.org/10.1016/j.neuron.2006.03.036

Addis, D. R. (2020). Mental time travel? A neurocognitive model of event simulation. *Review of Philosophy and Psychology*, *11*(2), 233–259. https://doi.org/10.1007/s13164-020-00470-0

Addis, D. R., Pan, L., Vu, M. A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, *47*(11), 2222–2238. https://doi.org/10.1016/j.neuropsychologia.2008.10.026

Addis, D. R., Wong, A. T., & Schacter, D. L. (2008). Age-related changes in the episodic simulation of future events. *Psychological Science*, *19*(1), 33–41. https://doi.org/10.1111/j.1467-9280.2008.02043.x

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, *82*(4), 463–496. https://doi.org/10.1037/h0076860

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., & Skipper, J. I. (2020). *A "Naturalistic Neuroimaging Database" for understanding the brain using ecological stimuli* (p. 2020.05.22.110817) [Preprint]. bioRxiv. https://www.biorxiv.org/content/10.1101/2020.05.22.110817v1

Alink, A., Walther, A., Krugliak, A., Bosch, J. J. F. van den, & Kriegeskorte, N. (2015). *Mind the drift—Improving sensitivity to fMRI pattern information by accounting for temporal pattern drift* (p. 032391) [Preprint]. bioRxiv. https://www.biorxiv.org/content/10.1101/032391v2

Altgassen, M., Rendell, P. G., Bernhard, A., Henry, J. D., Bailey, P. E., Phillips, L. H., & Kliegel, M. (2015). Future thinking improves prospective memory performance and

plan enactment in older adults. *Quarterly Journal of Experimental Psychology*, *68*(1), 192–204. https://doi.org/10.1080/17470218.2014.956127

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron*, *65*(4), 550–562. https://doi.org/10.1016/j.neuron.2010.02.005

Antony, J. W., Hartshorne, T. H., Pomeroy, K., Gureckis, T. M., Hasson, U., McDougle, S. D., & Norman, K. A. (2021). Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron*, *109*(2), 377-390.e7. https://doi.org/10.1016/j.neuron.2020.10.029

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Bach, D. R., Castegnetti, G., Korn, C. W., Gerster, S., Melinscak, F., & Moser, T. (2018). Psychophysiological modeling: Current state and future directions. *Psychophysiology*, *55*(11), e13214. https://doi.org/10.1111/psyp.13209

Bach, D. R., Flandin, G., Friston, K. J., & Dolan, R. J. (2009). Time-series analysis for rapid event-related skin conductance responses. *Journal of Neuroscience Methods*, *184*(2), 224–234. https://doi.org/10.1016/j.jneumeth.2009.08.005

Bach, D. R., & Friston, K. J. (2013). Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology*, *50*(1), 15–22. https://doi.org/10.1111/j.1469-8986.2012.01483.x

Baeyens, F., Crombez, G., Van den Bergh, O., & Eelen, P. (1988). Once in contact always in contact: Evaluative conditioning is resistant to extinction. *Advances in Behaviour Research and Therapy*, *10*(4), 179–199. https://doi.org/10.1016/0146-6402(88)90014-8

Baeyens, F., Diaz, E., & Ruiz, G. (2005). Resistance to extinction of human evaluative conditioning using a between-subjects design. *Cognition & Emotion*, *19*, 245–268. https://doi.org/10.1080/02699930441000300

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., Shadlen, M. N., & Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *ELife*, *8*, e46080. https://doi.org/10.7554/eLife.46080

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709-721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Bao, X., Gjorgieva, E., Shanahan, L. K., Howard, J. D., Kahnt, T., & Gottfried, J. A. (2019). Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron*, *102*(5), 1066–1075. https://doi.org/10.1016/j.neuron.2019.03.034

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, *16*(10), 1492–1498. https://doi.org/10.1038/nn.3515

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beck, A. T., Brown, G., Steer, R. A., Eidelson, J. I., & Riskind, J. H. (1987). Differentiating anxiety and depression: A test of the cognitive content-specificity hypothesis. *Journal of Abnormal Psychology*, *96*(3), 179–183. https://doi.org/10.1037/0021-843X.96.3.179

Behrens, T. E. J., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund, J. L., Deuker, L., Schröder, T. N., & Doeller, C. F. (2016). Grid-cell representations in mental simulation. *ELife*, *5*, e17089. https://doi.org/10.7554/eLife.17089

Bellmund, J. L., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415), eaat6766. https://doi.org/10.1126/science.aat6766

Benoit, R. G., Berkers, R. M. W. J., & Paulus, P. C. (2018). An adaptive function of mental time travel: Motivating farsighted decisions. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X1700125X

Benoit, R. G., Davies, D. J., & Anderson, M. C. (2016). Reducing future fears by suppressing the brain mechanisms underlying episodic simulation. *Proceedings of the National Academy of Sciences*, *113*(52), E8492–E8501. https://doi.org/10.1073/pnas.1606604114

Benoit, R. G., Gilbert, S. J., & Burgess, P. W. (2011). A neural mechanism mediating the impact of episodic prospection on farsighted decisions. *Journal of Neuroscience*, *31*(18), 6771–6779. https://doi.org/10.1523/JNEUROSCI.6559-10.2011

Benoit, R. G., Gilbert, S. J., Volle, E., & Burgess, P. W. (2010). When I think about me and simulate you: Medial rostral prefrontal cortex and self-referential processes. *NeuroImage*, *50*(3), 1340–1349. https://doi.org/10.1016/j.neuroimage.2009.12.091

Benoit, R. G., Paulus, P. C., & Schacter, D. L. (2019). Forming attitudes via neural activity supporting affective episodic simulations. *Nature Communications*, *10*(1), 2215. https://doi.org/10.1038/s41467-019-09961-w

Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia*, *75*, 450–457. https://doi.org/10.1016/j.neuropsychologia.2015.06.034

Benoit, R. G., Szpunar, K. K., & Schacter, D. L. (2014). Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proceedings of the National Academy of Sciences*, *111*(46), 16550–16555. https://doi.org/10.1073/pnas.1419274111

Bertossi, E., Aleo, F., Braghittoni, D., & Ciaramelli, E. (2016). Stuck in the here and now: Construction of fictitious and future experiences following ventromedial prefrontal damage. *Neuropsychologia*, *81*, 107–116. https://doi.org/10.1016/j.neuropsychologia.2015.12.015

Bertossi, E., Tesini, C., Cappelli, A., & Ciaramelli, E. (2016). Ventromedial prefrontal damage causes a pervasive impairment of episodic memory and future thinking. *Neuropsychologia*, *90*, 12–24. https://doi.org/10.1016/j.neuropsychologia.2016.01.034

Birbaumer, N. (Ed.). (1977). *Psychophysiologie der Angst* (2nd ed.). Urban & Schwarzenberg.

Bjork, E. L., & Bjork, R. A. (1988). On the adaptive aspects of retrieval failure in autobiographical memory. In *Practical aspects of memory: Current research and issues, Vol. 1: Memory in everyday life* (pp. 283–288). John Wiley & Sons.

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., roth, W. T., Dawson, M. E., Filion, D. L., & Society for Psychophysiological Research ad hoc committee on electrodermal measures. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Boyer, P. (2008). Evolutionary economics of mental time travel? *Trends in Cognitive Sciences*, *12*(6), 219–224. https://doi.org/10.1016/j.tics.2008.03.003

Braun, E. K., Wimmer, G. E., & Shohamy, D. (2018). Retroactive and graded prioritization of memory by reward. *Nature Communications*, *9*(1), 4886. https://doi.org/10.1038/s41467-018-07280-0

Brewer, G. A., & Marsh, R. L. (2010). On the role of episodic future simulation in encoding of prospective memories. *Cognitive Neuroscience*, *1*(2), 81–88. https://doi.org/10.1080/17588920903373960

Brod, G., & Shing, Y. L. (2018). Specifying the role of the ventromedial prefrontal cortex in memory formation. *Neuropsychologia*, *111*, 8–15. https://doi.org/10.1016/j.neuropsychologia.2018.01.005

Brown, V. A. (2020). *An introduction to linear mixed effects modeling in R*. https://doi.org/10.31234/osf.io/9vghm

Bryant, R. A., Moulds, M. L., Guthrie, R. M., Dang, S. T., & Nixon, R. D. V. (2003). Imaginal exposure alone and imaginal exposure with cognitive restructuring in treatment of posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, *71*(4), 706–712. https://doi.org/10.1037/0022-006X.71.4.706

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, *61*, 27–48, C1-8. https://doi.org/10.1146/annurev.psych.60.110707.163508

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*, 1–38. https://doi.org/10.1196/annals.1440.011

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57. https://doi.org/10.1016/j.tics.2006.11.004

Bulley, A., Henry, J. D., & Suddendorf, T. (2017). Thinking about threats: Memory and prospection in human threat management. *Consciousness and Cognition*, *49*, 53–69. https://doi.org/10.1016/j.concog.2017.01.005

Bulley, A., & Schacter, D. L. (2020). Deliberating trade-offs with the future. *Nature Human Behaviour*, *4*(3), 238–247. https://doi.org/10.1038/s41562-020-0834-9

Burgess, P. W., & Shallice, T. (1996). Confabulation and the control of recollection. *Memory*, *4*(4), 359–411. https://doi.org/10.1080/096582196388906

Burnham, K. P. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Burnham, K. P., Anderson, D. R., & Burnham, K. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed). Springer.

Buzsáki, G. (1996). The hippocampo-neocortical dialogue. *Cerebral Cortex*, *6*(2), 81–92. https://doi.org/10.1093/cercor/6.2.81

Camille, N., Griffiths, C. A., Vo, K., Fellows, L. K., & Kable, J. W. (2011). Ventromedial frontal lobe damage disrupts value maximization in humans. *Journal of Neuroscience*, *31*(20), 7527–7532. https://doi.org/10.1523/JNEUROSCI.6527-10.2011

Carr, M. F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, *14*(2), 147–153. https://doi.org/10.1038/nn.2732

Castegnetti, G., Zurita, M., & Martino, B. D. (2021). How usefulness shapes neural representations during goal-directed behavior. *Science Advances*, *7*(15), eabd5363. https://doi.org/10.1126/sciadv.abd5363

Chan, A. W.-Y., Kravitz, D. J., Truong, S., Arizpe, J., & Baker, C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature Neuroscience*, *13*(4), 417–418. https://doi.org/10.1038/nn.2502

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. https://doi.org/10.1073/pnas.1402594111

Chen, C., Lu, Q., Beukers, A., Baldassano, C., & Norman, K. A. (2021). Learning to perform role-filler binding with schematic knowledge. *PeerJ*, *9*, e11046. https://doi.org/10.7717/peerj.11046

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. https://doi.org/10.1038/nn.4450

Cherkassky, V., & Ma, Y. (2003). Comparison of model selection for regression. *Neural Computation*, *15*(7), 1691–1714. https://doi.org/10.1162/089976603321891864

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, *29*(39), 12315–12320. https://doi.org/10.1523/JNEUROSCI.2575-09.2009

Ciaramelli, E., De Luca, F., Monk, A. M., McCormick, C., & Maguire, E. A. (2019). What "wins" in VMPFC: Scenes, situations, or schema? *Neuroscience and Biobehavioral Reviews*, *100*, 208–210. https://doi.org/10.1016/j.neubiorev.2019.03.001

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, *26*(8), 3563–3579. https://doi.org/10.1093/cercor/bhw135

Clark, D. M., & Wells, A. (1995). A cognitive model of social phobia. In *Social phobia: Diagnosis, assessment, and treatment* (pp. 69–93). The Guilford Press.

