Violation of Expectation in Naturalistic Stimuli

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Vincent Plikat

aus Essen

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	20.02.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Andreas Bartels
2. Berichterstatter/-in:	Prof. Dr. Hartmut Leuthold

Contents

List of Figures III

Acknowledgment 2

Abbreviations 5

Abstract 6

Zusammenfassung 7

1 Introduction 8

1.1 Expectations, Experience and Memory 8

1.2 Predictive Coding 9

1.2.1 The Feed Forward Model of Perception 10

1.2.2 A Hierarchy in Predictions 10

1.2.3 Testing Prediction Errors 12

1.3 Magic in Research 13

1.3.1 Magic and Predictive Coding 14

1.4 The Memory Color Effect 15

1.4.1 The Memory Color Effect during Ambiguous Visual Information 16

1.4.2 Cognitive Penetrability of the Memory Color Effect 18

1.5 Scientific Scope 18

2 Research Aims 19

3 Results and Discussion 21

3.1 Hierarchical Surprise Signals in Naturalistic Violation of Expectations 21

3.1.1 Introduction and Design 21

3.1.2 Results 22

3.1.3 Conclusion 28
3.2 How Can We be Moved by Magic 29
3.2.1 Epistemic Emotions 30
3.2.2 The Paradox of Theatrical Magic 30
3.2.3 Fractured Beliefs 31
3.2.4 Conclusion 32
3.3 Memory color influences conscious object perception 33
3.3.1 Introduction and Design 33
3.3.2 Results and Discussion 34
3.3.3 Conclusion 36

4 Grand Discussion **37**

5 Outlook **40**

Bibliography **43**

Publications **63**

List of Figures

3.1 Experimental Design 22
3.2 Univariate Contrasts 24
3.3 Multivariate Results 27
3.4 Stimuli and Experimental Design 34
3.5 Results for color diagnostic objects 36

Acknowledgment

Meine Arbeit wurde von der Barbara-Wengeler Stiftung für zwei Jahre mit einem Stipendium finanziert. Für diese finanzielle Unterstützung und Unabhängigkeit bin ich zutiefst dankbar. Sie hat mir erlaubt mich voll und ganz auf meine Arbeit zu konzentrieren und gab mir darüber hinaus den Anstoß mich eingehender mit der Philosophie zu beschäftigen, was meine Arbeit sehr bereichert hat.

Ich möchte mich bedanken bei all meinen Freunden und bei meiner Familie, die mich während der Jahre der Promotion begleitet haben. Es war nicht immer leicht, eine Promotion hat Höhen und Tiefen - danke, dass ihr bei beiden dabei wart.

Des Weiteren gebührt Prof. Dr. Andreas Bartels mein Dank dafür, dass ich bei ihm spannenden, abwechslungsreichen und auch unkonventionellen wissenschaftlichen Fragen nachgehen konnte.

Bei dem Rest der AG Vision and Cognition möchte ich mich für die schöne Zeit, die ich mit euch hatte, bedanken. Mit euch hat nicht nur das *woran* ich gearbeitet hab Spaß gemacht, sondern auch alles andere drum herum.

Mein größter Dank jedoch muss an Dr. Pablo Grassi gehen. Von dir habe ich fast alles während meiner Promotion gelernt, du hast mich in jedem Moment unterstützt, gefordert, gefördert und warst dabei noch ein guter Freund!

Ich durfte über Zauberei promovieren. Wer hätte da “nein” gesagt?

Abbreviations

ACC	Anterior Cingulate Cortex
BR	Binocular Rivalry
CN	Caudate Nucleus
DAN	Dorsal Attention Network
DMN	Default Mode Network
FEF	Frontal Eyefield
FFA	Fusiform face area
FFG	Fusiform gyrus
fMRI	functional Magnetic Resonance Imaging
IPS	Intraparietal sulcus
LOC	Lateral occipital complex
MCE	Memory color effect
MVPA	Multivariate Pattern Analysis
OFA	Occipital face area
PC	Predictive Coding
PCC	Posterior Cinguclate Cortex
PCUN	Precuneus
PFC	Prefrontal Cortex
POS	Parieto-occipital sulcus
PPA	Parahippocampal place area
PPC	Posterior Parietal Cortex
ROI	Region of Interest
SFG	Superior Frontal Gyrus
V1	primary visual cortex
V2	secondary visual cortex
V3	tertiary visual cortex
VOE	Violation of Expectation

Abstract

Our visual perceptual system receives information about the outside world solely in the form of light, that projects a two-dimensional image of the three-dimensional world on our two retinas. However, one three-dimensional scene can result in infinitely different two-dimensional images and one two-dimensional image can be the result of many different three-dimensional sources. In order to veridically and accurately represent the outer world, our perceptual system relies on memory-based predictions about the true source of sensory information. These predictions can either result from recent events, forming precise and specific expectations or from consistent experiences we make over a longer period of time, forming abstract concepts that give less precise and less specific expectations. To investigate the workings of our perceptual system, it is thus fertile to *violate* the predictions made by an observer to see which parts of the brain are involved in processing the predicted stimulus. However, since it is more difficult to violate predictions based on abstract concepts, scientific investigations of prediction errors mostly focused on violating predictions based on newly learned associations or patterns. The aim of my doctoral thesis is to fill this gap by investigating the influence of violations of expectations, that are based on the most consistent experiences we make throughout our lives. Inspecting these violations, I gained insight about how predictions modulate sensory processing, our conscious experience and, in specific cases, why we feel the way we feel, when confronted with such violations of expectations.

Zusammenfassung

Unser visuelles Verarbeitungssystem erhält Informationen über die Welt nur in Form von Licht, dass zwei-dimensionale Bilder der drei-dimensionalen Welt auf unsere Retinas projiziert. Allerdings kann eine drei-dimensionale Szene unendlich viele verschiedene zwei-dimensionale Bilder erzeugen und ein zwei-dimensionales Bild kann durch die Projektion unendlich vieler verschiedener drei dimensionaler Szenen entstanden sein. Um die Welt wahrheitsgetreu zu repräsentieren, muss unser visuelles Verarbeitungssystem erinnerungsbasierte Vorhersagen über den Ursprung des Retinabildes nutzen. Diese Vorhersagen können das Resultat von jüngsten Ereignissen sein, die spezifische und genaue Erwartungen formen, oder sie können das Resultat von vielen konsistenten Erfahrungen sein, die abstrakte Konzepte formen. Um unser visuelles Verarbeitungssystem zu erforschen, ist es daher nützlich die Vorhersagen eines Betrachters zu *verletzten*, um zu sehen welche Teile des Gehirns den erwarteten Stimulus mitverarbeiten. Da es allerdings schwierig ist Erwartungen basierend auf abstrakten Konzepten zu verletzten, lag der Fokus der Forschung eher darin Erwartungen basierend auf neuen Informationen zu verletzten. Das Ziel meiner Arbeit ist es diese Lücke zu schließen und den Einfluss von Vorhersagefehler basierend auf abstrakten Konzepten zu erforschen. Meine Arbeit hat gezeigt, dass diese unerwarteten Ereignisse Verarbeitungsprozesse und unsere bewusste Wahrnehmung beeinflussen und in speziellen Fällen sogar erklären können, warum wir uns fühlen, wie wir uns fühlen, wenn wir mit solchen Ereignissen konfrontiert sind.

1 Introduction

1.1 Expectations, Experience and Memory

Expectations, experience, and memory guide and influence us throughout our lives. One example: I expect my friend to come over and visit me - so I clean my apartment. Over time, I made the experience that my friend is often late - so I decide I have time to go buy groceries. I remember that I used all the milk in the morning - so I buy a new one.

These rather trivial examples illustrate how past events shape our actions; however, expectations, experience, and memory also influence how we perceive the world in a much more subtle way. As perception is an automatic process, we take it very much for granted, however, the fact that we constantly perceive a vivid and clear three-dimensional image of the world is an extraordinary feat.

This might become clearer considering the following three aspects of perception (only focusing on the visual domain): First, the visual system is confronted with a plethora of information that is at the same time noisy and incomplete. Second, the three-dimensional world needs to be reconstructed from light projected onto our two-dimensional retinas. Lastly, the mind needs to infer which external cause created the visual sensation, even though every visual input can be generated by an infinite number of external causes (Kersten & Yuille, 2003) and one external cause can lead to numerous different visual inputs. This poses perception as an ill posed problem of inverse inference (Bertero et al., 1988; Di Lollo, 2012; Pizlo, 2001; Spratling, 2017).

How the human mind deals with this problem has been a matter of debate for centuries, but it is assumed that it incorporates memory-based information to narrow down the possible external sources of sensory input.

The idea that perception is engaged in some form of unconscious inference (as already proposed by Helmholtz 1867) gained new popularity in cognitive (neuro-) science with the predictive coding framework.

1.2 Predictive Coding

In 1999 Rao and Ballard published their seminal work "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects" in which they showed, using a computational model for visual processing, that a neural network implementing predictive feedback from higher to lower levels and prediction error signals from lower to higher levels, creates neurons with extra-classical properties found in the primary visual cortex V1 (Bolz & Gilbert, 1986; Hubel & Wiesel, 1968).

Following this finding, many papers argued how predictive feedback could explain behavioral and neuroimaging findings (Auksztulewicz & Friston, 2016; Baldeweg et al., 2004; Friston, 2002; Kok & de Lange, 2014; Kok et al., 2012; Kumar et al., 2011; Muckli, 2010; Murray et al., 2002; Schellekens et al., 2016; Wolpert & Flanagan, 2001) and lead to the development of a predictive coding framework suggesting that the brain is constantly engaged in Bayesian inference. The basic idea behind this framework is that the brain constructs a generative model of the world that constantly predicts incoming sensory information, compares prediction with current input and uses the difference (prediction error) to constantly update the model using Bayesian inference (Clark, 2013), thereby laying the fundamental mechanism for learning (Friston & Kiebel, 2009; Friston, Karl, 2005; Schultz et al., 1997).

The goal of the brain, according to the predictive coding framework, is to reduce the prediction error and thereby keep the world predictable and the internal processes efficient (Ali et al., 2022). In some iterations this framework even aims to not only explain perception but to unify perception, action, emotion, attention, consciousness and even more (Clark, 2013; Friston, 2008, 2009, 2010; Hohwy, 2012).

The predictive coding framework claims that the brain constantly generates a model of the world and only uses sensory information to compare it and to update the model. The way the predictive coding framework understands the process of

perception is thus orthogonal to the traditional feed-forward model of perception.

1.2.1 The Feed Forward Model of Perception

The traditional view on perception was that it is the result of sensory information propagating up the hierarchy of processing stages in a feed-forward manner, that responds to increasingly complex, larger, and specific stimuli (Felleman & Van Essen, 1991; Hubel & Wiesel, 1962; Riesenhuber & Poggio, 1999; Thorpe et al., 1996).

The hierarchy follows a posterior-anterior gradient. With neurons responding to low-level features like contrast and edges in posterior occipital areas such as primary visual cortex (V1) (Movshon et al., 1978; Tootell et al., 1998) over to more anterior areas responding to more complex features, such as color in V4 (Bartels & Zeki, 2000; Zeki et al., 1998) or motion in V5 (Bartels et al., 2008; Zeki, 2015), to areas responsive to high-level visual stimuli such as faces in the fusiform face area (FFA)(Kanwisher et al., 1997) or facial features in the occipital face area (OFA) (Ambrus et al., 2017; Gschwind et al., 2012), objects in lateral occipital complex (LOC) (Grill-Spector et al., 2001; Kourtzi & Kanwisher, 2001), tools in the intraparietal sulcus (IPS) (Mahon et al., 2010; Mruczek et al., 2013), or entire scenes in the parahippocampal place area (PPA) (Epstein & Kanwisher, 1998).

At some stage (or possibly multiple stages) of this hierarchy the stimulus is (consciously) perceived. In the traditional view, top-down predictive signals were thought of as an enhancing signal to facilitate the feed-forward processing of expected stimuli (Ekman et al., 2017).