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. https://doi.org/10.1093/scan/nst106

Cohen, M. S. (1997). Parametric Analysis of fMRI Data Using Linear Systems Methods. *NeuroImage*, *6*(2), 93–103. https://doi.org/10.1006/nimg.1997.0278

Collingridge, D. S. (2013). A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research*, *7*(1), 81–97. https://doi.org/10.1177/1558689812454457

Constantinescu, A. O., OReilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

D'Argembeau, A., & Van der Linden, M. (2004). Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and Cognition*, *13*(4), 844–858. https://doi.org/10.1016/j.concog.2004.07.007

D'Argembeau, A., & Van der Linden, M. (2006). Individual differences in the phenomenology of mental time travel: The effect of vivid visual imagery and emotion

regulation strategies. *Consciousness and Cognition*, *15*(2), 342–350. https://doi.org/10.1016/j.concog.2005.09.001

D'Argembeau, A., Xue, G., Lu, Z.-L., Van der Linden, M., & Bechara, A. (2008). Neural correlates of envisioning emotional events in the near and far future. *NeuroImage*, *40*(1), 398–407. https://doi.org/10.1016/j.neuroimage.2007.11.025

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews*, *19*(1), 55–68. https://doi.org/10.1080/07474930008800459

de Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science*, *28*(1), 59–66. https://doi.org/10.1177/0963721418806512

Demblon, J., & D'Argembeau, A. (2016). Networks of prospective thoughts: The organisational role of emotion and its impact on well-being. *Cognition and Emotion*, *30*(3), 582–591. https://doi.org/10.1080/02699931.2015.1015967

Dimsdale-Zucker, H. R., & Ranganath, C. (2018). Representational Similarity Analyses. In *Handbook of Behavioral Neuroscience* (Vol. 28, pp. 509–525). Elsevier. https://doi.org/10.1016/B978-0-12-812028-6.00027-6

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657–661. https://doi.org/10.1038/nature08704

Driskell, J. E., Copper, C., & Moran, A. (1994). Does mental practice enhance performance? *Journal of Applied Psychology*, *79*(4), 481–492. https://doi.org/10.1037/0021-9010.79.4.481

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy*, *38*(4), 319–345. https://doi.org/10.1016/s0005-7967(99)00123-0

Eichenbaum, H. (2017). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, *18*(9), 547–558. https://doi.org/10.1038/nrn.2017.74

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*(1), 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Elward, R. L., Vilberg, K. L., & Rugg, M. D. (2015). Motivated memories: Effects of reward and recollection in the core recollection network and beyond. *Cerebral Cortex (New York, NY)*, *25*(9), 3159–3166. https://doi.org/10.1093/cercor/bhu109

Enkavi, A. Z., Weber, B., Zweyer, I., Wagner, J., Elger, C. E., Weber, E. U., & Johnson, E. J. (2017). Evidence for hippocampal dependence of value-based decisions. *Scientific Reports*, *7*(1), 17738. https://doi.org/10.1038/s41598-017-18015-4

Euston, D. R., Gruber, A. J., & McNaughton, B. L. (2012). The role of medial prefrontal cortex in memory and decision making. *Neuron*, *76*(6), 1057–1070. https://doi.org/10.1016/j.neuron.2012.12.002

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, *26*(8), 3508–3526. https://doi.org/10.1093/cercor/bhw157

Farovik, A., Place, R. J., McKenzie, S., Porter, B., Munro, C. E., & Eichenbaum, H. (2015). Orbitofrontal cortex encodes memories within value-based schemas and represents contexts that guide memory retrieval. *Journal of Neuroscience*, *35*(21), 8333–8344. https://doi.org/10.1523/JNEUROSCI.0134-15.2015

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Gunther, M., Glasser, M. F., Miller, K. L., Ugurbil, K., & Yacoub, E. (2010). Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging. *PloS One*, *5*(12), e15710. https://doi.org/10.1371/journal.pone.0015710

Fellows, L. K. (2019). The functions of the frontal lobes: Evidence from patients with focal brain damage. *Handbook of Clinical Neurology*, *163*, 19–34. https://doi.org/10.1016/B978-0-12-804281-6.00002-1

Foa, E. B., & Hearst-Ikeda, D. (1996). Emotional dissociation in response to trauma: An information-processing approach. In *Handbook of dissociation: Theoretical, empirical, and clinical perspectives* (pp. 207–224). Plenum Press. https://doi.org/10.1007/978-1-4899-0310-5_10

Forester, G., Halbeisen, G., Walther, E., & Kamp, S.-M. (2020). Frontal ERP slow waves during memory encoding are associated with affective attitude formation. *International Journal of Psychophysiology*, *158*, 389–399. https://doi.org/10.1016/j.ijpsycho.2020.11.003

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews Neuroscience*, *6*(2), 119–130. https://doi.org/10.1038/nrn1607

Friston, K. J., Frith, C. D., Frackowiak, R. S., & Turner, R. (1995). Characterizing dynamic brain responses with fMRI: A multivariate approach. *NeuroImage*, *2*(2), 166–172. https://doi.org/10.1006/nimg.1995.1019

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C., Frackowiak, R. S., & Turner, R. (1995). Analysis of fMRI time-series revisited. *NeuroImage*, *2*(1), 45–53. https://doi.org/10.1006/nimg.1995.1007

Friston, K. J., Jezzard, P., & Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, *1*(2), 153–171. https://doi.org/10.1002/hbm.460010207

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., & Turner, R. (1996). Movement-related effects in fMRI time-series. *Magnetic Resonance in Medicine*, *35*(3), 346–355. https://doi.org/10.1002/mrm.1910350312

Gabbott, P. L. A., Warner, T. A., Jays, P. R. L., Salway, P., & Busby, S. J. (2005). Prefrontal cortex in the rat: Projections to subcortical autonomic, motor, and limbic centers. *The Journal of Comparative Neurology*, *492*(2), 145–177. https://doi.org/10.1002/cne.20738

Gagnepain, P., Vallée, T., Heiden, S., Decorde, M., Gauvain, J.-L., Laurent, A., Klein-Peschanski, C., Viader, F., Peschanski, D., & Eustache, F. (2020). Collective memory shapes the organization of individual memories in the medial prefrontal cortex. *Nature Human Behaviour*, *4*(2), 189–200. https://doi.org/10.1038/s41562-019-0779-z

Gamble, B., Moreau, D., Tippett, L. J., & Addis, D. R. (2019). Specificity of future thinking in depression: A meta-analysis. *Perspectives on Psychological Science*, *14*(5), 816–834. https://doi.org/10.1177/1745691619851784

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *37*(7), 1297–1310. https://doi.org/10.1016/j.neubiorev.2013.03.023

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife*, *6*, e17086. https://doi.org/10.7554/eLife.17086

Gerraty, R. T., Davidow, J. Y., Wimmer, G. E., Kahn, I., & Shohamy, D. (2014). Transfer of learning relates to intrinsic connectivity between hippocampus, ventromedial prefrontal cortex, and large-scale networks. *Journal of Neuroscience*, *34*(34), 11297–11303. https://doi.org/10.1523/JNEUROSCI.0185-14.2014

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLOS Computational Biology*, *11*(11), e1004567. https://doi.org/10.1371/journal.pcbi.1004567

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Ghosh, V. E., Moscovitch, M., Colella, B. M., & Gilboa, A. (2014). Schema representation in patients with ventromedial PFC lesions. *Journal of Neuroscience*, *34*(36), 12057–12070. https://doi.org/10.1523/JNEUROSCI.0740-14.2014

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science*, *317*(5843), 1351–1354. https://doi.org/10.1126/science.1144161

Gilboa, A., & Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends in Cognitive Sciences*, *21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013

Gilboa, A., & Moscovitch, M. (2017). Ventromedial prefrontal cortex generates pre-stimulus theta coherence desynchronization: A schema instantiation hypothesis. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *87*, 16–30. https://doi.org/10.1016/j.cortex.2016.10.008

Gilboa, A., Sekeres, M., Moscovitch, M., & Winocur, G. (2014). Higher-order conditioning is impaired by hippocampal lesions. *Current Biology: CB*, *24*(18), 2202–2207. https://doi.org/10.1016/j.cub.2014.07.078

Gold, J. J., Smith, C. N., Bayley, P. J., Shrager, Y., Brewer, J. B., Stark, C. E. L., Hopkins, R. O., & Squire, L. R. (2006). Item memory, source memory, and the medial temporal lobe: Concordant findings from fMRI and memory-impaired patients. *Proceedings of the National Academy of Sciences*, *103*(24), 9351–9356. https://doi.org/10.1073/pnas.0602716103

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments & Computers*, *26*(4), 381–386. https://doi.org/10.3758/BF03204653

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, *130*(5), 769–792. https://doi.org/10.1037/0033-2909.130.5.769

Gregory, W. L., Cialdini, R. B., & Carpenter, K. M. (1982). Self-relevant scenarios as mediators of likelihood estimates and compliance: Does imagining make it so? *Journal of Personality and Social Psychology*, *43*(1), 89–99.

Grueschow, M., Polania, R., Hare, T. A., & Ruff, C. C. (2015). Automatic versus choice-dependent value representations in the human brain. *Neuron*, *85*(4). https://doi.org/10.1016/j.neuron.2014.12.054

Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *Journal of Neuroscience*, *27*(52), 14365–14374. https://doi.org/10.1523/JNEUROSCI.4549-07.2007

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, *104*(5), 1726–1731. https://doi.org/10.1073/pnas.0610561104

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, *11*(7), 299–306. https://doi.org/10.1016/j.tics.2007.05.001

Hassabis, D., & Maguire, E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1263–1271. https://doi.org/10.1098/rstb.2008.0296

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, *24*(8), 1979–1987. https://doi.org/10.1093/cercor/bht042

Hasson, U., & Honey, C. J. (2012). Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, *62*(2), 1272–1278. https://doi.org/10.1016/j.neuroimage.2012.02.004

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hayne, H. (2004). Infant memory development: Implications for childhood amnesia. *Developmental Review*, *24*(1), 33–73. https://doi.org/10.1016/j.dr.2003.09.007

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hebscher, M., & Gilboa, A. (2016). A boost of confidence: The role of the ventromedial prefrontal cortex in memory, decision-making, and schemas. *Neuropsychologia*, *90*, 46–58. https://doi.org/10.1016/j.neuropsychologia.2016.05.003

Hendriks, M. H. A., Daniels, N., Pegado, F., & Op de Beeck, H. P. (2017). The Effect of Spatial Smoothing on Representational Similarity in a Simple Motor Paradigm. *Frontiers in Neurology*, *8*, 222. https://doi.org/10.3389/fneur.2017.00222

Henson, R., Büchel, C., Josephs, O., & Friston, K. J. (1999). The slice-timing problem in event-related fMRI. *NeuroImage*, *9*, 125.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

Holmes, E. A., Deeprose, C., Fairburn, C. G., Wallace-Hadrill, S. M. A., Bonsall, M. B., Geddes, J. R., & Goodwin, G. M. (2011). Mood stability versus mood instability in

bipolar disorder: A possible role for emotional mental imagery. *Behaviour Research and Therapy*, *49*(10), 707–713. https://doi.org/10.1016/j.brat.2011.06.008

Holmes, E. A., & Mathews, A. (2010). Mental imagery in emotion and emotional disorders. *Clinical Psychology Review*, *30*(3), 349–362. https://doi.org/10.1016/j.cpr.2010.01.001

Huettel, S. A., Song, A. W., & McCarthy, G. (2008). *Functional magnetic resonance imaging* (2nd ed). Sinauer Associates.

Inostroza, M., & Born, J. (2013). Sleep for preserving and transforming episodic memory. *Annual Review of Neuroscience*, *36*(1), 79–102. https://doi.org/10.1146/annurev-neuro-062012-170429

Irish, M., Addis, D. R., Hodges, J. R., & Piguet, O. (2012). Considering the role of semantic memory in episodic future thinking: Evidence from semantic dementia. *Brain*, *135*(7), 2178–2191. https://doi.org/10.1093/brain/aws119

Jezzard, P., & Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic Resonance in Medicine*, *34*(1), 65–73. https://doi.org/10.1002/mrm.1910340111

Jones, C. R., Olson, M. A., & Fazio, R. H. (2010). Evaluative conditioning: The "how" question. *Advances in Experimental Social Psychology*, *43*, 205–255. https://doi.org/10.1016/S0065-2601(10)43005-1

Josselyn, S. A., & Frankland, P. W. (2018). Memory allocation: Mechanisms and function. *Annual Review of Neuroscience*, *41*(1), 389–413. https://doi.org/10.1146/annurev-neuro-080317-061956

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633. https://doi.org/10.1038/nn2007

Kappes, H. B., & Morewedge, C. K. (2016). Mental simulation as substitute for experience. *Social and Personality Psychology Compass*, *10*(7), 405–420. https://doi.org/10.1111/spc3.12257

Kappes, H. B., & Oettingen, G. (2011). Positive fantasies about idealized futures sap energy. *Journal of Experimental Social Psychology*, *47*(4), 719–729. https://doi.org/10.1016/j.jesp.2011.02.003

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. https://doi.org/10.1038/nature06713

Kensinger, E. A., & Ford, J. H. (2020). Retrieval of emotional events from memory. *Annual Review of Psychology*, *71*(1), 251–272. https://doi.org/10.1146/annurev-psych-010419-051123

Klein, S. B., Loftus, J., & Kihlstrom, J. F. (2002). Memory and temporal experience: The effects of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future. *Social Cognition*, *20*(5), 353–379. https://doi.org/10.1521/soco.20.5.353.21125

Kliegel, M., McDaniel, M. A., & Einstein, G. O. (Eds.). (2007). *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives*. Psychology Press. https://doi.org/10.4324/9780203809945

Knudsen, E. B., & Wallis, J. D. (2021). Hippocampal neurons construct a map of an abstract value space. *Cell*, *184*(18), 4640–4650. https://doi.org/10.1016/j.cell.2021.07.010

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex*, *20*(12), 2916–2925. https://doi.org/10.1093/cercor/bhq042

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, *3*(3), 363–373. https://doi.org/10.3389/neuro.01.035.2009

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*, 245. https://doi.org/10.3389/fpsyg.2012.00245

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron*, *63*(6), 889–901. https://doi.org/10.1016/j.neuron.2009.07.030

Kurczek, J., Wechsler, E., Ahuja, S., Jensen, U., Cohen, N. J., Tranel, D., & Duff, M. (2015). Differential contributions of hippocampus and medial prefrontal cortex to self-projection and self-referential processing. *Neuropsychologia*, *73*, 116–126. https://doi.org/10.1016/j.neuropsychologia.2015.05.002

Kwan, D., Craver, C. F., Green, L., Myerson, J., Gao, F., Black, S. E., & Rosenbaum, R. S. (2015). Cueing the personal future to reduce discounting in intertemporal choice: Is episodic prospection necessary? *Hippocampus*, *25*(4), 432–443. https://doi.org/10.1002/hipo.22431

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54–64. https://doi.org/10.1038/nrn1825

Lahey, B. B. (2009). Public health significance of neuroticism. *The American Psychologist*, *64*(4), 241–256. https://doi.org/10.1037/a0015309

Landi, S. M., & Buffalo, E. A. (2022). Value representation in the monkey hippocampus. *Trends in Cognitive Sciences*, *26*(1), 4–5. https://doi.org/10.1016/j.tics.2021.10.007

Lebreton, M., Bertoux, M., Boutet, C., Lehericy, S., Dubois, B., Fossati, P., & Pessiglione, M. (2013). A critical role for the hippocampus in the valuation of imagined outcomes. *PLOS Biology*, *11*(10), e1001684. https://doi.org/10.1371/journal.pbio.1001684

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation system in the human brain: Evidence from functional neuroimaging. *Neuron*, *64*(3), 431–439. https://doi.org/10.1016/j.neuron.2009.09.040

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging*, *17*(4), 677–689. https://doi.org/10.1037/0882-7974.17.4.677

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences*, *15*(8), 343–351. https://doi.org/10.1016/j.tics.2011.06.004

Lewis-Peacock, J. A., & Norman, K. A. (2014). Multi-Voxel Pattern Analysis of fMRI Data. In M. Gazzaniga & R. Mangun (Eds.), *The Cognitive Neurosciences* (pp. 911–920). MIT Press.

Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2013). Stimulus value signals in ventromedial PFC reflect the integration of attribute value signals computed in fusiform gyrus and

posterior superior temporal gyrus. *Journal of Neuroscience*, *33*(20), 8729–8741. https://doi.org/10.1523/JNEUROSCI.4809-12.2013

Lin, W.-J., Horner, A. J., Bisby, J. A., & Burgess, N. (2015). Medial prefrontal cortex: Adding value to imagined scenarios. *Journal of Cognitive Neuroscience*, *27*(10), 1957–1967. https://doi.org/10.1162/jocn_a_00836

Lin, W.-J., Horner, A. J., & Burgess, N. (2016). Ventromedial prefrontal cortex, adding value to autobiographical memories. *Scientific Reports*, *6*, 28630. https://doi.org/10.1038/srep28630

Lipp, O. V., Oughton, N., & LeLievre, J. (2003). Evaluative learning in human Pavlovian conditioning: Extinct, but still there? *Learning and Motivation*, *34*(3), 219–239. https://doi.org/10.1016/S0023-9690(03)00011-0

Litt, A., Plassmann, H., Shiv, B., & Rangel, A. (2011). Dissociating valuation and saliency signals during decision-making. *Cerebral Cortex*, *21*(1), 95–102. https://doi.org/10.1093/cercor/bhq065

Liu, Z.-X., Grady, C., & Moscovitch, M. (2017). Effects of prior-knowledge on brain activation and connectivity during associative memory encoding. *Cerebral Cortex*, *27*(3), 1991–2009. https://doi.org/10.1093/cercor/bhw047

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157. https://doi.org/10.1038/35084005

Lyons, A. D., Henry, J. D., Rendell, P. G., Corballis, M. C., & Suddendorf, T. (2014). Episodic foresight and aging. *Psychology and Aging*, *29*(4), 873–884. https://doi.org/10.1037/a0038130

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-019-13930-8

MacLeod, A. K., & Byrne, A. (1996). Anxiety, depression, and the anticipation of future positive and negative experiences. *Journal of Abnormal Psychology*, *105*(2), 286–289. https://doi.org/10.1037/0021-843X.105.2.286

Madan, C., & Kensinger, E. (2021). *Exploring the generalisation of affect across related experiences: A study of affective bleed and memory precision* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jyu93

Maguire, E. A., & Mullally, S. L. (2013). The hippocampus: A manifesto for change. *Journal of Experimental Psychology. General*, *142*(4), 1180–1189. https://doi.org/10.1037/a0033650

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*(3), 1233–1239.

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., & Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, *113*(44), 12574–12579. https://doi.org/10.1073/pnas.1608282113

Martin, V. C., Schacter, D. L., Corballis, M. C., & Addis, D. R. (2011). A role for the hippocampus in encoding simulations of future events. *Proceedings of the National Academy of Sciences*, *108*(33), 13858–13863. https://doi.org/10.1073/pnas.1105816108

Masuda, A., Sano, C., Zhang, Q., Goto, H., McHugh, T. J., Fujisawa, S., & Itohara, S. (2020). The hippocampus encodes delay and value information during delay-discounting decision making. *ELife*, *9*, e52466. https://doi.org/10.7554/eLife.52466

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the

successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McCormick, C., Barry, D. N., Jafarian, A., Barnes, G. R., & Maguire, E. A. (2020). VmPFC drives hippocampal processing during autobiographical memory recall regardless of remoteness. *Cerebral Cortex*, *30*(11), 5972–5987. https://doi.org/10.1093/cercor/bhaa172

McNamara, Q., De La Vega, A., & Yarkoni, T. (2017). Developing a comprehensive framework for multimodal feature extraction. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1567–1574. https://doi.org/10.1145/3097983.3098075

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19. https://doi.org/10.1037/0033-295X.106.1.3

Michelmann, S., Staresina, B. P., Bowman, H., & Hanslmayr, S. (2019). Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nature Human Behaviour*, *3*(2), 143–154. https://doi.org/10.1038/s41562-018-0491-4

Milivojevic, B., Vicente-Grabovetsky, A., & Doeller, C. F. (2015). Insight reconfigures hippocampal-prefrontal memories. *Current Biology*, *25*(7), 821–830. https://doi.org/10.1016/j.cub.2015.01.033

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Milner, B. (1972). Disorders of learning and memory after temporal lobe lesions in man. *Neurosurgery*, *19*(CN_suppl_1), 421–446. https://doi.org/10.1093/neurosurgery/19.CN_suppl_1.421

Miloyan, B., Pachana, N. A., & Suddendorf, T. (2014). The future is here: A review of foresight systems in anxiety and depression. *Cognition & Emotion*, *28*(5), 795–810. https://doi.org/10.1080/02699931.2013.863179

Miloyan, B., & Suddendorf, T. (2015). Feelings of the future. *Trends in Cognitive Sciences*, *19*(4), 196–200. https://doi.org/10.1016/j.tics.2015.01.008

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience*, *24*(21), 4912–4917. https://doi.org/10.1523/JNEUROSCI.0481-04.2004

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655–663. https://doi.org/10.1016/j.neuron.2006.03.040

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Uğurbil, K. (2010). Multiband multislice GE-EPI at 7 Tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, *63*(5), 1144–1153. https://doi.org/10.1002/mrm.22361

Momennejad, I. (2020). Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, *32*, 155–166. https://doi.org/10.1016/j.cobeha.2020.02.017

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Montague, P. R., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*(1), 417–448. https://doi.org/10.1146/annurev.neuro.29.051605.112903

Morewedge, C. K., Huh, Y. E., & Vosgerau, J. (2010). Thought for food: Imagined consumption reduces actual consumption. *Science*, *330*(6010), 1530–1533. https://doi.org/10.1126/science.1195701

Morton, N. W., Sherrill, K. R., & Preston, A. R. (2017). Memory integration constructs maps of space, time, and concepts. *Current Opinion in Behavioral Sciences*, *17*, 161–168. https://doi.org/10.1016/j.cobeha.2017.08.007

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic memory and beyond: The hippocampus and neocortex in transformation. *Annual Review of Psychology*, *67*(1), 105–134. https://doi.org/10.1146/annurev-psych-113011-143733

Mueller, E. M., Sperl, M. F. J., & Panitz, C. (2019). Aversive imagery causes de novo fear conditioning. *Psychological Science*, *30*(7), 1001–1015. https://doi.org/10.1177/0956797619842261

Mullally, S. L., Intraub, H., & Maguire, E. A. (2012). Attenuated boundary extension produces a paradoxical memory advantage in amnesic patients. *Current Biology: CB*, *22*(4), 261–268. https://doi.org/10.1016/j.cub.2012.01.001

Mullally, S. L., & Maguire, E. A. (2014). Memory, imagination, and predicting the future: A common brain mechanism? *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *20*(3), 220–234. https://doi.org/10.1177/1073858413495091

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00128

Murty, V. P., & Adcock, R. A. (2014). Enriched encoding: Reward motivation organizes cortical networks for hippocampal detection of unexpected events. *Cerebral Cortex*, *24*(8), 2160–2168. https://doi.org/10.1093/cercor/bht063

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Neroni, M. A., Gamboz, N., & Brandimonte, M. A. (2014). Does episodic future thinking improve prospective remembering? *Consciousness and Cognition*, *23*, 53–62. https://doi.org/10.1016/j.concog.2013.12.001

Nili, H., Walther, A., Alink, A., & Kriegeskorte, N. (2020). Inferring exemplar discriminability in brain representations. *PLOS ONE*, *15*(6), e0232551. https://doi.org/10.1371/journal.pone.0232551

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, *87*(24), 9868–9872. https://doi.org/10.1073/pnas.87.24.9868

Overwalle, F. V. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Palombo, D. J., Elizur, L., Tuen, Y. J., Te, A. A., & Madan, C. R. (2021). Transfer of negative valence in an episodic memory task. *Cognition*, *217*, 104874. https://doi.org/10.1016/j.cognition.2021.104874

Palombo, D. J., Keane, M. M., & Verfaellie, M. (2015). The medial temporal lobes are critical for reward-based decision making under conditions that promote episodic future thinking. *Hippocampus*, *25*(3), 345–353. https://doi.org/10.1002/hipo.22376

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, *1*(5), 1–7. https://doi.org/10.1038/s41562-017-0072

Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning & Memory*, *24*(1), 65–69. https://doi.org/10.1101/lm.042978.116

Pauling, L., & Coryell, C. D. (1936). The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, *22*(4), 210–216.

Paulus, P. C., Castegnetti, G., & Bach, D. R. (2016). Modeling event-related heart period responses. *Psychophysiology*, *53*(6), 837–846. https://doi.org/10.1111/psyp.12622

Paulus, P. C., Charest, I., & Benoit, R. G. (2020). *Value shapes the structure of schematic representations in the medial prefrontal cortex* [Preprint]. bioRxiv. https://www.biorxiv.org/content/10.1101/2020.08.21.260950v3

Paulus, P. C., Dabas, A., Felber, A., & Benoit, R. G. (2021). *Simulation-based learning influences real-life attitudes* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/k8nxm

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical parametric mapping: The analysis of functional brain images*. Elsevier.

Peters, J., & Büchel, C. (2009). Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. *Journal of Neuroscience*, *29*(50), 15727–15734. https://doi.org/10.1523/JNEUROSCI.3489-09.2009

Peters, J., & Büchel, C. (2010a). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, *66*(1), 138–148. https://doi.org/10.1016/j.neuron.2010.03.026

Peters, J., & Büchel, C. (2010b). Neural representations of subjective reward value. *Behavioural Brain Research*, *213*(2), 135–141. https://doi.org/10.1016/j.bbr.2010.04.031

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, Trans.). W.W. Norton & Co.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, *27*(37), 9984–9988. https://doi.org/10.1523/JNEUROSCI.2131-07.2007

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge Univ. Press.

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(17), R764–R773. https://doi.org/10.1016/j.cub.2013.05.041

Price, J. L., & Drevets, W. C. (2010). Neurocircuitry of mood disorders. *Neuropsychopharmacology*, *35*(1), 192–216. https://doi.org/10.1038/npp.2009.104

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rameson, L. T., Satpute, A. B., & Lieberman, M. D. (2010). The neural correlates of implicit and explicit self-relevant processing. *NeuroImage*, *50*(2), 701–708. https://doi.org/10.1016/j.neuroimage.2009.12.098

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): *Diagnostica*, *51*(4), 195–206. https://doi.org/10.1026/0012-1924.51.4.195

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews. Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Reagh, Z. M., & Ranganath, C. (2021). *A cortico-hippocampal scaffold for representing and recalling lifelike events* (preprint). bioRxiv. https://www.biorxiv.org/content/10.1101/2021.04.16.439894v1

Renner, F., Ji, J. L., Pictet, A., Holmes, E. A., & Blackwell, S. E. (2017). Effects of engaging in repeated mental imagery of future positive events on behavioural activation in individuals with major depressive disorder. *Cognitive Therapy and Research*, *41*(3), 369–380. https://doi.org/10.1007/s10608-016-9776-y

Renner, F., Schwarz, P., Peters, M. L., & Huibers, M. J. H. (2014). Effects of a best-possible-self mental imagery exercise on mood and dysfunctional attitudes. *Psychiatry Research*, *215*(1), 105–110. https://doi.org/10.1016/j.psychres.2013.10.033

Renoult, L., Davidson, P. S. R., Palombo, D. J., Moscovitch, M., & Levine, B. (2012). Personal semantics: At the crossroads of semantic and episodic memory. *Trends in Cognitive Sciences*, *16*(11), 550–558. https://doi.org/10.1016/j.tics.2012.09.003

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Richter, F. R., Chanales, A. J. H., & Kuhl, B. A. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *NeuroImage*, *124*, 323–335. https://doi.org/10.1016/j.neuroimage.2015.08.051

Ritchey, M., Libby, L. A., & Ranganath, C. (2015). Cortico-hippocampal systems involved in memory and cognition: The PMAT framework. *Progress in Brain Research*, *219*, 45–64. https://doi.org/10.1016/bs.pbr.2015.04.001

Robin, J., Buchsbaum, B. R., & Moscovitch, M. (2018). The primacy of spatial context in the neural representation of events. *Journal of Neuroscience*, *38*(11), 2755–2765. https://doi.org/10.1523/JNEUROSCI.1638-17.2018

Rösch, S. A., Stramaccia, D. F., & Benoit, R. G. (2021). *Promoting farsighted decisions via episodic future thinking: A meta-analysis* [Preprint]. PsyArXiv. https://psyarxiv.com/53ju2/

Rosenbaum, R. S., Köhler, S., Schacter, D. L., Moscovitch, M., Westmacott, R., Black, S. E., Gao, F., & Tulving, E. (2005). The case of K.C.: Contributions of a memory-impaired person to memory theory. *Neuropsychologia*, *43*(7), 989–1021. https://doi.org/10.1016/j.neuropsychologia.2004.10.007

Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition*, *203*, 104269. https://doi.org/10.1016/j.cognition.2020.104269

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences*, *16*(3), 147–156. https://doi.org/10.1016/j.tics.2012.01.005

Rugg, M. D., & Vilberg, K. L. (2013). Brain networks underlying episodic memory retrieval. *Current Opinion in Neurobiology*, *23*(2), 255–260. https://doi.org/10.1016/j.conb.2012.11.005

Rugg, M. D., Vilberg, K. L., Mattson, J. T., Yu, S. S., Johnson, J. D., & Suzuki, M. (2012). Item memory, context memory and the hippocampus: FMRI evidence. *Neuropsychologia*, *50*(13), 3070–3079. https://doi.org/10.1016/j.neuropsychologia.2012.06.004

Rumelhart, D. E., & Ortony, A. (1976). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Erlbaum.