1.2.2 A Hierarchy in Predictions

Analogous to the hierarchy in visual processing, from small and simple to larger and more complex stimuli, there is a proposed hierarchy of predictions. This hierarchy follows a gradient of predictions that are mostly specific and precise to predictions that are mostly abstract and vague.

Consider someone activated a pendulum - the exact frequency and amplitude are completely unknown to you, but just after a few oscillations you have a very precise prediction of how the pendulum will swing and therefore how the edges and the contrast (processed at early stages of the visual hierarchy) in the scene will

change. Now consider something more visually complex, say the face of a person you have known for a while. When seeing that person, you have certain predictions of how the face will look, but these predictions are less precise - are you seeing them from afar or nearby, from the side or up front, is the person smiling, frowning, or blinking? And in everyday life you are, most of the time, in highly complex and specific environments. While being indoors you might, for example, expect to see chairs, tables, or people but only have a vague expectation of their specific position, orientation, color, etc. Being outdoors, you deem a car passing you more or less likely depending on whether you are in the woods or on a street (Peelen et al., 2024).

At this point, it is important to note that the precision of predictions additionally relies on the consistency of the events. Being on a street very often predicts seeing cars, seeing corn almost always predicts perceiving the color yellow and seeing an unsupported heavy object always predicts seeing it fall to the ground.

The hierarchy of predictions described above can be tightly linked to the separation between time-variant aspects of experience and time-invariant aspects of experience. This division entails a useful implication that helps us understand how our perception is shaped and inferred. Time-variant aspects of experience come in the form of changes in our first-person experience, like moving objects or changes in our visual field due to head or eye movements. Time-invariant aspects of experience are more abstract and focus more on states of the world that do not depend on how the world changes at the exact moment, like the arrangement of faces or that leaves grow on trees. The predictive coding framework suggests that the higher, more abstract expectations set constraints on the more concrete predictions (Friston & Kiebel, 2009; Roark & Holt, 2022).

Let us again consider the example of seeing a person: Seeing a person creates the prediction of seeing a face in form of a top-down signal to face selective areas, such as the FFA (Kanwisher et al., 1997; Parvizi et al., 2012; Xu, 2005). The expectation of a face in turn creates the prediction of facial features (eyes, nose, mouth) to a lower visual areas like the OFA (J. Liu et al., 2010) and the expectation of facial features creates predictive signals to even lower visual areas regarding, e.g., color, edge orientation or contrast, processed in the earliest stages of visual processing

(Hubel & Wiesel, 1959; Movshon et al., 1978; Tootell et al., 1998).

1.2.3 Testing Prediction Errors

Following the logic of top-down prediction signals shaping our perception, investigating the process of perception and predictions made by the predictive coding framework can thus be done by violating predictions made by an observer. These investigations have traditionally focused on violating time-variant regularities such as repetition suppression (Grotheer & Kovács, 2016; Todorovic et al., 2011), odd-ball paradigms (Bubic et al., 2010; Modirshanechi et al., 2019) or newly learned associations (Iglesias et al., 2013). However, there has recently been an increasing interest in investigating time-invariant regularities (Danek et al., 2015; Fischer et al., 2016; Parris et al., 2009; Paulun et al., 2017; Pramod et al., 2022; Schwettmann et al., 2019). This focus on time-variant experiences is, from an experimental perspective, understandable since 1) they are easier to generate, manipulate, and parametrizise, and 2) it is difficult to manipulate time-invariant regularities in a controlled experimental set-up. The more predictions are based on slow regularities, the more they represent abstract concepts that need to be defined, identified, and then manipulated, ideally using naturalistic stimuli (one's prediction about abstract concepts might not hold in artificial environments like a computer simulation).

So how can we investigate the influence of predictions based on time-invariant aspects on visual processing and further on the conscious percept? Here I want to suggest two ways of experimental manipulation that violate expectations based on time-invariant experiences. One way would be to use deception to show people events that seemingly violate the laws of physics. Another way could be to violate strong object-color associations - so-called color diagnostic objects (Oliva & Schyns, 2000; Tanaka & Presnell, 1999) to violate object-feature expectation learned naturally throughout our lives.

In the next two subsections I will discuss the advantages of using deception in the form of magic tricks to investigate how violations of deeply held beliefs about the physical world are processed in the brain and the advantages of using color diagnostic objects to investigate the influence of predictions on conscious perception.

1.3 Magic in Research

One of the oldest performance art, that seems to be ubiquitous in human culture is magic (Macknik et al., 2008). However, compared to other forms of art like music, painting, or architecture, it has been largely neglected by science (though there were few exceptions in the past arguing to include magic in the scientific discourse, however with little effect - Binet, 1894; Dessoir, 1893; Jastrow, 1897; Triplett, 1900). On the one hand, this is not surprising since magicians and the magic community as a whole try to keep details of their workings secret (Luhrmann, 1989). On the other hand, it is astonishing that the scientific community did not take advantage of the insights into human cognition magicians gained over many generations and centuries to manipulate our attention, memory, and choices.

It has been argued that magicians follow a similar pattern of work as scientists (Kuhn et al., 2008): Based on experience, luck, intuition, or work of their peers they build a hypothesis on how they can deceive their audience and thereby create a new routine. By trial and error, they refine their hypothesis until the routine works.

This procedure has led to performances that manipulate how we perceive, attend and also the way we memorize these events. One could even argue that the famous phenomena of change blindness (Rensink et al., 1997; Simons & Levin, 1997) and inattentional blindness (Simons & Chabris, 1999) would have been discovered earlier if scientists had cooperated with the magic community, as objects appearing, changing or vanishing in plain sight without the spectator noticing until attention is drawn to it is not an uncommon thing in magic performances.

At the start of the 21st century, however, there was a rise in interest towards magic within the scientific community with different views on how exactly magic could best be used to gain new scientific insights (Kuhn et al., 2008; Lamont, 2017; Lamont et al., 2010; Macknik et al., 2008; Macknik & Martinez-Conde, 2009; Quiroga, 2016). One approach suggested gathering information about the workings of magic tricks and trying to explain them with known cognitive mechanisms. Any magic trick that cannot be explained, points towards an unknown cognitive mechanism that can be investigated (Kuhn et al., 2008). Following this idea a number of publications explained magic tricks in terms of cognitive mechanisms. So, the famous rope cutting routine in which a rope is cut in half and magically restored was connected to

Gestalt principles (Barnhart, 2010) and concealing to put away/fetching an object by giving the movement a different purpose was connected to the principle of exclusive allocation of movements to intentions (Van de Cruys et al., 2015), to name just two. Another approach was to just use magic tricks in scientific experiments to add evidence to known cognitive mechanisms, such that people usually follow the gaze of a person they pay attention to (Barnhart & Goldinger, 2014; Kuhn & Land, 2006; Kuhn & Rensink, 2016; Kuhn & Tatler, 2005; Kuhn et al., 2009; Quiroga, 2016), that the Einstellungs effect - a given solution to a problem blocks exploration for other solutions (Bilalić et al., 2008a, 2008b; Luchins, 1942) - works stronger than previously expected (Thomas & Didierjean, 2016) or that attention is reduced shortly after blinking (Wiseman & Nakano, 2016).

1.3.1 Magic and Predictive Coding

However, in light of the predictive coding framework discussed above, magic tricks pose a much more interesting potential. As described, the predictive coding framework suggests that statistical regularities on different timescales (time-in/variant) create predictions on different levels of abstraction (Honey et al., 2012; Jakob Hohwy, 2013; Tenenbaum et al., 2011), further the number and consistency of regularities define the strength of a prediction (Grassi & Bartels, 2021; Hsu & Hämäläinen, 2021).

Over our entire lifespan we are confronted with the laws of physics - objects are drawn down by gravity, solid objects cannot penetrate each other without breaking and objects do not appear out of nowhere or vanish into thin air. This, however, is exactly what is seemingly accomplished by magicians in their tricks.

Let us consider the experience of a simple magic trick within the Bayesian predictive coding framework - the appearance of a coin in a previously (seemingly) empty hand: A magician shows you their empty hands and that they do not have anything up their sleeves. They close their hand and upon reopening the hand, there is a coin.

In Bayesian predictive coding terms, your brain constructs a prior of what to see when the magician opens their hand (Grassi & Bartels, 2021). This includes, but is not restricted to, the color of the skin, the lines on their hands, the specific orien-

tation of fingers, and so on. The prediction is then compared to the actual sensory input. The occurrence of the coin creates a prediction error, which is propagated up the visual hierarchy to update the generative model from "the hand is empty" to "there is a coin in the hand". Since your prior "objects do not appear out of nowhere" is very strong, the prior probability of perceiving an object in the previously empty hand is extremely low (close to zero). Further, as the magician performed the trick in bright daylight and did not perform any suspicious or fast movements, the confidence in the sensory data is also very high. Then seeing the coin entails, on the one hand, a strong prediction error signal (Grassi & Bartels, 2021), that should be easily detectable using neuroimaging techniques, and on the other hand, a strong sensation of surprise.

Here it becomes clear that the experience of such an event depends on prior, memory-based expectations. Seeing a coin in a hand is not particularly surprising or interesting to see. Only if we had a strong memory-based expectation of the hand being empty, will we be surprised.

Given that magic tricks violate expectations based on the most abstract and consistent regularities (laws of physics), they provide a great means to investigate how our memory and experience based intuitive understanding of physics influences visual processing in a top-down manner.

1.4 The Memory Color Effect

As already discussed in this thesis, the predictive coding framework states that our perception is altered by our memory- and experience-based expectations. A generative model incorporates noisy and ambiguous information with memory-based expectations to represent the most likely external cause for our sensory input (Clark, 2013; Friston, 2008, 2018; Hohwy, 2012).

A central question is to which extent such processes affect conscious visual perception beyond their known effects on neural processing. On the one hand, predictions can support processing in noisy conditions and boost perception matching the predictions. On the other hand, novelty ought to attract attention and perception to aid learning (Press et al., 2020). An excellent model case to study the influence of

memory-based expectations on perception is the memory color effect (MCE) (Hering, 1920). Objects with strong color associations (e.g., bananas are yellow, grass is green, the sky is blue, etc.) alter how we perceive these objects, such that objects strongly associated with a specific color appear tinted in their associated color when presented objectively gray (Hansen et al., 2006; Kimura et al., 2013; Olkkonen et al., 2008; Witzel et al., 2011). This object-color association has further been shown to affect other cognitive processes like categorization (Mitterer & de Ruiter, 2008), object identification (D. E. Lewis et al., 2013; Teichmann et al., 2020), scene recognition (Oliva & Schyns, 2000) and visual search (Cutler et al., 2024).

Seen within the predictive coding framework, the actual sensory data (e.g., an objectively gray banana) was combined with the strong prior expectation (bananas are yellow) and resulted in the highest posterior probability for the hypothesis that "a lightly yellow dyed banana" was the cause for the sensory input. This is in line with neuroimaging studies showing that an object's associated color can be decoded in the early visual cortex, even when the object is presented in gray (Bannert & Bartels, 2013) or an ambiguous color (Vandenbroucke et al., 2016).

1.4.1 The Memory Color Effect during Ambiguous Visual Information

While the MCE has been investigated extensively, how it modulates conscious perception during ambiguous visual information, however, is yet to be discovered.

An ideal way to do so is to use binocular rivalry (BR), as it is a spontaneous automatic process in which conscious perception alternates between two incompatible images presented to both eyes (Logothetis et al., 1996; Tong et al., 2006; Wheatstone, 1997). BR therefore allows to investigate cognitive influences on perception as it gives the opportunity to keep confounding low-level visual factors (luminance, contrast, etc.) equal and minimize high-level cognitive confounding factors (attention, memorability, familiarity, etc.).