Sawamura, H., Orban, G. A., & Vogels, R. (2006). Selectivity of neuronal adaptation does not match response selectivity: A single-cell study of the fMRI adaptation paradigm. *Neuron*, *49*(2), 307–318. https://doi.org/10.1016/j.neuron.2005.11.028

Scarf, D., Gross, J., Colombo, M., & Hayne, H. (2013). To have and to hold: Episodic memory in 3- and 4-year-old children. *Developmental Psychobiology*, *55*(2), 125–132. https://doi.org/10.1002/dev.21004

Schacter, D. L. (1999). Insights From Psychology and Cognitive Neuroscience. *American Psychologist*, 22.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773–786. https://doi.org/10.1098/rstb.2007.2087

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, *8*(9), 657–661. https://doi.org/10.1038/nrn2213

Schacter, D. L., Benoit, R. G., & Szpunar, K. K. (2017). Episodic future thinking: Mechanisms and functions. *Current Opinion in Behavioral Sciences*, *17*, 41–50. https://doi.org/10.1016/j.cobeha.2017.06.002

Schacter, D. L., Gaesser, B., & Addis, D. R. (2013). Remembering the past and imagining the future in the elderly. *Gerontology*, *59*(2), 143–151. https://doi.org/10.1159/000342198

Schafer, M., & Schiller, D. (2018). Navigating social space. *Neuron*, *100*(2), 476–489. https://doi.org/10.1016/j.neuron.2018.10.006

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. https://doi.org/10.1038/nn.3331

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Phil. Trans. R. Soc. B*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*, 8151. https://doi.org/10.1038/ncomms9151

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Scoville, W. B., & Milner, B. (2000). Loss of recent memory after bilateral hippocampal lesions. 1957. *Journal of Neuropsychiatry and Clinical Neurosciences*, *12*(1), 103–113. https://doi.org/10.1176/jnp.12.1.103

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., & Wald, L. L. (2012). Blipped-Controlled Aliasing in Parallel Imaging (blipped-CAIPI) for simultaneous multi-slice EPI with reduced g-factor penalty. *Magnetic Resonance in Medicine*, *67*(5), 1210–1224. https://doi.org/10.1002/mrm.23097

Sharot, T., Riccardi, A. M., Raio, C. M., & Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. *Nature*, *450*(7166), 102–105. https://doi.org/10.1038/nature06280

Shenhav, A., Barrett, L. F., & Bar, M. (2013). Affective value and associative processing share a cortical substrate. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(1), 46–59. https://doi.org/10.3758/s13415-012-0128-4

Shohamy, D. (2011). Learning and motivation in the human striatum. *Current Opinion in Neurobiology*, *21*(3), 408–414. https://doi.org/10.1016/j.conb.2011.05.009

Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences*, *14*(10), 464–472. https://doi.org/10.1016/j.tics.2010.08.002

Shohamy, D., & Daw, N. D. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, *5*, 85–90. https://doi.org/10.1016/j.cobeha.2015.08.010

Simplicio, M. D., Renner, F., Blackwell, S. E., Mitchell, H., Stratford, H. J., Watson, P., Myers, N., Nobre, A. C., Lau-Zhu, A., & Holmes, E. A. (2016). An investigation of mental imagery in bipolar disorder: Exploring "the mind's eye." *Bipolar Disorders*, *18*(8), 669–683. https://doi.org/10.1111/bdi.12453

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In *New Methods in Cognitive Psychology*. Routledge.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, 208–219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, *3*(1), 3–9. https://doi.org/10.1016/j.wep.2014.03.001

Spencer, S. M., & Norem, J. K. (1996). Reflection and distraction: Defensive pessimism, strategic optimism, and performance. *Personality and Social Psychology Bulletin*, *22*(4), 354–365. https://doi.org/10.1177/0146167296224003

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience*, *21*(3), 489–510. https://doi.org/10.1162/jocn.2008.21029

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. https://doi.org/10.1037/0033-295X.99.2.195

Squire, L. R. (2009). The legacy of patient H.M. for neuroscience. *Neuron*, *61*(1), 6–9. https://doi.org/10.1016/j.neuron.2008.12.023

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650

Stahl, C., & Aust, F. (2018). Evaluative conditioning as memory-based judgment. *Social Psychological Bulletin*, *13*(3), 1–30. https://doi.org/10.5964/spb.v13i3.28589

Stawarczyk, D., & D'Argembeau, A. (2015). Neural correlates of personal goal processing during episodic future thinking and mind-wandering: An ALE meta-analysis. *Human Brain Mapping*, *36*(8), 2928–2947. https://doi.org/10.1002/hbm.22818

Suddendorf, T., & Busby, J. (2005). Making decisions with the future in mind: Developmental and comparative identification of mental time travel. *Learning and Motivation*, *36*(2), 110–125. https://doi.org/10.1016/j.lmot.2005.02.010

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *The Behavioral and Brain Sciences*, *30*(3), 299–313; discussion 313-351. https://doi.org/10.1017/S0140525X07001975

Szpunar, K. K., Addis, D. R., & Schacter, D. L. (2012). Memory for emotional simulations: Remembering a rosy future. *Psychological Science*, *23*(1), 24–29. https://doi.org/10.1177/0956797611422237

Szpunar, K. K., Jacques, P. L. S., Robbins, C. A., Wig, G. S., & Schacter, D. L. (2014). Repetition-related reductions in neural activity reveal component processes of mental simulation. *Social Cognitive and Affective Neuroscience*, *9*(5), 712–722. https://doi.org/10.1093/scan/nst035

Szpunar, K. K., & Schacter, D. L. (2013). Get real: Effects of repeated simulation and emotion on the perceived plausibility of future experiences. *Journal of Experimental Psychology: General*, *142*(2), 323–327. https://doi.org/10.1037/a0028877

Talmi, D., Dayan, P., Kiebel, S. J., Frith, C. D., & Dolan, R. J. (2009). How humans integrate the prospects of pain and reward during choice. *Journal of Neuroscience*, *29*(46), 14617–14626. https://doi.org/10.1523/JNEUROSCI.2026-09.2009

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A map for social navigation in the human brain. *Neuron*, *87*(1), 231–243. https://doi.org/10.1016/j.neuron.2015.06.011

Teismann, T., & Margraf, J. (2018). *Exposition und Konfrontation*. Hogrefe. https://doi.org/10.1026/02825-000

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The hippocampus encodes distances in multidimensional feature space. *Current Biology*, *29*(7), 1226-1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Thornton, M. A., & Mitchell, J. P. (2017). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, 1–16. https://doi.org/10.1093/cercor/bhx216

Thulborn, K. R. (2012). My starting point: The discovery of an NMR method for measuring blood oxygenation using the transverse relaxation time of blood water. *NeuroImage*, *62*(2), 589–593. https://doi.org/10.1016/j.neuroimage.2011.09.070

Thulborn, K. R., Waterton, J. C., Matthews, P. M., & Radda, G. K. (1982). Oxygenation dependence of the transverse relaxation time of water protons in whole blood at high field. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *714*(2), 265–270. https://doi.org/10.1016/0304-4165(82)90333-6

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*(3), 403–421. https://doi.org/10.1037/0033-295X.110.3.403

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82. https://doi.org/10.1126/science.1135935

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., & Morris, R. G. M. (2011). Schema-dependent gene activation and memory encoding in neocortex. *Science*, *333*(6044), 891–895. https://doi.org/10.1126/science.1205274

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Tustin, K., & Hayne, H. (2010). Defining the boundary: Age-related changes in childhood amnesia. *Developmental Psychology*, *46*(5), 1049–1061. https://doi.org/10.1037/a0020105

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

van Kesteren, M. T. R., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7550–7555. https://doi.org/10.1073/pnas.0914892107

van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., & Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *Journal of Neuroscience*, *30*(47), 15888–15894. https://doi.org/10.1523/JNEUROSCI.2674-10.2010

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Vertes, R. P., Hoover, W. B., Szigeti-Buck, K., & Leranth, C. (2007). Nucleus reuniens of the midline thalamus: Link between the medial prefrontal cortex and the hippocampus. *Brain Research Bulletin*, *71*(6), 601–609. https://doi.org/10.1016/j.brainresbull.2006.12.002

Viganò, S., Rubino, V., Soccio, A. D., Buiatti, M., & Piazza, M. (2021). Grid-like and distance codes for representing word meaning in the human brain. *NeuroImage*, 117876. https://doi.org/10.1016/j.neuroimage.2021.117876

Walther, E., Halbweisen, G., & Blask, K. (2018). What you feel is what you see: A binding perspective on evaluative conditioning. *Social Psychological Bulletin*, *13*(3), 1–18. https://doi.org/10.5964/spb.v13i3.27551

Whittington, J. C., McCaffary, D., Bakermans, J. J. W., & Behrens, T. E. J. (2022). *How to build a cognitive map: Insights from models of the hippocampal formation* [Preprint]. arXiv. http://arxiv.org/abs/2202.01682

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2019). *The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalisation in the hippocampal formation* (p. 770495) [Preprint]. bioRxiv. https://www.biorxiv.org/content/10.1101/770495v1

Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, *35*(7), 1092–1104. https://doi.org/10.1111/j.1460-9568.2012.08017.x

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, *338*(6104), 270–273. https://doi.org/10.1126/science.1223252

Winecoff, A., Clithero, J. A., Carter, R. M., Bergman, S. R., Wang, L., & Huettel, S. A. (2013). Ventromedial prefrontal cortex encodes emotional value. *Journal of Neuroscience*, *33*(27), 11032–11039. https://doi.org/10.1523/JNEUROSCI.4317-12.2013

Wittmann, B. C., Schiltz, K., Boehler, C. N., & Düzel, E. (2008). Mesolimbic interaction of emotional valence and reward improves memory formation. *Neuropsychologia*, *46*(4), 1000–1008. https://doi.org/10.1016/j.neuropsychologia.2007.11.020

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058

Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited—Again. *NeuroImage*, *2*(3), 173–181. https://doi.org/10.1006/nimg.1995.1023

Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*(6), 224–228. https://doi.org/10.1111/1467-8721.00154

Zaki, J., & Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Sciences*, *1167*(1), 16–30. https://doi.org/10.1111/j.1749-6632.2009.04601.x

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

Zhou, J., Gardner, M. P. H., Stalnaker, T. A., Ramus, S. J., Wikenheiser, A. M., Niv, Y., & Schoenbaum, G. (2019). Rat orbitofrontal ensemble activity contains multiplexed but dissociable representations of value and task structure in an odor sequence task. *Current Biology*, *29*(6), 897-907.e3. https://doi.org/10.1016/j.cub.2019.01.048

Abbreviations

AAL	automatic anatomical labeling
AC-PC	anterior commissure – posterior commissure
AIC	Akaike's information criterion
ALE	Activation Likelihood Estimate
ANOVA	analysis of variance
AV	affective value
BIC	Bayesian information criterion
BOLD	blood oxygen level dependent (signal)
CE	centrality
CS	conditioned stimulus
EPI	echo planar imaging
EX	experience
fMRI	functional magnetic resonance imaging
FoV	field of view
FSL	FMRIB software library
FWE	family wise error
GLM	general linear model
HPC	hippocampus
HRF	hemodynamic response function
Hz	hertz
LER	log evidence ratio
LMEM	linear mixed-effects model
LTI	linear time invariant
MDS	multidimensional scaling
MNI	montreal neurological institute
mPFC	rostral and ventral medial prefrontal cortex
MPRAGE	magnetization-prepared rapid gradient-echo
MR	magnetic resonance
MRI	magnetic resonance imaging
MVPA	multi-voxel pattern analysis
N	noise
n.s.	not significant
N+	sorted noise
PC	principal component