Consider MCE does significantly modulate perceptual dominance for color diagnostic objects during BR, another question still remains: which color will be more dominant? This question relates to the perceptual prediction paradox (Press et al., 2020), which contrasts two core principles that our perceptual system tries to fol-

low. On the one hand it has to use the noisy and incomplete sensory information to represent the outside world veridically by incorporating prior expectations, on the other hand it has to be informative, by highlighting events that deviate from our prediction.

In the case of BR (inherently noisy and ambiguous), following Hohwy et al. (2008) the expected input should be favored, as it has the higher posterior probability. There is a plethora of empirical evidence supporting this view, showing that predictions based on motion (Attarha & Moore, 2015; R. Denison et al., 2011; Hu et al., 2021), spatial frequency (Baker & Graf, 2009), priming (Mitchell et al., 2004) or cross-modal cues (Conrad et al., 2010; Lunghi & Alais, 2013; Lunghi et al., 2010; Lunghi et al., 2017; Ono et al., 2022; Zhou et al., 2010; Zhou et al., 2012) favor the expected stimulus. However, there is a small number of studies showing that the unexpected stimulus is preferred during BR, like an incongruent object in a scene (Mudrik et al., 2011; Zacharia et al., 2020) or an unexpected natural scene in a learned sequence (R. N. Denison et al., 2016).

This difference in modulation of the dominant percept might be related to the stimulus complexity and hence where in the visual hierarchy it is processed. The studies showing dominance for the expected stimulus used simplistic stimuli, such as gratings, moving dots or simple shapes like polygons, whereas the studies showing dominance for the unexpected stimulus used naturalistic images, with either an incongruent object within the scene (Mudrik et al., 2011; Zacharia et al., 2020) or an unexpected scene within a learned sequence (R. N. Denison et al., 2016).

It is still widely debated where exactly in the brain rivalry is resolved. It has long been argued that it happens as early as in monocular cells in primary visual cortex (Blake et al., 1980; Lee & Blake, 2004) (or even earlier in LGN Wunderlich et al., 2005). However, increasing evidence suggests that not only higher areas of the brain are involved (Dwarakanath et al., 2022; Leopold et al., 1999; Sterzer & Rees, 2008; Zaretskaya et al., 2010), but also stimulus specific areas are engaged during rivalry (Kim et al., 2020; Tong et al., 1998; Zaretskaya & Bartels, 2013).

If we assume that the stimulus with the strongest neural representation is dominant during rivalry, both outcomes (dominance for the expected or unexpected color) are possible. On the one hand, it has been shown that expectations boost

representation of stimulus features in early visual cortex (Kok et al., 2012). On the other hand, expectations dampen representation of object identity in higher visual areas along the ventral visual stream (Richter et al., 2018; Richter et al., 2024).

I therefore remain naive as to how object-color associations (if at all) modulate the conscious percept during ambiguous visual input.

1.4.2 Cognitive Penetrability of the Memory Color Effect

Asking whether the MCE has an influence on the conscious percept during BR tips into another ongoing debate, triggered by Firestone and Scholl (2016).

Firestone and Scholl (2016) argue that the MCE (and top-down modulations in general) do not modulate perception per se. The critique is that top-down signals do not actually change *how* we perceive an object, but merely influence our memory, attention, or our ability to recognize an object (Firestone & Scholl, 2016; Valenti & Firestone, 2019).

The authors avoid discussing top-down modulation, such as reward, attention or familiarity of perceptual dominance during BR as a case of cognitive penetrability. Nonetheless, or even specifically because of that, it is worth investigating whether the MCE modulates conscious perception during ambiguous visual information as it adds important evidence regarding the cognitive penetrability of perception.

1.5 Scientific Scope

My doctoral thesis aims to investigate how strong expectations based on long lasting experience shape the processing and perception of naturalistic stimuli. By violating these expectations, I aim to investigate how prediction errors modulate neural activation and conscious perception.

2 Research Aims

In the last few decades, a large number of studies have been conducted in context of predictive coding approaches to perception. However, most of this empirical research focused on the generation and violation of lower-level predictions using simple behavioral paradigms usually involving sequentially presented artificial stimuli, such as repetition suppression (Grotheer & Kovács, 2016; Todorovic et al., 2011), odd-ball paradigms (Bubic et al., 2009; Modirshanechi et al., 2019) or newly learnt associations (Iglesias et al., 2013). Indeed, very few studies used ecologically valid stimuli to test predictive coding hypotheses. There are however known differences between the more abstract, artificial and simplistic stimuli used in experimental settings compared to more natural scenes and real-life situations (Bartels & Zeki, 2004; Güçlütürk et al., 2018; Schmuckler, 2001; Sonkusare et al., 2019).

Moreover, far less experiments focused their research on the investigation of higher-level priors based on our general knowledge about the world, (e.g., the light from above prior, the bigger-is-heavier prior, gravity, etc.), because it is often difficult to exploit and manipulate this prior knowledge in a controlled ecologically valid experimental setting. In this thesis I used naturalistic stimuli to induce violations of expectations that are formed through our daily life experience.

Study I

In my first study I used videos showing either magic tricks or control videos in an fMRI experiment. Using this approach, I investigated how violations of expectations based on our intuitive understanding of physics are processed using naturalistic, real-world stimuli.

Study II

In a psychologically informed philosophical discussion, I explored the paradox of theatrical magic. The paradox asks the question of why we experience a cognitive

incongruity in the light of a magic performances when we can easily "explain" away everything that happens on stage with our knowledge that what we believe to observe (e.g., someone is cut in half) is not real, but an illusion.

Study III

The last study conducted in this thesis is concerned with the question of how strong object-color associations shape our conscious perception during ambiguous visual information. To this end I conducted a binocular rivalry experiment showing color diagnostic objects in their congruent color on one eye and in their incongruent color to the other eye.

3 Results and Discussion

The focus of my doctoral thesis laid on the experience of unexpected visual input. I therefore investigated three different aspects of it. First, I performed a neuroimaging experiment with magic tricks as stimuli to see how unexpected events are processed in the brain (Plikat et al., 2025). Secondly, the question arose why we experience surprise during the experience of a magic trick, which I discussed in philosophical terms (Grassi et al., 2023). Lastly, I wanted to investigate how the experience of unexpected visual input changes our conscious perception by presenting objects with a strong color association in a binocular rivalry experiment (in review). In the following section I will summarize and discuss the three different projects' approaches.

3.1 Hierarchical Surprise Signals in Naturalistic Violation of Expectations

3.1.1 Introduction and Design

Intuitive knowledge about the laws of physics guides our actions throughout our lives. We know solid heavy objects do not float in the air and even toddlers learn within the first year of their life that objects do not appear out of nowhere or vanish into thin air (Simon et al., 1995; Wynn, 1992).

Predictions about what we are supposed to perceive are thus informed by our intuitive understanding of physics - seeing an object thrown up in the air creates the prediction to see the object coming down again. In an fMRI study I used videos of a magician performing either magic tricks or comparable control actions in order to find the neural correlates to violations of deeply held beliefs. The magic tricks showed either an object appearing out of nowhere, changing the color, or vanishing

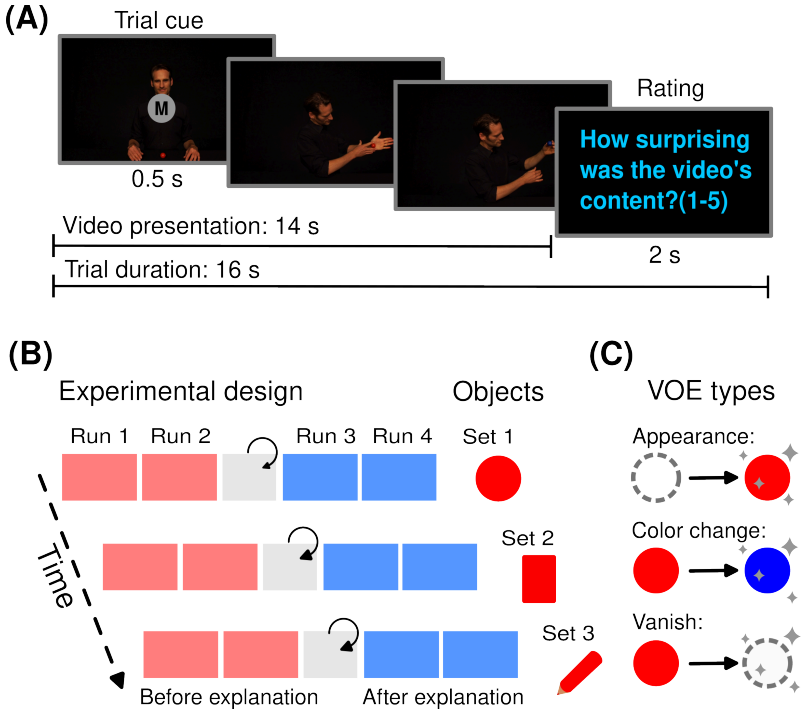


Figure 3.1: **(A)**, timeline of a single trial. Every trial started with a cue telling subjects whether the video will show a magic trick (M) or not(X). Video presentation lasted 14 sec and response time was 2 sec. **(B)**, experimental design: The experiment consisted of three sets, each associated with one object. Each set had four runs, between the second and third run videos were shown explaining how the magic tricks were performed. **(C)**, different types of magic tricks: an object appearing, changing color or vanishing.

into thin air. In an additional step I manipulated subjects' prior knowledge of the underlying mechanisms. This led to a 3 (appear, change, vanish) × 2 (magic, control) × 2 (with, without prior knowledge) repeated measures design (see Fig. 3.1 for a detailed depiction of the experiment)

3.1.2 Results

Univariate Results

As a first question I wanted to know which regions in the brain respond to violations of expectations (VOEs) based on our intuitive understanding of physics. To answer this question, I performed a univariate comparison between the BOLD responses to magic tricks vs control videos before subjects learned how the tricks were performed. This analysis revealed a large network of frontal and parietal areas, including the anterior and posterior cingulate cortex (ACC and PCC respectively), medial prefrontal cortex (mPFC), superior frontal gyrus (SFG) and posterior parietal cortex (PPC), but also the subcortical structures Thalamus and Caudate Nucleus (Fig. 3.2

A).

To refine the analysis and find areas that respond to violation of intuitive physics regardless of which expectation was violated specifically, I performed a univariate conjunction analysis, contrasting each specific VOE with its corresponding control condition before revelation. This more conservative analysis revealed significant clusters of activity in ventral ACC and mPFC as well as in precuneus (Fig. 3.2 B).

After establishing which areas in the brain respond to violations of high-level expectations regardless of the specific expectation that was violated, I wanted to answer the inverted question: Which areas respond only to one specific VOE, but not the others? To answer this question, I performed two VOE specific sets of contrasts. First, I contrasted each specific VOE with its corresponding control condition and next I performed VOE specific conjunction analyses, testing where in the brain do I find higher activation for one VOE than for both other VOEs separately (e.g., $Appear_{pre} > Change_{pre} \cap Appear_{pre} > Vanish_{pre}$). Both sets of contrasts showed overlapping activation for the specific VOE types in mostly feature specific areas, such as the intraparietal sulcus (IPS) for when an object appeared out of nowhere or fusiform gyrus when an object unexpectedly changed its color (Fig. 3.2 C).

These two opposing sets of analyses show compelling evidence for a hierarchy in prediction error propagation with prediction errors in feature specific areas when an expectation about their specific feature is violated (e.g., an object's color) to higher and more frontal areas in the brain responding to general violations of expectations.

Here it is important to notice that even though our experimental design confronts subjects with seemingly impossible events, I could not find a dedicated area that seems to respond only to this form of prediction errors. mPFC as well as ACC are known to be involved in prediction error detection and conflict monitoring (Alexander & Brown, 2011, 2019; Fouragnan et al., 2018), PPC is known to be involved in spatial attention (Bressler et al., 2008; Cavanna & Trimble, 2006) and SFG is known to be involved in problem solving (Jack et al., 2013).

After testing for generic and specific responses to seemingly impossible violations of expectations, I wondered whether the response changes when the seemingly impossible action performed (e.g., an object vanishing into thin air) is disenchanted. To this end I showed subjects videos explaining how the magic tricks are performed

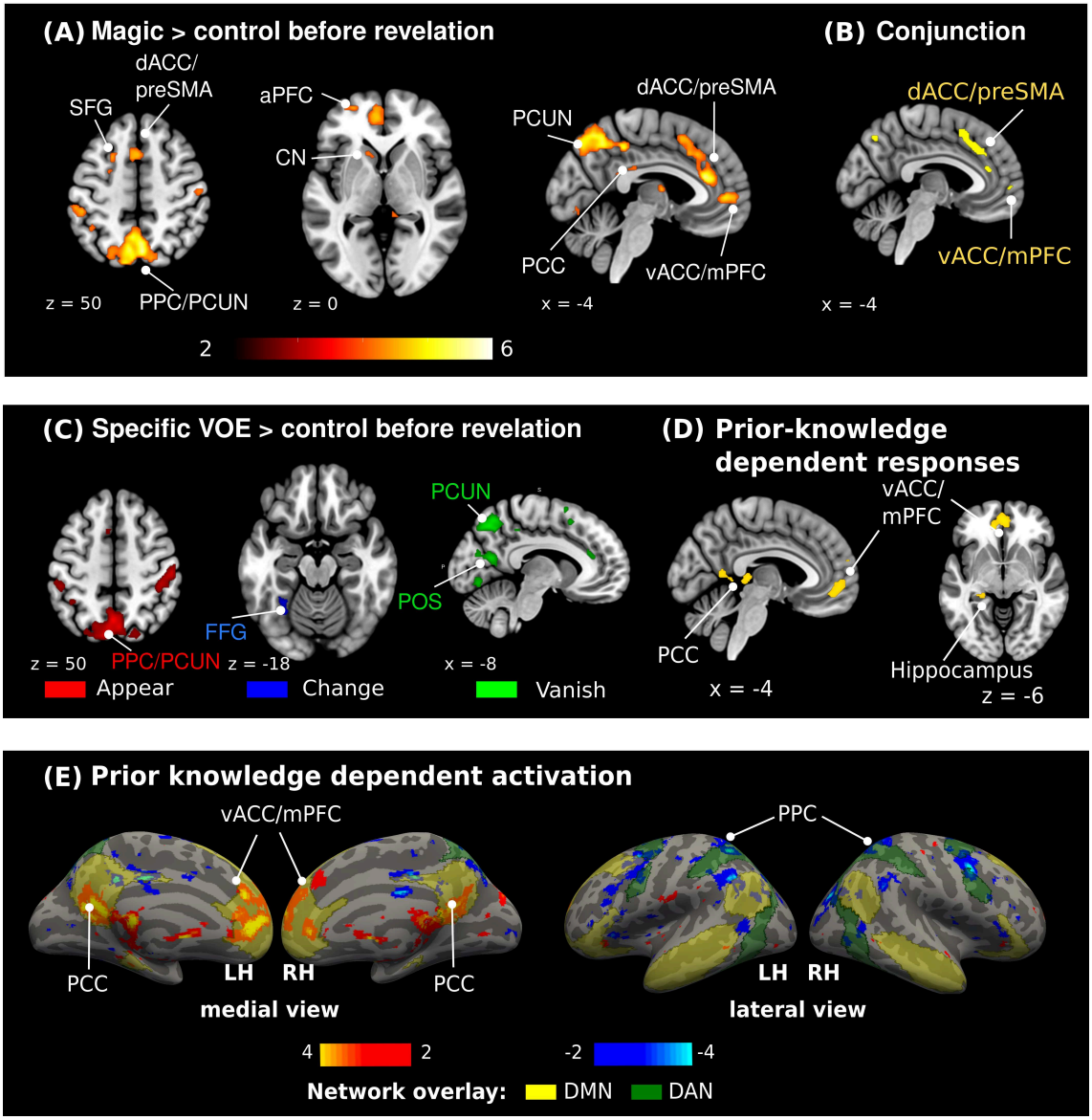


Figure 3.2: Univariate analyses results. **(A)**, shown are the results of comparison between magic and control video pre revelation. The analysis revealed a widespread network of activation in frontal and parietal areas. **(B)**, shows the result of the conjunction analysis for generic activation. **(C)**, shows effect specific activation **(D)**, shows results for the prior knowledge dependent analysis $(Magic_{pre} > Control_{pre}) > (Magic_{post} > Control_{post})$. **(E)**, shows results for the prior knowledge dependent analysis (activation and deactivation) with DMN and DAN as overlays. SFG: superior frontal gyrus; ACC: anterior cingulate cortex; dACC: dorsal ACC; preSMA: pre-supplementary motor area; PFC: prefrontal cortex; aPFC: anterior PFC; CN: caudate nucleus; PCUN: precuneus; PCC: posterior cingulate cortex; vACC: ventral ACC; mPFC: medial PFC; FFG: fusiform gyrus; POS: parieto occipital sulcus; PPC: posterior parietal cortex

and compared the BOLD signal to the perception of magic tricks vs the perception of control videos.

Surprisingly, the resulting clusters of activation overlapped strongly with the ones resulting from the same analysis using data before revelation with strong activation in PPC, ACC and Caudate Nucleus. To better disentangle the response to seemingly impossible events with and without prior knowledge of how they came to be, I performed an interaction analysis between the video- and the prior knowledge condition, i.e. $(Magic_{pre} > Control_{pre}) > (Magic_{post} > Control_{post})$. The interaction contrast revealed a handful of clusters showing higher activation in vACC/mPFC, PPC and PCC (Fig. 3.2 D). This pattern of activation overlaps strikingly with the default mode network (DMN) (Fig. 3.2 E).

Intriguingly, inverting the contrast, i.e. $(Magic_{post} > Control_{post}) > (Magic_{pre} > Control_{pre})$ shows activation overlapping with the dorsal attention network (DAN). At first glance this might seem surprising and counter intuitive, since the DMN is known to be engaged during mind wandering, daydreaming, meditation and to be generally deactivated during demanding tasks (Brewer et al., 2011; Menon, 2023; Raichle, 2015; Smallwood et al., 2021), whereas the DAN is known to be involved during top-down attention (Corbetta & Shulman, 2002). At first one would expect higher activation in the DAN before subjects know the workings behind the magic tricks, as subjects engage in trying to figure out how the seemingly impossible actions are performed. After revealing the methods behind the magic tricks, one would expect higher activation in the DMN as subjects would relax and maybe even start daydreaming during the experiment. However, recently it has been shown that the DMN does not solely ramp up activation during idle states but engages in processing structured events that unfold over time (Baldassano et al., 2018; Jack et al., 2013; Regev et al., 2013; Simony et al., 2016) and surprising events (Brandman et al., 2021; Jääskeläinen et al., 2016). My findings add to a proposed shift from seeing the DMN less as an "intrinsic" network, but more of a "sense-making" network, integrating information from different modalities with prior knowledge (Stawarczyk et al., 2021; Yeshurun et al., 2021).

An increase in the DAN after explaining the methods behind the magic tricks might be explained by assuming that before the explanation of the magic tricks, the

magician's misdirection worked so well that subjects could not even begin to reason how any of the seemingly impossible feats were accomplished. It has already been shown that spectators of a well performed magic trick are often completely clueless of how the trick was performed (Danek et al., 2014; Thomas & Didierjean, 2016; Van de Cruys et al., 2015), even after repeated exposures (Caffaratti et al., 2016; Ekroll et al., 2018). Coupled with the fact that in the experiment there was a very short inter trial interval of two seconds before the next video started, subjects had next to no time to reason how the trick they just saw was performed.

Taken together our univariate results revealed a hierarchy in prediction error signals from generic responses in frontal and parietal areas, to specific responses in feature-specific sensory areas. The generic responses were modulated by prior knowledge in areas overlapping with the DMN, adding evidence to the idea that the DMN is engaged in understanding scenes that unfold over time.

Multivariate Results

To test whether responses to the different VOEs were differentially modulated by prior knowledge, I performed the same interaction analysis for each specific VOE separately (e.g., $(Magic_{App_{pre}} > Control_{App_{pre}}) > (Magic_{App_{post}} > Control_{App_{post}})$). Against my expectations, I did not find knowledge dependent modulation in feature specific visual areas.

The next step was then to move from univariate net comparisons in BOLD signal to multivariate pattern analysis (MVPA). I used a linear discriminant analysis classifier to decode the specific VOE that was perceived. Decoding analyses were performed on a set of regions of interest (ROI) that comprised of two subsets: The first subset was derived from significant responses to magic videos compared to control videos of two previous fMRI experiments (Danek et al., 2015; Parris et al., 2009). The second subset consisted of visual ROIs from early visual cortex (V1, V2, V3) over ventral (ventral occipital, V4) and dorsal (V3A, V3B, lateral occipital and IPS) to the frontal eye field (FEF) in prefrontal cortex.

Decoding analyses using data before explaining the workings behind the tricks, revealed that from each specific VOE an activation pattern emerged that could be decoded significantly above chance level in all posterior visual ROIs (i.e., all visual

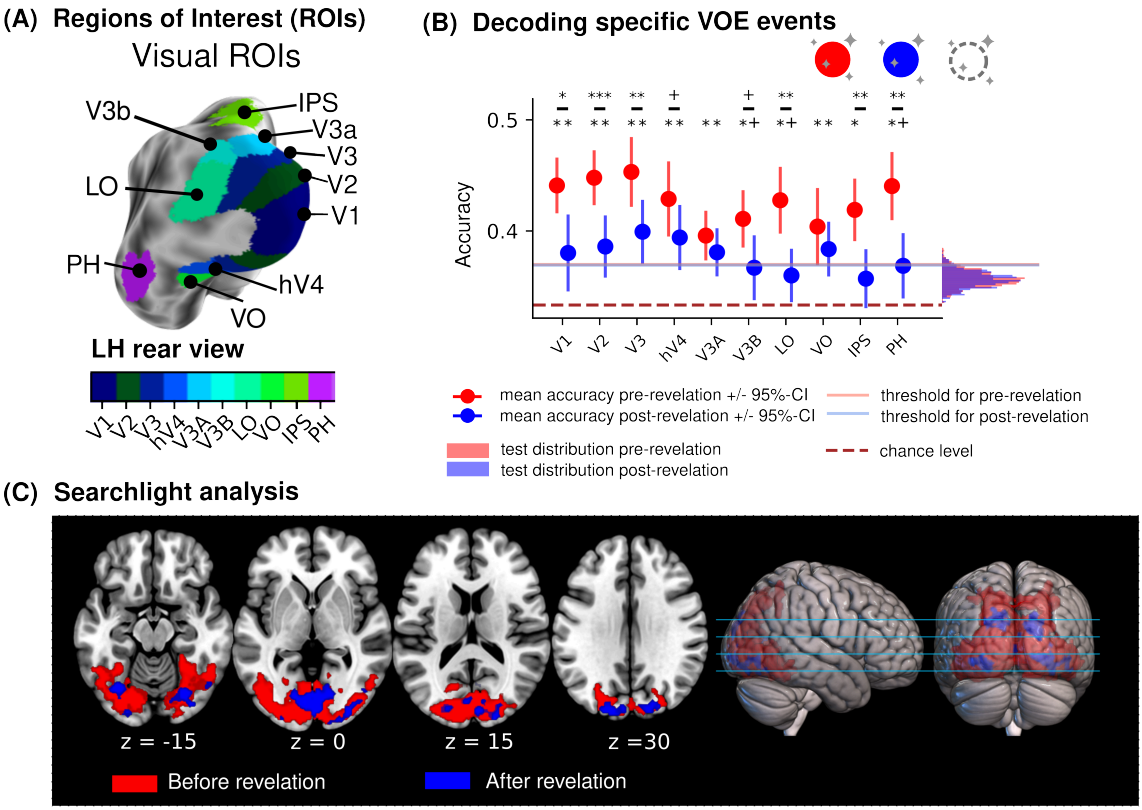


Figure 3.3: **(A)**, shows posterior regions of interest (ROIs), which showed significant decoding of the magic effect. **(B)**, shows decoding accuracies for posterior ROIs pre (red) and post (blue) revelation. Results show significant differences in decoding accuracy in early visual cortex (V1, V2, V3) and higher-level visual areas LO, IPS and PH. **(C)**, shows results of the searchlight decoding analysis pre (red) and post (blue) revelation.