PCA	principal component analysis
PCC	posterior cingulate cortex
PFC	prefrontal cortex
RDM	representational *dis*similarity matrix
RF	radiofrequency
ROI	region of interest
RSA	representational similarity analysis
RSM	representational similarity matrix
s	seconds
SCR	skin conductance response
SD	standard deviation
SE	standard error
SEM	standard error of the mean
SPM	statistical parametric mapping
TE	echo time
TR	repetition time
US	unconditioned stimulus
vmPFC	ventromedial prefrontal cortex

A Supplements Study 1

A.1 Matching liked and disliked people on familiarity for the replication study

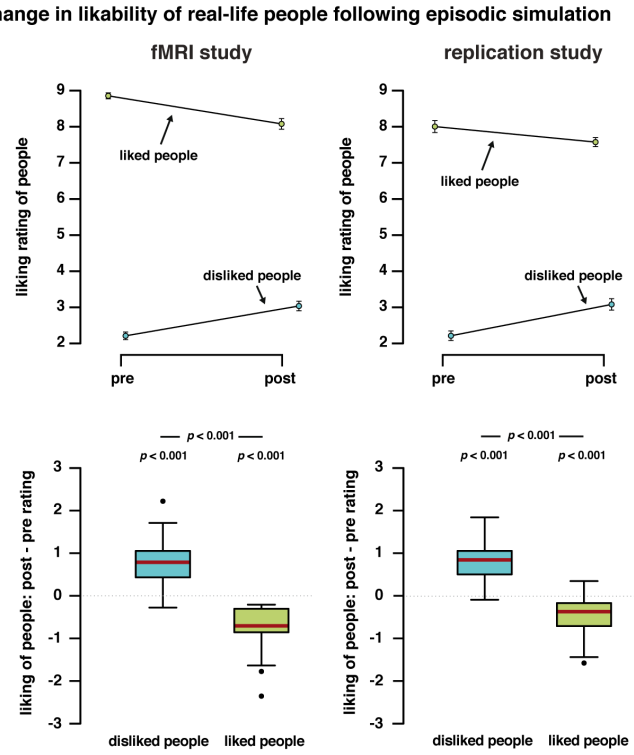
The overall procedure for the replication study was identical to the main study albeit for the omission of phases I and III. Participants provided names of 100 places and of 150 people that they were personally familiar with. We then asked participants to rate on a 9-point scale (i) how familiar they were with each person and place, and (ii) how much they liked each individual item. We then selected 28 neutral places (i.e., rating of 5 and, if necessary, additional items with the next lower and greater ratings).

Based on the liking and familiarity ratings, we selected two sets of people (i.e., the liked and disliked) such that we (i) maximized the difference in liking and (ii) minimized the difference in familiarity between the sets. To this end, we used a stepwise selection approach. First, we used linear regression to remove shared variance of familiarity and liking from the raw liking scores. We then selected the 14 people with highest and the 14 people with lowest residual liking scores. We then checked whether the two sets were of equal average familiarity. If this was not the case, we sought to match the two sets by replacing people from the set with the lower average familiarity. That is, of the least familiar people we took either the most liked person (for the disliked set) or the least liked person (for the liked set) and exchanged it with the respective next best person (i.e., the most familiar person with a liking rating smaller than four or greater than 6, respectively). This person was included in the set if it increased the mean familiarity of the set. This approach continued until the sets were matched on familiarity or when the difference could no further be minimized. Finally, we checked if we could include other people of identical familiarity that would further maximize the difference in liking between both sets. We then randomly paired each of the selected liked and disliked people with a unique neutral place to create the critical 28 pairings.

A.2 Behavioral results

	Familiarity of people		Familiarity of places		Plausibility		Pleasantness	
	liked	**disliked**	**'liked'**	**'disliked'**	**liked**	**disliked**	**liked**	**disliked**
fMRI study (n = 18)								
Mean	7.5	5.3	5.6	5.7	4.6	3.6	7.8	3.2
Std. Deviation	0.8	1.5	1.0	0.8	1.0	1.2	0.7	0.8
Minimum	5.8	3.2	3.1	4.1	3.3	1.6	6.2	1.6
Maximum	8.4	8.4	7.6	7.0	6.6	5.9	8.9	4.3
Behavioral study (n = 30)								
Mean	5.5	5.5	6.5	6.7	4.6	3.6	7.7	3.1
Std. Deviation	1.2	1.2	1.0	0.9	1.1	1.1	0.8	0.8
Minimum	3.4	3.4	4.4	5.0	2.7	2.1	5.3	1.9
Maximum	8.3	8.3	8.0	8.5	6.2	5.9	8.9	5.1

A.3 Change in likability of real-life people following episodic simulation



Though the studies were designed to examine changes in the liking of the neutral places, we also explored concomitant changes in the liking of the paired liked and disliked people. Consistent across the two studies, the liking of the liked people decreased, whereas the liking of the disliked people increased from the pre- to the post test (with significant differences between the two respective difference scores) (study 1: liked people: Wilcoxon test: $W_{17} = 0$, $p < 0.001$, matched rank serial correlation $r = -1$, because of significant Shapiro-Wilk: $W = 0.83$, $p = 0.004$; disliked people: $t_{17} = 5.98$, $p < 0.001$, $d = 1.41$; difference: Wilcoxon test: $W_{17} = 0$, $p < 0.001$, matched rank serial correlation $r = -1$, because of significant Shapiro-Wilk: $W = 0.87$, $p = 0.021$) (study 2: liked people: $t_{29} = -5.16$, $p < 0.001$, $d = -0.94$; disliked people: $t_{29} = 10.7$, $p < 0.001$, $d = 1.95$; difference: $t_{29} = -11.08$, $p < 0.001$, $d = -2.02$). Though this pattern is consistent with the hypothesized transfer of value between the constituting elements of a simulation (i.e., the valenced person and the neutral place), we caution any interpretation. The studies were not designed to include a proper baseline to evaluate the changes for the people, and we therefore cannot rule out simple explanations such as regression to the mean (i.e., from either very positive or very negative to the neutral "mean"). Error bars in the pre- vs. post panels indicate the respective standard error of the means. Boxplots indicate the median, central quartiles, and +/- 2.7 SD. The dots indicate outliers beyond that range.

A.4 Detailed results of the replication study

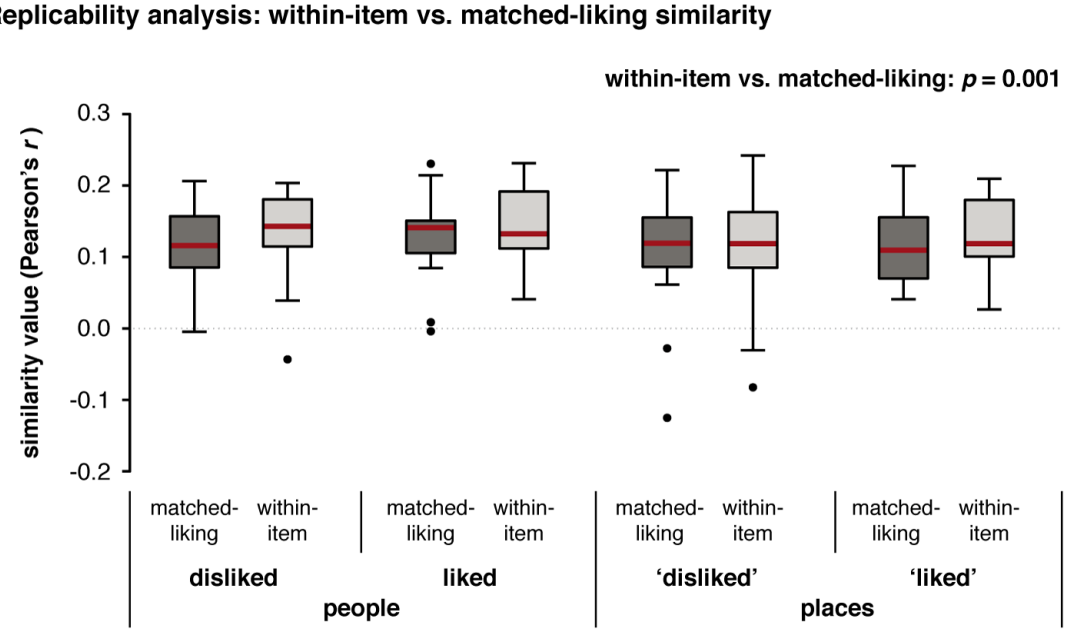
As in the fMRI study, we selected places that participants felt neutral towards and paired these with either much liked or much disliked people (difference in liking: $t_{29} = 23.06, p < 0.001, d = 4.21$). Importantly, this time, the liked people were not more familiar than the disliked people ($t_{29} = 1.21, p = 0.238, d = 0.22$) (Supplementary Table A.2). (In fact, they were exactly matched for 28 of the 30 participants). The places in the two conditions did also not differ on this dimension (Shapiro-Wilk: $W = 9.26, p = 0.038$, hence Wilcoxon: $W_{29} = 143.5, p = 0.11$, matched rank biserial correlation r = -0.38).

Of the simulated episodes, participants judged those featuring liked people as more plausible ($t_{29} = 4.18, p < 0.001, d = 0.76$) and, importantly, also as more pleasant ($t_{29} = 19.02, p < 0.001, d = 3.47$).

We replicated the observation that both kinds of places were deemed more positive following simulation. In the pre-registration, we had specified that we would use *t*-tests, which were indeed significant (paired with liked people: $t_{29} = 6.24, p < 0.001, d = 1.14$; paired with disliked people: $t_{29} = 3.47, p = 0.002, d = 0.63$). However, for the change scores of the places paired with liked people, a Shapiro-Wilk test indicated a deviation from normality ($W = 0.93, p = 0.046$). We therefore additionally analyzed these data with a Wilcoxon test, which also yielded a significant effect ($W = 439.5, p < 0.001$, matched rank biserial correlation $r = 0.89$).

Critically, as predicted (https://aspredicted.org/blind.php?x=th9zv6), we also replicated the critical finding of a more positive shift in attitude for places that had been imagined with liked people ($t_{29} = 3.77, p < 0.001, d = 0.69$) (Figure 3B).

A.5 Control analysis – Pattern replicability



Our fMRI results indicate that univariate vmPFC activity is sensitive to value (Figure 4 and Figure 5)(see also Bartra et al., 2013; Litt et al., 2011). This univariate effect, in turn, may drive differences between multivariate activity patterns for items that differ in value. The current analysis examines whether we can observe evidence for unique representations for individual people and places, even if we compare their within-item similarity not broadly to all other items of the same category (as we had done with the between-item similarity). Instead, we compare the within-item similarity to the similarity of items that are exactly matched in terms of value (i.e., matched-liking similarity). We therefore attempted to pair each item with another item (of the same category and from the same functional run) that had received the identical liking rating on the post test. We thus ensure that the comparison of the item with itself (within-item similarity) versus with its paired item (matched-liking similarity) is not biased by possible value differences. This analysis is based on 78.08% of the items for which it was possible to assign a matched 'partner'. For many items, there was more than one possible match. Therefore, on each of 1000 iterations, we randomly drew one of the possible matches as a partner before then computing the similarity between the items and their respective partners. We finally averaged these similarity scores across all iterations, which we took as an estimate of the matched-liking similarity. Critically, this approach yielded the predicted larger within-item than matched-liking similarity ($F_{1,17} = 14.47$, $p = 0.001$, $\eta^2 = 0.46$), i.e., a greater within-item similarity even when we had directly controlled for effects of value. The control analysis thus further demonstrates that the vmPFC codes for individual elements from our environment. In addition, only the main effect of material (people vs. places, $F_{1,17} = 6.63$, $p = 0.02$, $\eta^2 = 0.28$) but no interactions including the comparison factor (i.e., no interaction with within-item vs. matched-item, all $F_{1,17} < 0.51$, all $p > 0.48$, all $\eta^2 < 0.03$) were significant. Boxplots indicate the median, central quartiles, and +/- 2.7 SD. Dots denote outliers beyond that range.

A.6 Parametric modulation by affective value

Region	approx. BA	Hemisphere	MNI (peak)			Voxels	Z(max)
			x	*y*	*z*		
positive modulation by liking							
vmPFC	11	L/R	6	26	-14	2477	4.93
			6	0	-10	same cluster	4.79
			-10	2	-12	same cluster	4.77
vmPFC	10, 11, 25	L/R	6	26	-14	991*	4.93
			4	24	-10	same cluster	4.74
			-4	20	-16	same cluster	4.24
			-10	38	-16	same cluster	4.02
			-6	26	-12	same cluster	3.99
			-12	40	-10	same cluster	3.84
			4	42	-16	same cluster	3.66
			10	44	-12	same cluster	3.63
			6	16	-20	same cluster	3.61
			12	40	-14	same cluster	3.61
			10	48	-10	same cluster	3.44
			12	50	-4	same cluster	3.32
negative modulation by liking							
vPC/dPC	1, 39	R	36	-30	32	921	3.96
			38	-22	36	same cluster	3.68
			46	-58	46	same cluster	3.55
vPC	7, 39, 40	L	-32	-46	38	459	3.70
			-38	-48	44	same cluster	3.55
			-30	-64	42	same cluster	3.50

Note. Thresholded at $p < 0.05$ FWE-cluster corrected with a cluster forming threshold of $p < 0.001$ and at least 15 contiguous voxels. vmPFC = ventromedial prefrontal cortex, dPC = dorsal parietal cortex, vPC = ventral parietal cortex; * = significant following small-volume-correction for the vmPFC region-of-interest.