ROIs except FEF) and in PH, a ROI located in inferior temporal sulcus (temporo-occipital division), derived from the univariate result of Danek et al. (2015). As the three different VOE types show three different visual scenes (a scene with a red object, a blue object or no object at all), these results do not come as a surprise. However, the same decoding analysis was performed again on data after explaining the workings behind the magic tricks and decoding accuracies significantly dropped in early visual cortex (V1,V2, V3), LO, IPS and PH.

Given that the visual input remained the same and only the contextual knowledge of participants changed, these results indicate that predictive information about what is going to happen in a complex naturalistic scene, is propagated down to feature specific mid- and early level visual areas (Heilbron & Lange, 2025).

3.1.3 Conclusion

In this fMRI study I used naturalistic video stimuli showing magic tricks and control stimuli to violate predictions based on deeply held beliefs about the workings of the physical world.

Our stimuli induced three different types of violation of expectation, through either showing an object appearing out of nowhere, vanishing into thin air or changing its color unexpectedly. Further I manipulated subjects' prior knowledge about the workings behind the tricks, by providing videos explaining the magic tricks.

Using univariate analysis approaches I found a network of frontal and parietal areas that respond to VOEs regardless of the specific type of VOE. Most of these areas were insensitive to changes in prior knowledge, however a small subset in mPFC and PCC (overlapping with the DMN) showed modulation by prior knowledge.

In contrast, posterior sensory areas showed VOE specific responses. These areas showed no prior knowledge modulation in net BOLD responses. Advancing to multivariate analysis approaches I found that activation patterns in posterior sensory areas are sensitive to the type of VOE. What is even more intriguing is the fact that this sensitivity decreases with prior knowledge.

Taken together our results show a hierarchy of surprise signal processing from generic in frontal and parietal to specific in posterior sensory areas. Both steps of the hierarchy are modulated by prior knowledge, however generic responses dropped in net BOLD signal in the DMN, whereas VOE specific responses change in activation pattern.

3.2 How Can We be Moved by Magic

Whenever we are confronted with an event we did not expect, we are surprised. This can range from a startled response due to an unexpected insect entering our visual field, over to a pleasant feeling of surprise when seeing someone we did not expect to see, up to the bafflement when observing something we deem to be impossible, like a chair levitating in the air.

In my thesis, together with my colleagues Dr. Grassi and Prof. Dr. Wong, I discussed the specific experience of something seemingly impossible happening during a magic performance. To better reflect on this particular experience let's illustrate it with a small example:

> Vincent sits in a show by world famous magician David Copperfield where he revives his famous flying performance. Copperfield lies on the ground and with a swim-like movement of his hands he is lifted up into the air levitating a couple of centimeters over the ground. Some more movements and Copperfield stands two meters above the ground in the air, he performs a salto, swings from one end of the stage to the other.
>
> Vincent thinks to himself: "Well, this is very impressive, but he clearly needs to be hold by thin and strong threads. I've seen Peter Pan doing the same on stage last week!" But right after he finished the thought, Copperfield flies through a set of rings and then ascends up high into the air, just to come down into a glass box. The top of the box is closed, and Copperfield levitates in a seemingly sealed glass container. There is no way Copperfield is held by threads in a sealed container!
>
> Vincent is astonished, looks closely whether he can find any hint of how this could have been accomplished. When he cannot find any, he turns to his friend asking: "How is this possible?" before joining the audience in applauding this extraordinary performance.

It is important to note here that Vincent in this example (and the majority of people) does not believe in real magic, he does not believe that anyone on earth can levitate and fly through the air. He knows that there ought to be an alternative explanation, though he does not know its details. Yet he experiences a strong feeling of surprise.

Here lies a paradoxical contradiction: Why are spectators moved by something they do not believe to be real, but illusory? This question can tightly be linked to another paradox in philosophy - the paradox of fictional emotions (Radford & Weston, 1975; Walton, 1978). Similarly to the paradox of fictional emotions, asking "how can we be moved by something we know is not real?" we wanted to ask, "how can we be moved by something we know to be illusory?"

3.2.1 Epistemic Emotions

To better characterize the experience of a theatrical magic trick I will shortly break down the sequence of emotions we believe a spectator typically experiences.

When seeing something seemingly impossible (e.g. Copperfield flying through a set of rings), we first are surprised. Our jaws drop, we squint and blink to assure ourselves that we did not misinterpret what is in front of us. We then look for possible alternative explanations - we get curious. However, when we cannot find any, we are confused until we finally accept "it must be magic".

These are "epistemic emotions" as they all are triggered by cognitive incongruity, and we believe that they are central to the aesthetics of magic - but how does the cognitive incongruity arise?

3.2.2 The Paradox of Theatrical Magic

The cognitive incongruity that arises from seeing a magic performance can best be described as a conflict of two beliefs. The first belief is based on prior knowledge about how the world works and based on our intuitive understanding of physics - "p cannot happen". p in this case might be "a person levitating unsupported in the air". As one sees David Copperfield levitating and flying over the stage, the second belief is formed - "p is happening". This belief is based on the current sensory information one is processing at the time. Given that (most) people do not really believe in supernatural powers there is a third belief: "p is an illusion". This last belief is crucial for the paradox discussed in this thesis as one should not have epistemic emotions towards something they know/believe is not real. The paradox can thus be derived from the following premises:

(A) People have a cognitive incongruity when experiencing something seemingly impossible (cognitive incongruity premise).

(B) Audiences of magic shows know that the seemingly impossible events are not real, but illusions (knowledge premise).

(C) People do not have a cognitive incongruity when they know that the seemingly impossible events are not real, but illusions (incompatibility premise).

Each statement on its own sounds plausible and intuitively true, yet together they are incompatible. We argue that the paradox, which we call the "paradox of theatrical magic", arises from accepting premise (C) and leave premises (A) and (B) untouched.

3.2.3 Fractured Beliefs

We discussed a set of alternative attempts to solve the paradox of theatrical magic which also reject premise (C). These attempts included viewing magic performances as "fictional puzzles", the epistemic emotions as reactions to "non-assertive thoughts" (Lamarque, 1981) and as epistemic reactions as "habituated responses" (Leddington, 2016). However, we argue that these attempts do not properly capture the aesthetics of magic and provide our own solution to the paradox.

The basis of our argumentation is to view beliefs as more nuanced, that are more fragmented (Borgoni et al., 2021) and need not be held in a single coherent web of belief. Understanding beliefs in this way is compatible with people believing in contradictions (D. Lewis, 1982), in natural and supernatural explanations for the same phenomena (Legare & Gelman, 2008) or holding implicit beliefs that are not endorsed or correlated to explicit beliefs.

Following that line of argumentation we claim that spectators of magic shows believe both, that "p is happening" (David Copperfield flying on stage) and that "p is an illusion" (David Copperfield is not really flying but is somehow supported). This somewhat counterintuitive claim becomes clearer when we accept that there can be beliefs on different levels of abstraction. Specifically, we argue that "p is happening" is a perceptual belief, whereas "p is an illusion" is a cognitive belief.

Based on the sensory information there is an automatic acceptance of what we see to be real (Mandelbaum, 2014; Mandelbaum & Quilty-Dunn, 2015), however contextual background knowledge that one is in a magic show prevents any form of belief updating (supernatural powers are real) or action. The different modes in which the two beliefs are acquired play a crucial role here. Seeing David Copperfield levitating in the air *immediately* and *effortlessly* creates the belief that he is flying - we take what we experience at face value. The cognitive belief that what we experience is an illusion must be actively endorsed, and we need to remind ourselves that we are in a magic show. Importantly, the latter knowledge driven belief does not interfere with the automatic perceptual belief. If it did, people would not be epistemically moved.

This cognitive belief even though not actively endorsed throughout, is not dismissed or idle and still influences our behavior. When seeing someone cut in half on stage or make a 100$ bill appear out of nowhere, we neither call the police nor ask them for infinite wealth.

3.2.4 Conclusion

We discussed the possibility of fractured and eventually contradicting beliefs that people can hold at the same time, though not endorsed simultaneously. This view is compelling as it does not solely apply to the scenario of a magic performance but can be applied to other situations where people's beliefs seem to contradict themselves.

3.3 Memory color influences conscious object perception

3.3.1 Introduction and Design

When two sufficiently different images are presented to the two eyes, the conscious percept alternates between these two images. Which image dominates the conscious percept depends on low-level visual properties such as luminance, contrast, spatial frequency (Brascamp et al., 2015; Levelt, 1965), as well as on more high-level cognitive mechanisms such as attention (Dieter & Tadin, 2011; Meng & Tong, 2004; Mitchell et al., 2004), reward (Marx & Einhäuser, 2015), familiarity (Yu & Blake, 1992), value (Balcetis et al., 2012) and also expectation (Attarha & Moore, 2015; R. Denison et al., 2011).

Based on the Bayesian predictive coding framework, expectation should boost the conscious perception of the expected stimulus, as it has a higher prior probability (Hohwy et al., 2008). Indeed, most of the empirical evidence supports this view, showing that in temporal sequences expected motion, (Hu et al., 2024) gratings, (Attarha & Moore, 2015; R. Denison et al., 2011; Lawler & Silver, 2023) or polygons (Hu et al., 2024) dominates in conscious perception. However, using complex visual scenes as rivaling stimuli, renders the unexpected stimulus more dominant (R. N. Denison et al., 2016; Mudrik et al., 2014; Zacharia et al., 2020).

Given that the representation of stimuli already processed on low levels of the visual hierarchy is enhanced by prediction (Kok et al., 2012), whereas representation of more complex stimuli in higher areas of the visual hierarchy is dampened by expectation (Richter et al., 2018), it is reasonable to believe that predictive influence on dominance depends where in the brain the rivaling stimuli is processed and where a predictive signal originated from.

In a BR experiment using color diagnostic objects I aimed to investigate in which direction the conscious percept is biased when a low-level feature (processed early in the visual hierarchy) is predicted by memory based contextual expectations.

Throughout our life we learn, that certain objects are tinted in one particular color - grass is green, bananas are yellow, and strawberries are red. These objects are called color diagnostic objects, and their strong object-color associations alter

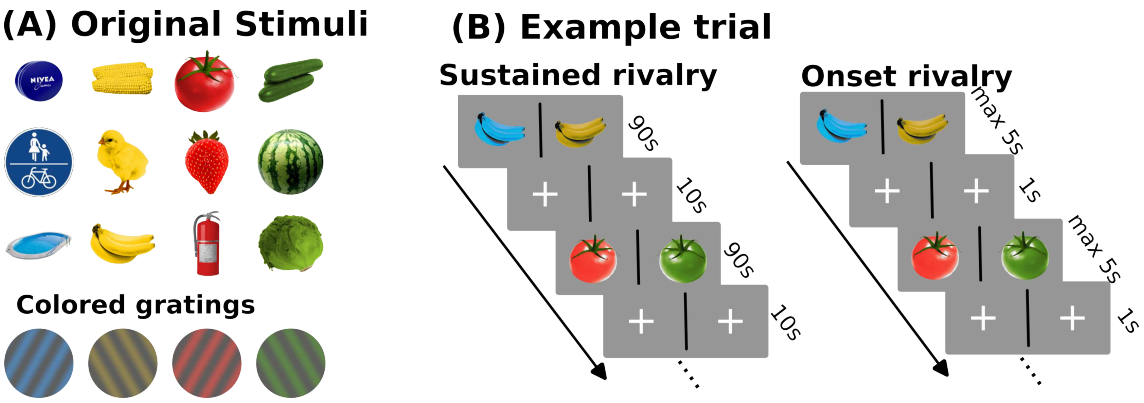


Figure 3.4: **(A)**, shows the stimuli used in our experiment. The upper part shows color diagnostic objects, the lower part shows the colored gratings.**(B)**, shows a sketch of an experimental trial in the sustained rivalry part (left) and onset rivalry part (right).

how they are perceived. Most famously Hansen et al. (2006) showed that color diagnostic objects that are presented in gray, appear to be tinted in their associated color.

This memory color effect (MCE) proved to further modulate cognitive processes like object identification (D. E. Lewis et al., 2013; Teichmann et al., 2020), scene recognition (Oliva & Torralba, 2007) or visual search (Cutler et al., 2024) and an objects associated color can be decoded in the earliest stages of visual processing, even when the objects is presented in gray (Bannert & Bartels, 2013) or an ambiguous color (Vandenbroucke et al., 2016).

Accordingly, I combined MCE with BR to test whether a prediction based on a higher-level semantic information (object identity) modulates a lower-level visual feature (color) during ambiguous sensory information. If it does, the question remains, whether it boosts the perception of the expected (as one would suggest based on findings like R. Denison et al., 2011) or of the unexpected color (as one would suggest based on findings like Mudrik et al., 2011).

3.3.2 Results and Discussion

In a binocular rivalry experiment using twelve different color diagnostic objects of four different colors (red, green, blue and yellow), I found that the congruent color of an object dominated at rivalry onset, during prolonged rivalry and showed longer median dominance durations (Figure 3.5). This boost for the expected color of an object during ambiguous sensory information was mostly driven by objects

associated with yellow or green and could not be systematically found in objects associated with red or blue. The results cannot be explained by general color biases since the congruent color on one object was the same as the incongruent color on another object (e.g., the congruent red on a strawberry was the same incongruent red on a lettuce).

These results contribute to two different and independent discussions: First, our findings contribute to the discussion of how predictions on different levels of abstraction alter the conscious percept during ambiguous visual information. Predictions about objects based on scene context boost the unexpected object (Mudrik et al., 2011; Zacharia et al., 2020), whereas predictions about edge orientation based on motion or sequential information boosts the expected orientation (R. Denison et al., 2011; Hu et al., 2024; Lawler & Silver, 2023).

In a similar vein Press et al. (2020) argues that processing of expected stimuli is only enhanced when the deviation of the expectation is rather small or the sensory information is noisy. If the difference between expectation and sensory information is large enough, processing of the unexpected is enhanced. Seen within this context one could argue that the level of prediction (object identity) and the level of processing of the stimulus (color) is not high enough to enhance the conscious perception of the unexpected stimulus. Similarly, one could say that the unexpected color does not yield a prediction error large enough to enhance processing of the unexpected stimulus.

Second, there is critique on the idea that cognitive top-down effects - and the MCE in particular - modulate perception (Firestone & Scholl, 2016). The argument is that top-down effects modulate other cognitive mechanisms, such as attention, recognition, memory, etc., but not *perception* itself. Here however, we show that top-down expectations based on object knowledge modulate the conscious percept of color.

We believe that our findings cannot be explained by modulation of other cognitive factors, as no judgment, emotion, or recognition was involved, nor is it likely that memory color modulated attention rather than perception (we would expect an inverted effect otherwise - Cutler et al., 2024), or that subjects complied with the experimenters expectation (we had no directed hypothesis).

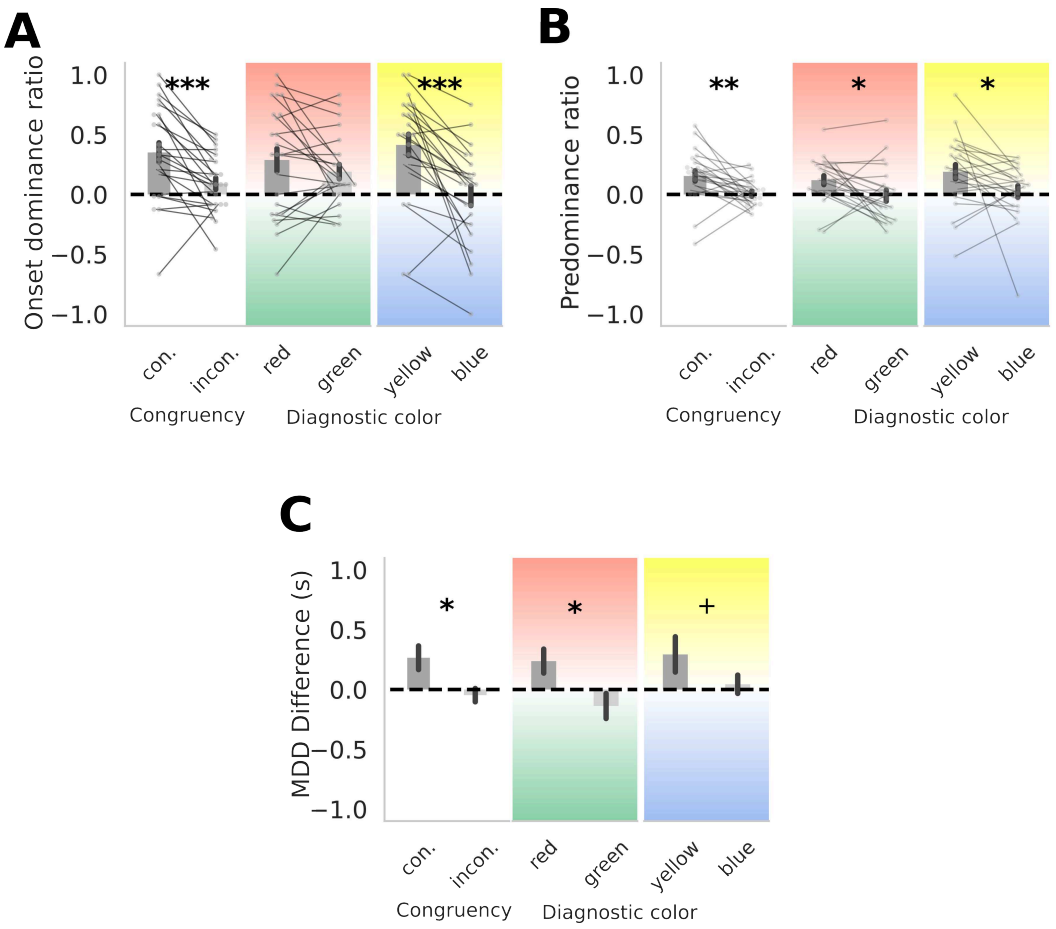


Figure 3.5: (A, B and C), show results for onset dominance, predominance and difference in median dominance duration (MDD) averaged over all colors (left), for the red-green (middle) and yellow-blue color pair (right). **D and E** show MDDs for each color within objects (e.g., yellow on a banana vs blue on a banana) and acros objects (yellow on a banana vs yellow on a nivea tin) respectively

3.3.3 Conclusion

In a behavioral binocular rivalry study, I demonstrated that predictions naturally learned throughout our lives drive how we see the world in ambiguous visual situations. These results set a building block in the discussion on how predictions shape our perception.

4 Grand Discussion

During my doctoral thesis I looked at how expectations about the world we live in, expectations we built over the course of our lives, affect perceptual processing on a neuroscientific level, how these expectations shape our experience from the perspective of aesthetics and how these expectations shape our conscious perception.

In light of the predictive coding framework my work has shown that violations of predictions about complex naturalistic scenes spread over a large network of cortical and subcortical areas in the brain. These predictions are modulated by contextual background knowledge and most intriguingly affect the earliest stages of visual processing (such as primary visual cortex V1). Violations of these high-level expectations, that stand in contrast to how we believe the world works, also open up the interpretation that our beliefs do not need to be constantly aligned but can contradict themselves. Finally, my work has shown that expectations based on contextual world knowledge influence our conscious perception.

My work has mostly focused on the *violation* of expectations, expectations that we built organically during interactions with the world. I investigated violations of expectations due to magic performances and due to the exposure of incongruent colors on a color diagnostic object. The implications of my findings, however, exceed the mere understanding of the perception of magic tricks or color diagnostic objects.

For instance, we found a hierarchy in prediction error signals from more parietal and frontal areas responding to unexpected events irregardless of the form of VOE to more sensory areas responding to specific forms of VOE. One area, the ACC, known to be involved in error detection and conflict monitoring (Alexander & Brown, 2011, 2014, 2019; Botvinick et al., 2004; Fouragnan et al., 2018), was significantly activated by *each* surprising event (unexpected object appearance, disappearance and color change). This might point to a hub in ACC that responds

to *any* form of unexpected input, in line with a recent finding by Grundei et al. (2023). Even more intriguing, we could show that errors of highly complex context specific predictions are distinguishable in the earliest processing steps of the visual hierarchy (V1/V2/V3). The prediction error signal was sensitive to prior knowledge - i.e., when subjects knew how the magic trick was performed the prediction error signals became less distinguishable from each other.

The philosophically discussed question "How can we be moved by Magic" and our answer does not solely contribute to the current discussion on how we experience theatrical magic (Bourne & Bourne, 2025; Leddington, 2016; Robieux et al., 2025; Windsor, 2019) but opens up new ways to look at how our expectations are structured. We proposed that our *beliefs* do not need to be coherent and aligned all the time but can be fractured or contradict themselves as long as they are not endorsed simultaneously.

Finally, moving away from how our memory and expectations shape (neural) processing of unexpected events or the question *why do we feel the way we feel, when seeing certain unexpected events?* my work moved towards the question *how do expectations shape our conscious perception?*. It turned out that memory-based expectations about an object's color boosts the conscious perception of the associated color during ambiguous visual input.

In light of the predictive coding framework my work has shown that violations of expectations based on experiences we made throughout our lives affect processing in large parts of our brain, our conscious perception and in specific cases make us feel a set of epistemic emotions (surprise, curiosity, confusion). If we assume that the brain engages in Bayesian inference to constantly update a model of our world, the prior probability of perceiving a color-diagnostic object in an incongruent color is lower than perceiving it in the congruent color. Perceiving the outcome of a magic trick (e.g., a ball appearing out of nowhere) has a prior probability as close to zero as anything can get (Grassi & Bartels, 2021). Or in light of the proposed hierarchy in predictions and the consistency related precision of predictions (see section 1.2.2),the color of a color-diagnostic object is mostly time invariant (the perceived color only changes with different sources of light) and the consistency is fairly high (we mostly see red tomatoes, but at times see a green one). In contrast the laws of physics are

as time-invariant as anything can get (matter does not cease to exist or is created out of nothing) and the consistency is as high as possible (we are only in space unaffected by gravity). In my work I could show that predictions of mostly time invariant aspects of life already impact the latest stage of perception - the *conscious percept* - and predictions of the most time invariant aspects of life already impact the earliest stages of visual perceptual processing (V1/V2/V3).

Together my work highlights the importance of expectations, experience and memory for perception. Coming back to my example from the beginning where expectations, experience and memory drove my actions. It could continue like this: In the store they changed the color of the milk container. Due to this violation of my expectations, it takes me longer to see the milk. Back home when the doorbell rings and I open the door, in front of me there is not the friend I expected, but someone I thought I would never see again. This violation of expectation leads to surprise, confusion and curiosity as my belief that the person is gone and perceptual belief formed through the sight of their face contradict themselves. This contradiction significantly alters perceptual processes in the earliest region of my visual cortex and activates surprise related areas in parietal, prefrontal and subcortical areas.

These less trivial examples show how the violations of expectations shape our conscious perception, how we feel, and the neural processing of information informed by the work presented here.