We provide coordinates of individual peaks to better characterize the extend of the significant clusters. However, because the results were obtained using cluster-correction, one should not infer that all of the individual peaks are necessarily significantly activated by themselves (Woo et al., 2014).

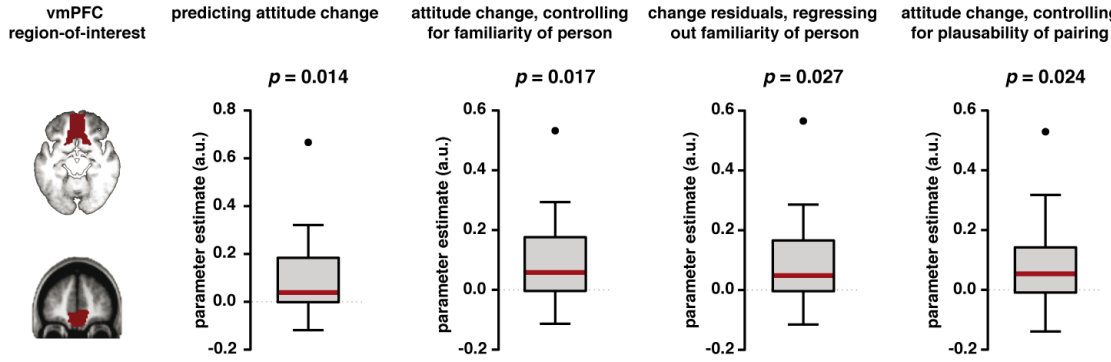
A.7 Parametric modulation by the value of the US and by change in value for the CS

Region	approx. BA	Hemisphere	MNI (peak)			Voxels	Z(max)
			x	*y*	*z*		
positive modulation by liking of the person							
dlPFC	8	L	-22	26	38	235	3.71
			-20	38	46	same cluster	3.37
vmPFC; CN	25	L/R	-6	14	-18	856	4.95
			8	10	-14	same cluster	4.49
			10	26	6	same cluster	4.22
Precuneus	7	L/R	14	-50	52	648	4.08
			-6	-50	54	same cluster	4.07
			-8	-52	64	same cluster	3.74
dOC	19, 39	L	-44	-76	30	262	4.05
			-40	-70	26	same cluster	3.96
			-36	-84	24	same cluster	3.77
vmPFC	10, 11	L	-6	52	-6	90*	3.46
			-6	44	-14	same cluster	3.17
vmPFC	11, 25	L	-6	14	-20	85*	4.76
			-10	16	-16	same cluster	4.05
			-2	26	-20	same cluster	3.16
negative modulation by liking of the person							
none							
positive modulation by subsequent change in liking of the place							
vmPFC	11	L	-6	16	-22	102*	4.19
			-8	22	-20	same cluster	3.86
			-10	26	-20	same cluster	3.70
			-12	30	-20	same cluster	3.52
			-12	32	-16	same cluster	3.23
negative modulation by subsequent change in liking of the place							
none							

Note. Thresholded at $p < 0.05$ FWE-cluster corrected with a cluster forming threshold of $p < 0.001$ and at least 15 contiguous voxels. vmPFC = ventromedial prefrontal cortex, dlPFC = dorsolateral prefrontal cortex, CN = Caudate Nucleus, dOC = dorsal occipital cortex; * = significant following small-volume-correction for the vmPFC region-of-interest.

We provide coordinates of individual peaks to better characterize the extend of the significant clusters. However, because the results were obtained using cluster-correction, one should not infer that all of the individual peaks are necessarily significantly activated by themselves (Woo et al., 2014).

A.8 Average contrast estimates from the vmPFC region of interest



Average contrast estimates from the vmPFC region-of-interest indicating that BOLD signal in this region was modulated by the subsequent change in liking of the CS (i.e., the place). This was also the case when controlling for the familiarity of the paired US (i.e., the person), either by including this effect as a first parametric regressor or by using the residual change values after regressing out possible effects of familiarity. Moreover, the effect was also present when controlling for the plausibility of the CS-US pairing. Boxplots indicate the median, central quartiles, and +/- 2.7 SD. The dots denote an outlier beyond that range.

B Supplements Study 2

B.1 Description of the selection algorithm for the people

As in previous studies (Benoit et al., 2014, 2019; Paulus et al., 2020), disliked people tended to also be less familiar than liked ones. In our selection procedure, we thus sought to simultaneously maximize differences in liking while minimizing differences in familiarity (based on the pre-simulation ratings).

In a first step, we split all provided people into disliked (liking-rating of 1-3), neutral (liking-rating of 4-6), and liked (liking-rating of 7-9) people. If there were less than ten people in any of these initial sets, the participant was compensated for the time spent at the institute, and not invited to return for the simulation session (n = 3).

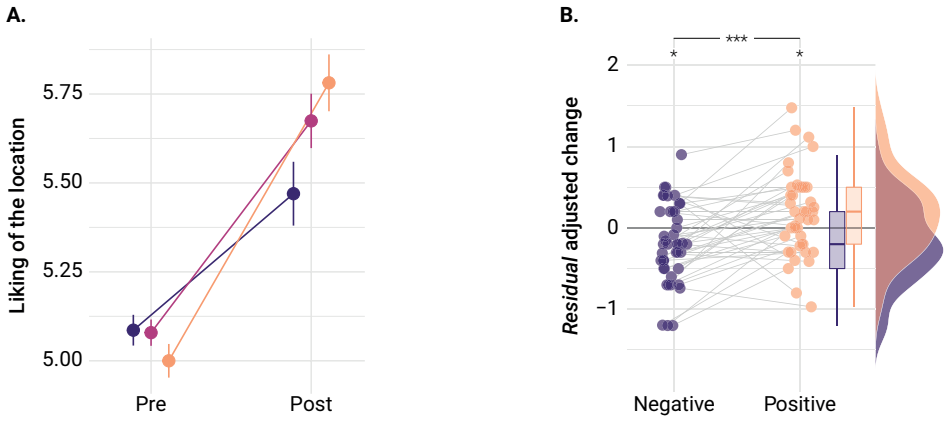
In a next step, we then identified the person with the lowest liking rating and – if possible – selected people with matching familiarity in the neutral and liked sets. Importantly, we always tried to select those people that were respectively closest to the neutral center (rating of 5) or the positive end of the scale (rating of 9). We repeated this process for all items in the set of disliked people.

If this procedure did not yield complete sets of disliked, neutral, and liked people, we drew 25'000 random samples from the remaining pool of people that would complete the sets. For each of these random samples, we determined the sum of squared familiarity differences between the three conditions in each resulting set. We then selected the sample with the smallest sum of squared familiarity differences between the conditions. This selection was then used in the main part of the experiment.

B.2 Control analysis: Familiarity of the people

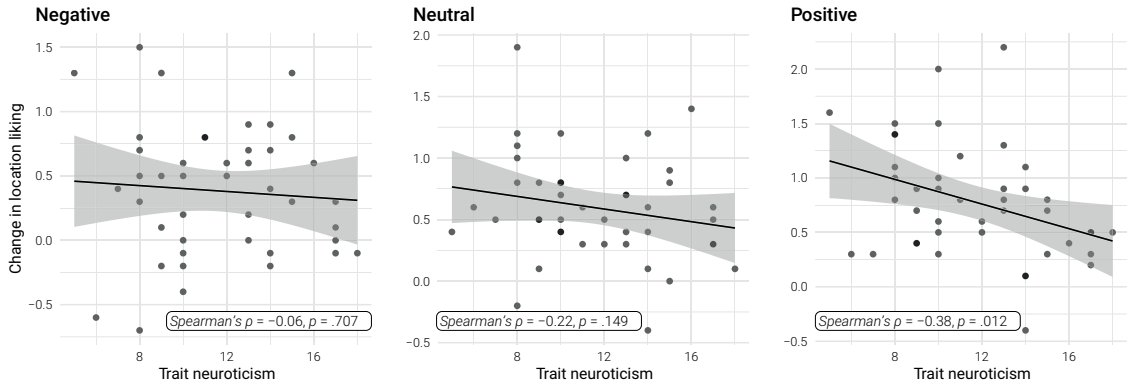
Despite our selection process, the people (i.e., the US) still differed in familiarity ($F(1.39, 58.31) = 5.88$, $p = .011$, $\eta^2_G = .01$) (see also Benoit et al., 2019). We thus sought to statistically control for this possible confound. Specifically, across all valence conditions, we first regressed the familiarity of the people on the change in liking of the locations. This was done separately for each participant. We then subjected the residuals of this regression analysis to another rANOVA. Importantly, this analysis retained the effect of *Valence* ($F(1.91, 80.16) = 10.44$, $p < .001$, $\eta^2_G = .2$) as well as the relative upward and downward shifts following simulations with liked versus disliked people ($ps < .05$; Supplement B.3). The observed effects are thus unlikely to be accounted for by the familiarity of the people.

B.3 Liking change of the locations and residual adjusted liking change



A. Episodic simulations induce attitude changes toward the simulated locations. **B.** Differences in familiarity between the sets of selected people (US) cannot account for differences in the change in liking of the locations (CS).

B.4 Reduced simulation-based learning in individuals with high trait neuroticism



Following simulations with liked people, individuals higher in neuroticism exhibited a weaker upward shift in liking of the locations.

C Supplements Study 3

C.1 Searchlight analysis – Node coding

Region (peak)	approx. BA (peak)	Hemi-sphere	MNI (peak) *x*	*y*	*z*	Voxels	*Z(max)*
PCC, precuneus	7, 23, 31	L/R	-2	-58	24	10,335	7.19
			-2	-50	28		7.14
			-2	-62	42		7.08
Medial PFC	10, 11	L	-4	56	2	353	6.54
			-10	68	0		5.30
			-8	44	-14		5.04
Dorsolateral PFC	45, 46	L	-52	34	10	289	6.48
			-46	32	18		5.58
			-40	38	18		5.41
Lateral PFC	6, 44	L	-48	8	32	297	6.38
			-42	4	44		5.71
			-52	16	26		5.07
Dorsal mPFC	9, 10	R	4	50	20	227	6.03
			10	56	28		5.57
			4	58	18		5.40
Fronto-parietal cortex	6, 8	L	-26	24	46	401	5.92
			-22	24	56		5.40
			-28	4	58		5.35
Early visual	visual assoc	L	16	-90	0	133	5.67
Fronto-parietal cortex	6	R	34	8	58	148	5.66
			40	2	50		5.48
			26	12	58		5.46
Early visual	visual assoc	L	-12	-92	-2	53	5.55
			-6	-98	-6		5.13
Lateral PFC	44, 45, 46	R	58	22	18	146	5.54
			56	24	10		5.51
			44	38	10		5.20
Dorsolateral PFC	8	R	32	28	40	65	5.43
			32	22	48		4.91
Medial PFC	10	R	6	54	-8	65	5.35
			12	50	-4		4.99
Temporal cortex	21	L	-56	-20	-10	40	5.34
Lateral PFC	9	R	36	42	24	50	5.21
			42	36	20		5.15
Temporal cortex	21, fusiform	R	64	-44	4	137	5.17
			56	-46	-6		5.17
			58	-52	2		5.13
Dorsal mPFC	9, 10	L	-20	50	28	38	5.16
			-14	54	32		5.12

Region (peak)	approx. BA (peak)	Hemi-sphere	MNI (peak) *x*	*y*	*z*	Voxels	*Z(max)*

Note. Thresholded at $P < .05$, voxel FWE corrected and at least 30 contiguous voxels. BA = Brodmann area.

C.2 Correlation of the principal component in anatomical ROIs

		Descriptive statistics		Shapiro Wilk test		Significance test		
ROI	**Category**	***Mean***	***SD***	***W***	***P***	t_{35}	P_{Holm}	***d***
vmPFC (LGM)	people	0.037	0.054	0.97	.47	4.16	< .001	0.69
	places	0.019	0.045	0.97	.41	2.52	.033	0.42
	people vs. places	0.019	0.061	0.96	.17	1.83	.076	0.31
mPFC (BN)	people	0.044	0.062	0.97	.35	4.26	< .001	0.71
	places	0.023	0.043	0.96	.21	3.25	.005	0.54
	people vs. places	0.020	0.074	0.95	.09	1.67	.104	0.28
PCC (BN)	people	0.040	0.069	0.99	.96	3.47	.004	0.58
	places	0.025	0.044	0.95	.13	3.41	.004	0.57
	people vs. places	0.015	0.082	0.95	.14	1.11	.276	0.18

Note. Correlation coefficient is Kendall's τ_a. *P*-values are adjusted for multiple comparisons using Holm's method. ROI = Region of interest, *SD* = standard deviation, *d* = Cohen's *d,* vmPFC = ventromedial prefrontal cortex, mPFC = medial prefrontal cortex, PCC = posterior cingulate cortex, LGM = Liu, Grady, Moscovitch (see also Benoit et al., 2019; Liu et al., 2017), BN = Brainnetome (Fan et al., 2016).