5 Outlook

The work done in this thesis touched on three different topics using distinct methods. With neuroimaging techniques, I looked into the processing of events that violate predictions of deeply held beliefs, in a philosophical discussion I asked and answered the question why we are surprised by theatrical magic performances and in a behavioral BR experiment I investigated the influence of object-color knowledge on the conscious percept. All three projects answered intriguing questions, however at the same time they open up new questions and research possibilities.

In the neuroimaging project I found that violations of expectations that are based in the intuitive understanding of physics, alter neural activity in the earliest stages of visual processing. However, with only three different forms of VOE that were all connected to *how an object appeared to be* (present, absent, colored), we do not know whether the distinctiveness results from an internal model of the world, similar to a "physics engine" (Battaglia et al., 2013) or e.g., from violations in object representations. There is current work investigating the understanding of intuitive physics (Fischer et al., 2016; S. Liu et al., 2024; Paulun & Fleming, 2020; Paulun et al., 2015; Paulun et al., 2017; Pramod et al., 2022; Schwettmann et al., 2019), however, to the best of my knowledge none of the neuroimaging studies investigated whether violations of different intuitive physical rules emitted distinctive neural activation in early visual areas. It would be worthwhile to replicate our experiment with different effects (e.g., levitation, penetration, destroy and reconstruct) to find out whether an automatically implemented physics engine-like predictive model alters early sensory areas.

In the psychologically informed philosophical discussion, I argued that beliefs do not need to be aligned but can be fractured or contradicting. Similarly, our expectations can be contradicting and fractured. We could have a high-level context

specific expectation based on the knowledge "I am in a magic show" that predicts something seemingly impossible is about to happen (a person levitating unsupported in the air). Simultaneously, we could have a deeply ingrained expectation based on experiences in our lives (gravity pulls heavy objects down to earth).

This view would explain why magic tricks surprise us, even when we know the outcome. When a magician asks us to pick a card, sign it and then loose it in the deck, we *know* the magician will later find the card even though (we believe) the deck was shuffled, the magician did not know our card and the deck was just an ordinary deck of cards. When they ultimately find the card, we are surprised *even though* this was exactly what we expected based on our knowledge "I am in a magic show, magicians do impossible things" (or maybe the magician even announced it). However, another expectation "there is no way they know my card and where it is" is violated, which results in the feeling of surprise. If the magician *does not* find the card and honestly confesses, they do not know what our card was and where it is, we are surprised too. The outcome based on our ingrained expectation that the magician cannot have the slightest clue where the card is and therefore will not find it, is satisfied. However, the outcome based on contextual knowledge that we are in a magic show and therefore the magician should find the card in some magical way is violated leading to a feeling of surprise as well.

The view of fractured and partially contradicting beliefs and expectations might not solely be applicable to the sensation of surprise. Picture someone in a haunted house with actors dressed as monsters. Even though on a contextual level they know that the "monsters" do not pose a real threat, but on a different level the sight of a creepy clown induces terror and fear. Or picture someone playing lotto. Even though they know that the chances of winning the big price and therefore the expectation of a win *should* be extremely low on an abstract higher level, but on a different emotional level they might *feel* that they got a lucky hand this time around.

Finally, in the BR study I showed that the expected color of an object dominates during rivalry. This finding is in line with predictive coding theories about rivalry (Hohwy et al., 2008) but in light of conflicting evidence showing dominance for unexpected naturalistic stimuli during rivalry (R. N. Denison et al., 2016; Mudrik et

al., 2011; Zacharia et al., 2020) the question remains whether the predictive bias is ubiquitous or depends on which feature of a stimulus is rivaling. To the best of my knowledge there has not been a systematic investigation of different expectations on the conscious percept. I showed that object-color associations influence perception of an object feature (color), but do object-color associations also influence object perception in the same way? Are all features of an object equally affected by contextual knowledge? Do expectations based on temporal and contextual information bias conscious perception in the same way?

It leaves to say that all the insights gained by my work open up new and exciting scientific questions that I am looking forward to answer.

Bibliography

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. https://doi.org/10.1038/nn.2921

Alexander, W. H., & Brown, J. W. (2014). A general role for medial prefrontal cortex in event prediction. *Frontiers in Computational Neuroscience*, *8*.

Alexander, W. H., & Brown, J. W. (2019). The Role of the Anterior Cingulate Cortex in Prediction Error and Signaling Surprise. *Topics in Cognitive Science*, *11*(1), 119–135. https://doi.org/10.1111/tops.12307

Ali, A., Ahmad, N., de Groot, E., Johannes van Gerven, M. A., & Kietzmann, T. C. (2022). Predictive coding is a consequence of energy efficiency in recurrent neural networks. *Patterns*, *3*(12), 100639. https://doi.org/10.1016/j.patter.2022.100639

Ambrus, G. G., Dotzer, M., Schweinberger, S. R., & Kovács, G. (2017). The occipital face area is causally involved in the formation of identity-specific face representations. *Brain Structure and Function*, *222*(9), 4271–4282. https://doi.org/10.1007/s00429-017-1467-2

Attarha, M., & Moore, C. M. (2015). Onset rivalry: Factors that succeed and fail to bias selection. *Attention, Perception & Psychophysics*, *77*(2), 520–535. https://doi.org/10.3758/s13414-014-0793-1

Auksztulewicz, R., & Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. *Cortex*, *80*, 125–140. https://doi.org/10.1016/j.cortex.2015.11.024

Baker, D. H., & Graf, E. W. (2009). Natural images dominate in binocular rivalry. *Proceedings of the National Academy of Sciences*, *106*(13), 5436–5441. https://doi.org/10.1073/pnas.0812860106

Balcetis, E., Dunning, D., & Granot, Y. (2012). Subjective value determines initial dominance in binocular rivalry. *Journal of Experimental Social Psychology*, *48*(1), 122–129. https://doi.org/10.1016/j.jesp.2011.08.009

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

Baldeweg, T., Klugman, A., Gruzelier, J., & Hirsch, S. R. (2004). Mismatch negativity potentials and cognitive impairment in schizophrenia. *Schizophrenia Research*, *69*(2), 203–217. https://doi.org/10.1016/j.schres.2003.09.009

Bannert, M. M., & Bartels, A. (2013). Decoding the Yellow of a Gray Banana. *Current Biology*, *23*(22), 2268–2272. https://doi.org/10.1016/j.cub.2013.09.016

WOS:000327417000025

Barnhart, A. S. (2010). The Exploitation of Gestalt Principles by Magicians. *Perception*, *39*(9), 1286–1289. https://doi.org/10.1068/p6766

Barnhart, A. S., & Goldinger, S. D. (2014). Blinded by magic: Eye-movements reveal the misdirection of attention. *Frontiers in Psychology*, *5*, 1461–1461. https://doi.org/10.3389/fpsyg.2014.01461

MAG ID: 2139471301

Bartels, A., & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: New results and a review *. *European Journal of Neuroscience*, *12*(1), 172–193. https://doi.org/10.1046/j.1460-9568.2000.00905.x

Bartels, A., Logothetis, N. K., & Moutoussis, K. (2008). fMRI and its interpretations: An illustration on directional selectivity in area V5/MT. *Trends in Neurosciences*, *31*(9), 444–453. https://doi.org/10.1016/j.tins.2008.06.004

Bartels, A., & Zeki, S. (2004). Functional brain mapping during free viewing of natural scenes. *Human Brain Mapping*, *21*(2), 75–85. https://doi.org/10.1002/hbm.10153

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

Bertero, M., Poggio, T., & Torre, V. (1988). Ill-posed problems in early vision. *Proceedings of the IEEE*, *76*(8), 869–889. https://doi.org/10.1109/5.5962

Bilalić, M., McLeod, P., & Gobet, F. (2008a). Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology*, *56*(2), 73–102. https://doi.org/10.1016/j.cogpsych.2007.02.001

Bilalić, M., McLeod, P., & Gobet, F. (2008b). Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition*, *108*(3), 652–661. https://doi.org/10.1016/j.cognition.2008.05.005

Binet, A. (1894). La Psychologie De La Prestidigitation. *Revue des Deux Mondes (1829-1971)*, *125*(4), 903–922.

Blake, R., Westendorf, D. H., & Overton, R. (1980). What is Suppressed during Binocular Rivalry? *Perception*, *9*(2), 223–231. https://doi.org/10.1068/p090223

Bolz, J., & Gilbert, C. D. (1986). Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature*, *320*(6060), 362–365. https://doi.org/10.1038/320362a0

Borgoni, C., Kindermann, D., & Onofri, A. (2021). *The Fragmented Mind*. Oxford University Press.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Bourne, C., & Bourne, E. C. (2025). Magic Tricks as Quasi-Miracles: A New Approach in the Aesthetics of Performance. *Journal of Performance Magic*, *8*(1). https://doi.org/10.5920/jpm.1559

Brandman, T., Malach, R., & Simony, E. (2021). The surprising role of the default mode network in naturalistic perception. *Communications Biology*, *4*(1), 1–9. https://doi.org/10.1038/s42003-020-01602-z

Brascamp, J. W., Klink, P. C., & Levelt, W. J. M. (2015). The 'laws' of binocular rivalry: 50 years of Levelt's propositions. *Vision Research*, *109*, 20–37. https://doi.org/10.1016/j.visres.2015.02.019

Bressler, S. L., Tang, W., Sylvester, C. M., Shulman, G. L., & Corbetta, M. (2008). Top-Down Control of Human Visual Cortex by Frontal and Parietal Cortex

in Anticipatory Visual Spatial Attention. *Journal of Neuroscience*, *28*(40), 10056–10061. https://doi.org/10.1523/JNEUROSCI.1776-08.2008

Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y.-Y., Weber, J., & Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 20254–20259. https://doi.org/10.1073/pnas.1112029108

Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00025

Bubic, A., von Cramon, D. Y., Jacobsen, T., Schroeger, E., & Schubotz, R. I. (2009). Violation of Expectation: Neural Correlates Reflect Bases of Prediction. *Journal of Cognitive Neuroscience*, *21*(1), 155–168. https://doi.org/10.1162/jocn.2009.21013

WOS:000262232000012

Caffaratti, H., Navajas, J., Rey, H. G., & Quian Quiroga, R. (2016). Where is the ball? Behavioral and neural responses elicited by a magic trick. *Psychophysiology*, *53*(9), 1441–1448. https://doi.org/10.1111/psyp.12691

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain*, *129*(3), 564–583. https://doi.org/10.1093/brain/awl004

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Conrad, V., Bartels, A., Kleiner, M., & Noppeney, U. (2010). Audiovisual interactions in binocular rivalry. *Journal of Vision*, *10*(10), 27. https://doi.org/10.1167/10.10.27

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Cutler, A., Rivest, J., & Cavanagh, P. (2024). The role of memory color in visual attention. *Attention, Perception, & Psychophysics*, *86*(1), 28–35. https://doi.org/10.3758/s13414-023-02714-4

Danek, A. H., Fraps, T., von Müller, A., Grothe, B., & Öllinger, M. (2014). Working Wonders? Investigating insight with magic tricks. *Cognition*, *130*(2), 174–185. https://doi.org/10.1016/j.cognition.2013.11.003

Danek, A. H., Öllinger, M., Fraps, T., Grothe, B., & Flanagin, V. L. (2015). An fMRI investigation of expectation violation in magic tricks. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00084

Denison, R., Piazza, E., & Silver, M. (2011). Predictive context influences perceptual selection during binocular rivalry. *Frontiers in Human Neuroscience*, *5*.

Denison, R. N., Sheynin, J., & Silver, M. A. (2016). Perceptual suppression of predicted natural images. *Journal of Vision*, *16*(13), 6. https://doi.org/10.1167/16.13.6

Dessoir, M. (1893). The Psychology of Legerdemain. *Revue Philosophique de la France Et de l'Etranger*, *36*, 659–660.