C.3 Correlations of centrality, experience, and affective value

			Descriptive statistics		Shapiro Wilk test		Significance test		
ROI	**Model**	**Category**	**Mean**	**SD**	**W**	***P***	***stat***	***P***	***d***
mPFC (SL)	CE	people	0.007	0.065	0.96	.24	0.67	.504	0.11
		places	-0.001	0.040	0.97	.44	-0.18	.859	0.03
	EX	people	0.041	0.062	0.96	.16	3.98	< .001	0.66
		places	0.023	0.047	0.97	.45	2.92	.006	0.49
	AV	people	0.036	0.061	0.96	.30	3.49	.001	0.58
		places	0.022	0.051	0.97	.38	2.61	.013	0.43
HPC (BN)	CE	people	-0.013	0.035	0.98	.59	-2.23	.032	0.37
		places	-0.001	0.036	0.94	.04	303[w]	.647	0.03
	EX	people	0.005	0.041	0.97	.31	0.68	.500	0.11
		places	0.006	0.037	0.97	.39	0.92	.363	0.15
	AV	people	0.007	0.043	0.97	.48	1.04	.304	0.17
		places	0.005	0.035	0.98	.80	0.82	.416	0.14
PCC (SL)	CE	people	0.005	0.064	0.95	.07	0.46	.647	0.08
		places	0.004	0.045	0.97	.41	0.59	.558	0.10
	EX	people	0.048	0.056	0.97	.33	5.15	< .001	0.86
		places	0.024	0.049	0.97	.54	2.93	.006	0.49
	AV	people	0.035	0.070	0.96	.23	3.03	.005	0.50
		places	0.018	0.063	0.96	.21	1.75	.089	0.29
vmPFC (LGM)	CE	people	0.003	0.051	0.98	.75	0.40	.694	0.07
		places	-0.004	0.039	0.86	< .001	373[w]	.539	0.10
	EX	people	0.039	0.052	0.98	.80	4.45	< .001	0.74
		places	0.017	0.041	0.98	.86	2.50	.017	0.42
	AV	people	0.034	0.053	0.97	.37	3.84	< .001	0.64
		places	0.015	0.049	0.97	.49	1.91	.064	0.32

ROI	Model	Category	Descriptive statistics		Shapiro Wilk test		Significance test		
			Mean	SD	W	*P*	*stat*	*P*	*d*
mPFC (BN)	CE	people	0.003	0.057	0.98	.72	0.29	.773	0.05
		places	0.001	0.043	0.87	<.001	388[w]	.396	0.01
	EX	people	0.049	0.057	0.97	.36	5.18	< .001	0.86
		places	0.021	0.045	0.98	.60	2.81	.008	0.47
	AV	people	0.037	0.061	0.98	.71	3.68	< .001	0.61
		places	0.018	0.050	0.97	.36	2.17	.037	0.36
PCC (BN)	CE	people	-0.001	0.061	0.94	.07	-0.12	.904	0.02
		places	0.007	0.050	0.98	.87	0.88	.383	0.15
	EX	people	0.043	0.058	0.98	.84	4.40	< .001	0.73
		places	0.026	0.047	0.99	.97	3.35	.002	0.56
	AV	people	0.032	0.071	0.96	.22	2.70	.011	0.45
		places	0.013	0.060	0.96	.17	1.35	.185	0.23

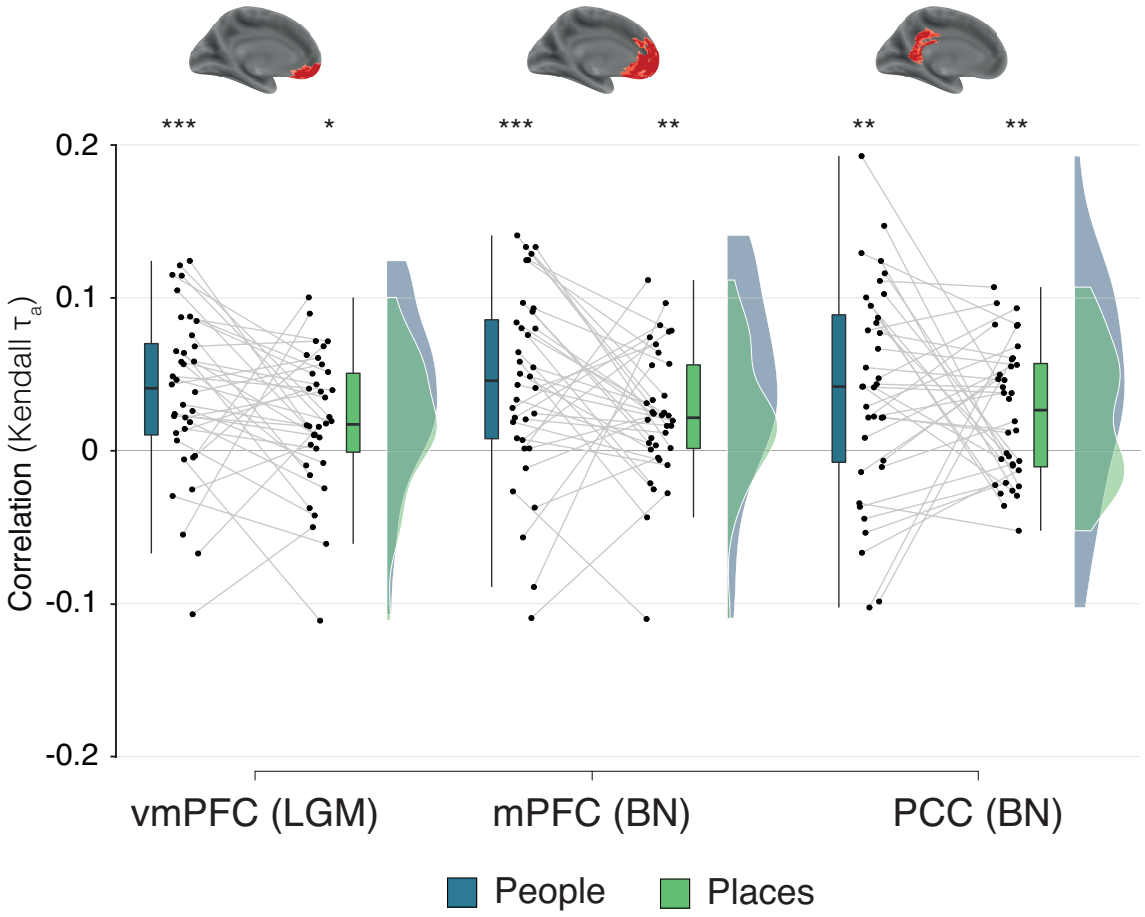
Note. Correlation coefficient is Kendall's τ_a. Reported *P*-vales are uncorrected for exploratory purposes. ROI = region of interest, SD = standard deviation, *stat* = test statistic, [w] – statistic *W* of a Wilcox test, used due to a deviation from normality as indicated by the Shapiro-Wilk-Test. All other statistics indicate the *t*-statistic of a simple *t*-test (*df* = 35) , *d* = Cohen's *d*. mPFC = medial prefrontal cortex, HPC = hippocampus, PCC = posterior cingulate cortex, vmPFC = ventromedial prefrontal cortex, SL = searchlight, LGM = Liu, Grady, Moscovitch (see also Benoit et al., 2019; Liu et al., 2017), BN = Brainnetome (Fan et al., 2016), CE = centrality, EX = experience, AV = affective value.

C.4 Linear mixed effects models – Model parameters of the winning models

ROI	Winning model(s)	Effect	β	*SE*	χ^2	*Pr(>χ^2)*	*Sig.*
mPFC (SL)	PC	$Category_{place}$	-0.026	0.008	11.61	< .001	***
		PC	0.048	0.012	17.12	< .001	***
		$Category_{place}$:PC	-0.005	0.008	0.36	.546	*n.s.*
HPC (BN)	CE	$Category_{place}$	0.020	0.004	31.49	< .001	***
		CE	-0.007	0.004	0.90	.342	*n.s.*
		$Category_{place}$:CE	0.007	0.004	3.04	.081	*n.s.*
	AV	$Category_{place}$	0.023	0.004	31.47	< .001	***
		AV	0.004	0.004	1.47	.225	*n.s.*
		$Category_{place}$:AV	-0.002	0.004	0.16	.686	*n.s.*
PCC (SL)	AV	$Category_{place}$	0.014	0.009	1.51	.219	*n.s.*
		AV	0.026	0.007	10.71	.001	**
		$Category_{place}$:AV	-0.010	0.005	3.92	.048	*
vmPFC (LGM)	PC	$Category_{place}$	-0.019	0.005	19.30	< .001	***
		PC	0.027	0.007	11.75	< .001	***
		$Category_{place}$:PC	-0.009	0.006	2.28	.131	*n.s.*
mPFC (BN)	PC	$Category_{place}$	-0.011	0.006	5.29	.021	*
		PC	0.037	0.007	23.62	< .001	***
		$Category_{place}$:PC	-0.008	0.006	1.56	.211	*n.s.*
	AV	$Category_{place}$	-0.010	0.006	4.32	.038	*
		AV	0.025	0.006	14.35	< .001	***
		$Category_{place}$:AV	-0.005	0.005	0.97	.326	*n.s.*
PCC (BN)	AV	$Category_{place}$	0.057	0.008	52.72	< .001	***
		AV	0.021	0.007	6.87	.009	**
		$Category_{place}$:AV	-0.010	0.005	4.84	.028	*

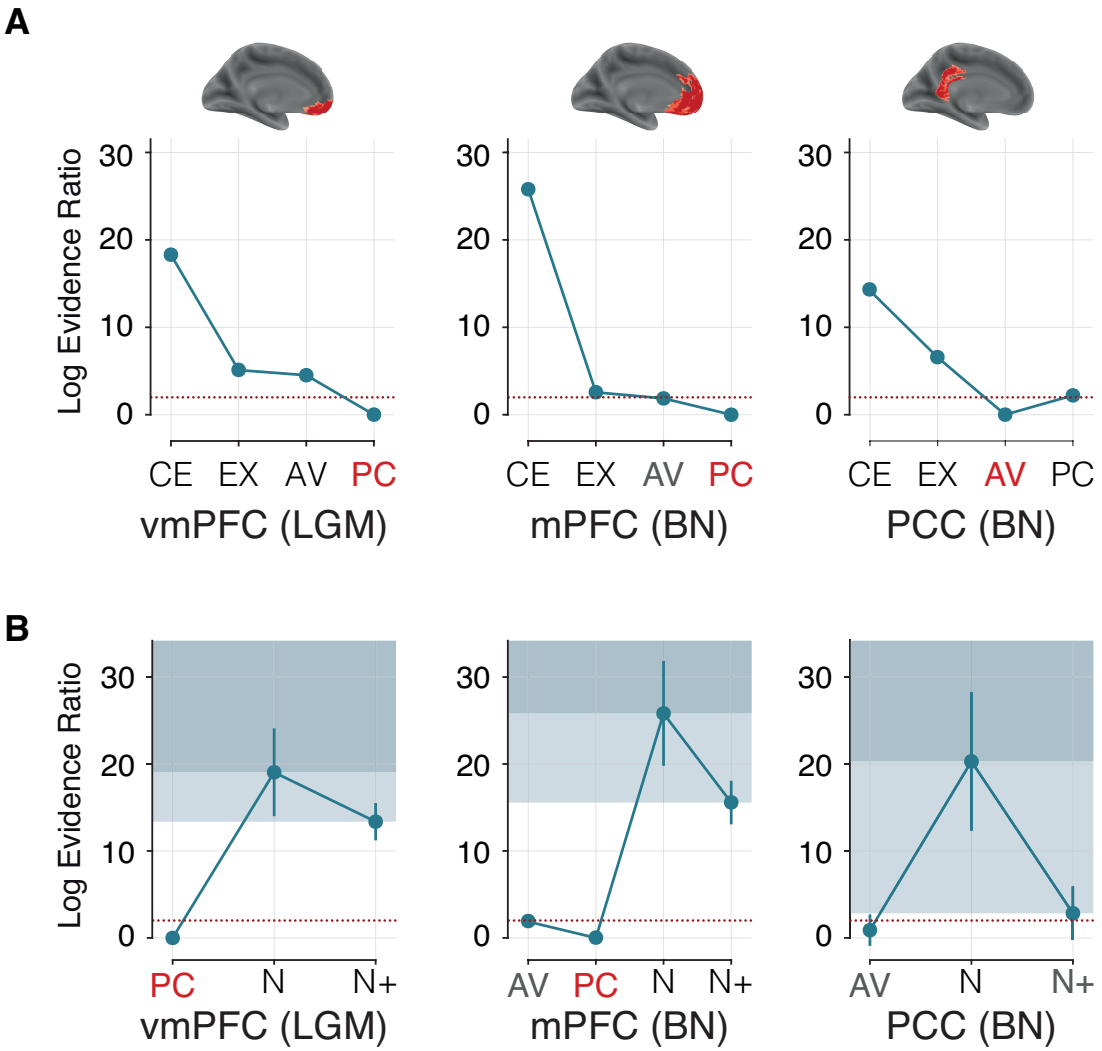
Note. mPFC = medial prefrontal cortex, HPC = Hippocampus, PCC = posterior cingulate cortex, vmPFC = ventromedial prefrontal cortex, SL = searchlight, BN = Brainnetome (Fan et al., 2016), LGM = Liu, Grady, Moscovitch (see also Benoit et al., 2019; Liu et al., 2017), CE = centrality, EX = experience, AV = affective value, PC = principal component, SE = standard error, Sig. = significance (***– $P < .001$, ** – $P < .01$, * – $P < .05$).

C.5 Correlation of the principal component in anatomical ROIs.



Asterisks denote the result of a *t*-test on the Fisher-Z transformed correlation coefficients (*** - $P_{Holm} < .001$, ** - $P_{Holm} < .01$, * - $P_{Holm} < .05$), vmPFC = ventromedial prefrontal cortex, mPFC = medial prefrontal cortex, PCC = posterior cingulate cortex, LGM = Liu, Grady, Moscovitch (see also Benoit et al., 2019; Liu et al., 2017), BN = Brainnetome (Fan et al., 2016). Box-plots: center line, median; box limits, first and third quartile; whiskers, 1.5x interquartile range.

C.6 Linear mixed effects models – Model selection in anatomical ROIs.



(**A**) Model comparison in three additional, anatomical ROIs, (**B**) Model comparisons of the winning models against random gaussian noise and sorted gaussian noise. Data points depict the mean model performance in comparisons with 1,000 random noise models (N) and sorted noise models (N+), whiskers indicate the standard deviation of the model performance. vmPFC = ventromedial prefrontal cortex, mPFC = medial prefrontal cortex, PCC = posterior cingulate cortex, LGM = Liu, Grady, Moscovitch (see also Benoit et al., 2019; Liu et al., 2017), BN = Brainnetome (Fan et al., 2016), CE = centrality, EX = experience, AV = affective value, PC = principal component, N = random noise, N+ = sorted noise.

C.7 Whole brain search light – Principal component

Region (peak)	approx. BA (peak)	Hemi-sphere	MNI (peak)			Voxel	*Z(max)*
			x	*y*	*z*		
Medial PFC	6, 8, 9	L/R	-4	32	50	9,375	5.91
			18	28	58		5.24
			10	58	30		5.16
Dorsolateral PFC	6, 8, 45	L	-52	24	8	1,189	4.79
			-38	14	48		4.72
			-44	20	40		4.24
Lateral parietal cortex	19, 39	L	-38	-60	20	1,087	4.70
			-50	-68	38		4.34
			-54	-64	30		4.30
Lateral PFC	44	R	54	18	14	420	4.68
			58	14	8		4.22
			50	10	-2		3.58
Cerebellum	-	L/R	38	-60	-40	2,005	4.44
			34	-64	-32		4.32
			34	-62	-48		4.26
posterior cingulate cortex	23	L/R	2	-44	36	1,126	4.40
			4	-46	28		4.33
			-4	-40	26		4.03
Lateral PFC	45, 47	R	38	34	-10	451	4.28
			46	30	-10		3.99
			52	30	4		3.43
Lateral parietal cortex	39	R	58	-56	28	725	4.26
			48	-46	28		4.04
			50	-54	34		3.95
Lateral temporal cortex	22, 41	R	60	-20	4	246	4.12
			60	-22	-4		3.75
Lateral temporal cortex	21	L	-62	-10	-24	59	3.99
Anterior temporal cortex	20, 21	R	56	-6	-36	224	3.94
			52	-2	-28		3.60
Temporal cortex	21, 22	L	-54	-24	-10	214	3.89
			-58	-26	0		3.51
			-54	-32	-4		3.44
Anterior temporal cortex	38	L	-50	-2	-32	58	3.84
Lateral parietal	40	L	-46	-32	14	76	3.79
			-48	-22	12		3.36
Anterior insula	47	L	-32	28	-8	134	3.77
			-38	22	-12		3.71

Region (peak)	approx. BA (peak)	Hemi-sphere	MNI (peak)			Voxel	*Z(max)*
			x	*y*	*z*		
			-30	20	-28		3.62
Thalamus	-	L	-4	-34	6	44	3.63
Lateral PFC	47	L	-48	42	-2	37	3.52
Cerebellum	-	L	-42	-66	-54	51	3.49
Ventromedial PFC	11	R	8	24	-16	32	3.43

Note. For exploratory purposes, thresholded at $P < 0.001$, uncorrected, and at least 30 contiguous voxels. BA = Brodmann area.

Curriculum Vitae

Name	Philipp Chrysostomos Paulus
Date of birth	16 July 1989
Place of birth	Stuttgart, Germany

Career

Since 2022	*Post-doctoral researcher* at the Chair of Neuropsychology (Prof. Dr. Monika Schoenauer), Institute of Psychology, University of Freiburg, Freiburg, Germany.
2016 – 2021	*PhD student* in the Max Planck Research Group: "Adaptive Memory" (Dr. Roland Benoit), Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany.
2015 – 2016	*Research fellow* at the Collaborative Research Centre 940 ("Volition and Cognitive Control"), Chair for Differential and Personality Psychology (Prof. Dr. Alexander Strobel), Dresden University of Technology, Dresden, Germany.
2013 – 2014	*Student intern* in the Comparative Emotion Research Group (Prof. Dr. Dr. Dominik Bach), Zurich University Hospital for Psychiatry, Zurich, Switzerland.
2009 – 2015	*Student* of Psychology (Diplom), Thesis: *Modeling event-related heart period responses*, Dresden University of Technology, Dresden, Germany.

Publications

Paulus, P.C.*, Meyer, A.K.*, Bernhard, I., & Benoit, R.G. (*in preparation*). Examining the impact of retrieval suppression on conditioned fear memories.

* *these authors contributed equally*

Paulus, P. C., Dabas, A., Felber, A., & Benoit, R. G. (2022). Simulation-based learning influences real-life attitudes. *Cognition*, *227*, 105202. https://doi.org/10.1016/j.cognition.2022.105202

Paulus, P. C., Charest, I., & Benoit, R. G. (2020). Value shapes the structure of schematic representations in the medial prefrontal cortex [Preprint]. *BioRxiv*, 2020.08.21.260950. https://doi.org/10.1101/2020.08.21.260950

Strobel, A., Wieder, G., **Paulus, P. C.**, Ott, F., Pannasch, S., Kiebel, S. J., & Kührt, C. (2020). Dispositional cognitive effort investment and behavioral demand avoidance: Are they related? *PLOS ONE*, *15*(10), e0239817. https://doi.org/10.1371/journal.pone.0239817

Benoit, R. G., **Paulus, P. C.**, & Schacter, D. L. (2019). Forming attitudes via neural activity supporting affective episodic simulations. *Nature Communications*, *10*(1), 2215. https://doi.org/10.1038/s41467-019-09961-w

Grass, J., Krieger, F., **Paulus, P. C**., Greiff, S., Strobel, A., & Strobel, A. (2019). Thinking in action: Need for Cognition predicts Self-Control together with Action Orientation. *PLOS ONE*, *14*(8), e0220282. https://doi.org/10.1371/journal.pone.0220282

Castegnetti, G., Tzovara, A., Staib, M., **Paulus, P. C.**, Hofer, N., & Bach, D. R. (2016). Modeling fear-conditioned bradycardia in humans. *Psychophysiology*, *53*(6), 930–939. https://doi.org/10.1111/psyp.12637

Paulus, P. C., Castegnetti, G., & Bach, D. R. (2016). Modeling event-related heart period responses. *Psychophysiology*, *53*(6), 837–846. https://doi.org/10.1111/psyp.12622

Talks and poster presentations

Paulus, P.C., Dabas, A., Felber, A., & Benoit, R.G. (2021, June 2-6). *Simulation-induced learning: Episodic simulations shape real-life attitudes* [Conference talk]. Annual meeting of the unit of "Biological Psychology and Neuropsychology" of the German Society for Psychology (DGPs), Tübingen, Germany. https://bit.ly/3mvKITN

Meyer, A.-K., Schmidt, T., Stramaccia, D.F., **Paulus, P.C.**, & Benoit, R.G. (2021, June, 2-6). *Category conditioning put to the test: A meta-analysis and successful replication* [Conference talk]. Annual meeting of the unit of "Biological Psychology and Neuropsychology" of the German Society for Psychology (DGPs), Tübingen, Germany. https://bit.ly/3lebv7S

Paulus, P.C., Dabas, A., Felber, A., & Benoit, R.G. (2020, cancelled due to Coronavirus) *Affective episodic simulations shape attitudes to elements from our everyday life*. [Poster submitted] Annual meeting of the unit of "Biological Psychology and Neuropsychology" of the German Society for Psychology (DGPs), Freiburg, Germany.

Paulus, P.C.*, Williams, A.N.*, & Benoit, R.G. (2020, March 3-4). *The structure of experience: Investigating the emergence of hippocampal versus medial prefrontal representations of a complex and novel environment* [Poster presentation, due to the Coronavirus held online]. Scientific Advisory Board evaluation of the Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. http://bit.ly/3BarHN5

** these authors contributed equally*

Paulus, P.C., Charest, I., & Benoit, R.G. (2020, March 3-4). *Revealing the structure of affective schematic representations in medial prefrontal cortex* [Poster presentation, due to the Coronavirus held online]. Scientific Advisory Board evaluation of the Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. http://bit.ly/3BarHN5

Paulus, P.C., Charest, I., & Benoit, R.G. (2020, May 2-5). *Revealing the structure of affective schematic representations in medial prefrontal cortex* [Poster presentation]. Annual meeting of the Cognitive Neuroscience Society (CNS), Boston, United States of America. http://bit.ly/3os4Xoc

Paulus, P.C. (2020, April 30). *Model-based analysis of heart period responses* [Talk in the online lecture series "Introduction to Psychophysiological Modelling"]. Psychophysiological Modeling. https://bit.ly/2Ys3h3r

Paulus, P.C., Charest, I., & Benoit, R.G. (2019, June 20-22). *Revealing the structure of affective schematic representations in medial prefrontal cortex* [Conference talk and co-chair in the symposium "Structured representations in the human brain"]. Annual meeting of the unit of "Biological Psychology and Neuropsychology" of the German Society for Psychology (DGPs), Dresden, Germany.

Paulus, P.C. (2019, June 5-6) *Model-based analysis of heart period responses* [Talk and host of the workshop "Introduction to Psychophysiological Modelling"]. International Max Planck Research School NeuroCom, Leipzig, Germany. https://bit.ly/2YhA9eD

Paulus, P.C., Charest, I., & Benoit R.G. (2018, June 26 - 28). *Episodic simulations reveal the structure of affective representations in ventromedial prefrontal cortex* [Poster presentation]. 8th IMPRS NeuroCom Summer School, Leipzig, Germany. https://bit.ly/3CXw7Zv

Paulus, P.C., Charest, I., & Benoit, R.G. (2018, Mai 30 - June 02). *Episodic simulations reveal the structure of affective representations in ventromedial prefrontal cortex* [Poster presentation, awarded with a poster prize]. Annual meeting of the unit of "Biological Psychology and Neuropsychology" of the German Society for Psychology (DGPs), Gießen, Germany. https://bit.ly/3wbrHfT

Paulus, P.C. (2018, Mai-June 30-2). *Model-based analysis of heart period responses* [Talk in the workshop "Introduction to Psychophysiological Modelling"]. Young researchers of

the unit of "Biological Psychology and Neuropsychology" of the German Society for Psychology (DGPs), Gießen, Germany. https://bit.ly/3wbrHfT

Paulus, P.C., Charest, I., & Benoit, R.G. (2018, March 24-27). *Episodic simulations reveal the structure of affective representations in ventromedial prefrontal cortex* [Poster presentation]. Annual meeting of the Cognitive Neuroscience Society (CNS), Boston, United States of America. http://bit.ly/3os4Xoc

Paulus, P.C., Charest, I., & Benoit R.G. (2017, July 12-14). *Affective representations in medial prefrontal cortex* [Poster presentation]. 7th IMPRS NeuroCom Summer School, London, UK. https://bit.ly/3q5WJBY

Paulus, P.C., Castegnetti, G., & Bach, D.R. (2016, July 4-6). *Modeling event-related heart period responses* [Poster presentation]. 6th IMPRS NeuroCom Summer School, Leipzig, Germany. https://bit.ly/3u2vlGx

Paulus, P.C. (2014, May 6). *Development of a heart-period response model* [Invited talk]. Emotion Club, Comparative Emotion Research Group, Prof. Bach, Zurich University Hospital for Psychiatry, Zurich, Switzerland.

Selbständigkeitserklärung

Hiermit versichere ich, Philipp Chrysostomos Paulus, geboren am 16. Juli 1989 in Stuttgart,

- dass die vorliegende Arbeit ohne unzulässige Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde und dass die aus fremden Quellen direkt oder indirekt übernommenen Gedanken in der Arbeit als solche kenntlich gemacht worden sind;

- dass weitere Personen bei der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt waren, insbesondere auch nicht die Hilfe eines Promotionsberaters in Anspruch genommen wurde und dass Dritte von dem Antragsteller weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;

- dass die vorgelegte Arbeit in gleicher oder in ähnlicher Form keiner anderen wissenschaftlichen Einrichtung zum Zwecke einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt und auch veröffentlicht wurde;

- dass keine früheren, erfolglosen Promotionsversuche stattgefunden haben.

Philipp Chrysostomos Paulus
Freiburg i. Br., den 4. April 2022