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences*, *16*(6), 317–321. https://doi.org/10.1016/j.tics.2012.04.007

Dieter, K., & Tadin, D. (2011). Understanding Attentional Modulation of Binocular Rivalry: A Framework Based on Biased Competition. *Frontiers in Human Neuroscience*, *5*.

Dwarakanath, A., Kapoor, V., Werner, J., Safavi, S., Fedorov, L. A., Logothetis, N. K., & Panagiotaropoulos, T. I. (2022, June 10). *Bistability of prefrontal states gates access to consciousness*. https://doi.org/10.1101/2020.01.29.924928

Ekman, M., Kok, P., & de Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, *8*(1), 15276. https://doi.org/10.1038/ncomms15276

Ekroll, V., De Bruyckere, E., Vanwezemael, L., & Wagemans, J. (2018). Never Repeat the Same Trick Twice—Unless it is Cognitively Impenetrable. *i-Perception*, *9*(6), 2041669518816711. https://doi.org/10.1177/2041669518816711

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, N.Y.*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, e229. https://doi.org/10.1017/S0140525X15000965

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*(34), E5072–E5081. https://doi.org/10.1073/pnas.1610344113

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*(7), 2887–2906. https://doi.org/10.1002/hbm.24047

Friston, K. (2002). Functional integration and inference in the brain. *Progress in Neurobiology*, *68*(2), 113–143. https://doi.org/10.1016/S0301-0082(02)00076-X

Friston, K. (2008). Hierarchical Models in the Brain. *PLOS Computational Biology*, *4*(11), e1000211. https://doi.org/10.1371/journal.pcbi.1000211

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2018). Does predictive coding have a future? *Nature Neuroscience*, *21*(8), 1019–1021. https://doi.org/10.1038/s41593-018-0200-7

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Friston, Karl. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Grassi, P. R., Plikat, V., & Wong, H. Y. (2023). How can we be moved by magic? *The British Journal of Aesthetics*, ayad026. https://doi.org/10.1093/aesthj/ayad026

Grassi, P. R., & Bartels, A. (2021). Magic, Bayes and wows: A Bayesian account of magic tricks. *Neuroscience & Biobehavioral Reviews*, *126*, 515–527. https://doi.org/10.1016/j.neubiorev.2021.04.001

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10), 1409–1422. https://doi.org/10.1016/S0042-6989(01)00073-6

Grotheer, M., & Kovács, G. (2016). Can predictive coding explain repetition suppression? *Cortex*, *80*, 113–124. https://doi.org/10.1016/j.cortex.2015.11.027

Grundei, M., Schmidt, T. T., & Blankenburg, F. (2023). A multimodal cortical network of sensory expectation violation revealed by fMRI. *Human Brain Mapping*, *44*(17), 5871–5891. https://doi.org/10.1002/hbm.26482

Gschwind, M., Pourtois, G., Schwartz, S., Van De Ville, D., & Vuilleumier, P. (2012). White-Matter Connectivity between Face-Responsive Regions in the Human Brain. *Cerebral Cortex*, *22*(7), 1564–1576. https://doi.org/10.1093/cercor/bhr226

Güçlütürk, Y., Güçlü, U., van Gerven, M., & van Lier, R. (2018). Representations of naturalistic stimulus complexity in early and associative visual and auditory cortices. *Scientific Reports*, *8*(1), 3439. https://doi.org/10.1038/s41598-018-21636-y

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*(11), 1367–1368. https://doi.org/10.1038/nn1794

Heilbron, M., & Lange, F. P. de. (2025, May 19). *Higher-level spatial prediction in natural vision across mouse visual cortex*. https://doi.org/10.1101/2025.05.15.654212

Helmholtz, H. von. (1867). *Handbuch der physilogischen Optik*.

Hering, E. (1920). *Grundzüge der Lehre vom Lichtsinn*.

Hohwy, J. (2012). Attention and Conscious Perception in the Hypothesis Testing Brain. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00096

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., & Hasson, U. (2012). Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron*, *76*(2), 423–434. https://doi.org/10.1016/j.neuron.2012.08.011

Hsu, Y.-F., & Hämäläinen, J. A. (2021). Both contextual regularity and selective attention affect the reduction of precision-weighted prediction errors but in distinct manners. *Psychophysiology*, *58*(3), e13753. https://doi.org/10.1111/psyp.13753

Hu, R., Li, S., Yuan, P., Wang, Y., & Jiang, Y. (2024). Temporal integration by multi-level regularities fosters the emergence of dynamic conscious experience. *Annals of the New York Academy of Sciences*, *1533*(1), 156–168. https://doi.org/10.1111/nyas.15099

Hu, R., Yuan, P., Wang, Y., Wang, Y., & Y, J. (2021). Visual temporal integration by multi-level regularities fosters the emergence of dynamic conscious experience. *bioRxiv*. https://doi.org/10.1101/2021.04.28.440365
MAG ID: 3158302083

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.2.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E. M., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain

and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530. https://doi.org/10.1016/j.neuron.2013.09.009

Jääskeläinen, I. P., Pajula, J., Tohka, J., Lee, H.-J., Kuo, W.-J., & Lin, F.-H. (2016). Brain hemodynamic activity during viewing and re-viewing of comedy movies explained by experienced humor. *Scientific Reports*, *6*(1), 27741. https://doi.org/10.1038/srep27741

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H., & Snyder, A. Z. (2013). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, *66*, 385–401. https://doi.org/10.1016/j.neuroimage.2012.10.061

Jakob Hohwy. (2013). *The Predictive Mind* (Vol. First edition). OUP Oxford.

Jastrow, J. (1897). Magic Stage Illusions and Scientific Diversions, Including Trick Photography. Compiled and edited by Albert A. Hopkins, with an introduction by Henry Ridgely Evans. New York, Munn & Co. 1897. With four hundred illustrations. Large 8vo. Pp. 556. Price, $2.50. *Science*, *6*(153), 850–851. https://doi.org/10.1126/science.6.153.850.c

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158. https://doi.org/10.1016/S0959-4388(03)00042-4

Kim, I., Hong, S. W., Shevell, S. K., & Shim, W. M. (2020). Neural representations of perceptual color experience in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *117*(23), 13145–13150. https://doi.org/10.1073/pnas.1911041117

Kimura, A., Wada, Y., Masuda, T., Goto, S.-i., Tsuzuki, D., Hibino, H., Cai, D., & Dan, I. (2013). Memory Color Effect Induced by Familiarity of Brand Logos. *PLOS ONE*, *8*(7), e68474. https://doi.org/10.1371/journal.pone.0068474

Kok, P., & de Lange, F. P. (2014). Shape Perception Simultaneously Up- and Downregulates Neural Activity in the Primary Visual Cortex. *Current Biology*, *24*(13), 1531–1535. https://doi.org/10.1016/j.cub.2014.05.042

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *2*(75), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Kourtzi, Z., & Kanwisher, N. (2001). Representation of Perceived Object Shape by the Human Lateral Occipital Complex. *Science*, *293*(5534), 1506–1509. https://doi.org/10.1126/science.1061133

Kuhn, G., Amlani, A. A., & Rensink, R. A. (2008). Towards a science of magic. *Trends in Cognitive Sciences*, *12*(9), 349–354. https://doi.org/10.1016/j.tics.2008.05.008

Kuhn, G., & Land, M. F. (2006). There's more to magic than meets the eye . *Current Biology*, *16*(22), R950–R951. https://doi.org/10.1016/j.cub.2006.10.012

Kuhn, G., & Rensink, R. A. (2016). The Vanishing Ball Illusion: A new perspective on the perception of dynamic events. *Cognition*, *148*, 64–70. https://doi.org/10.1016/j.cognition.2015.12.003

Kuhn, G., & Tatler, B. W. (2005). Magic and fixation: Now you don't see it, now you do. *Perception*, *34*(9), 1155–1161. https://doi.org/10.1068/p3409bn1
MAG ID: 1998966912

Kuhn, G., Tatler, B. W., & Cole, G. G. (2009). You look where I look! Effect of gaze cues on overt and covert attention in misdirection. *Visual Cognition*, *17*, 925–944. https://doi.org/10.1080/13506280902826775
MAG ID: 2127585516

Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., Patterson, R. D., Howard, M. A., Friston, K. J., & Griffiths, T. D. (2011). Predictive Coding and Pitch Processing in the Auditory Cortex. *Journal of Cognitive Neuroscience*, *23*(10), 3084–3094. https://doi.org/10.1162/jocn_a_00021

Lamarque, P. V. (1981). How Can We Fear and Pity Fictions? *British journal of aesthetics*, *21*, 291–304.

Lamont, P. (2017). A Particular Kind of Wonder: The Experience of Magic past and Present. *Review of General Psychology*, *21*(1), 1–8. https://doi.org/10.1037/gpr0000095

Lamont, P., Henderson, J. M., & Smith, T. J. (2010). Where Science and Magic Meet: The Illusion of a "Science of Magic". *Review of General Psychology*, *14*(1), 16–21. https://doi.org/10.1037/a0017157

Lawler, E. A., & Silver, M. A. (2023). Enhanced perceptual selection of predicted stimulus orientations following statistical learning. *Journal of Vision*, *23*(7), 3. https://doi.org/10.1167/jov.23.7.3

Leddington, J. (2016). The Experience of Magic. *The Journal of Aesthetics and Art Criticism*, *74*(3), 253–264. https://doi.org/10.1111/jaac.12290
MAG ID: 2507590215

Lee, S.-H., & Blake, R. (2004). A fresh look at interocular grouping during binocular rivalry. *Vision Research*, *44*(10), 983–991. https://doi.org/10.1016/j.visres.2003.12.007

Legare, C. H., & Gelman, S. A. (2008). Bewitchment, Biology, or Both: The Co-Existence of Natural and Supernatural Explanatory Frameworks Across Development. *Cognitive Science*, *32*(4), 607–642. https://doi.org/10.1080/03640210802066766

Leopold, D. A., Logothetis, N. K., Leopold, D. A., Logothetis, N. K., Leopold, D. A., Logothetis, N. K., Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, *3*(7), 254–264. https://doi.org/10.1016/S1364-6613(99)01332-7

Levelt, W. J. (1965). *On binocular rivalry*. Inst. Perception Rvo-Tno.

Lewis, D. (1982). Logic for Equivocators. *Noûs*, *16*(3), 431–441. https://doi.org/10.2307/2216219

Lewis, D. E., Pearson, J., & Khuu, S. K. (2013). The Color "Fruit": Object Memories Defined by Color. *PLOS ONE*, *8*(5), e64960. https://doi.org/10.1371/journal.pone.0064960

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of Face Parts and Face Configurations: An fMRI Study. *Journal of Cognitive Neuroscience*, *22*(1), 203–211. https://doi.org/10.1162/jocn.2009.21203

Liu, S., Lydic, K., Mei, L., & Saxe, R. (2024). Violations of physical and psychological expectations in the human adult brain. *Imaging Neuroscience*, *2*, 1–25. https://doi.org/10.1162/imag_a_00068

Logothetis, N. K., Leopold, D. A., & Sheinberg, D. L. (1996). What is rivalling during binocular rivalry? *Nature*, *380*(6575), 621–624. https://doi.org/10.1038/380621a0

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, *54*(6), i–95. https://doi.org/10.1037/h0093502

Luhrmann, T. M. (1989). The Magic of Secrecy. *Ethos*, *17*(2), 131–165.

Lunghi, C., & Alais, D. (2013). Touch Interacts with Vision during Binocular Rivalry with a Tight Orientation Tuning. *PLOS ONE*, *8*(3), e58754. https://doi.org/10.1371/journal.pone.0058754

Lunghi, C., Binda, P., & Morrone, M. C. (2010). Touch disambiguates rivalrous perception at early stages of visual analysis. *Current Biology*, *20*(4), R143–R144. https://doi.org/10.1016/j.cub.2009.12.015

Lunghi, C., Lo Verde, L., & Alais, D. (2017). Touch Accelerates Visual Awareness. *i-Perception*, *8*(1), 2041669516686986. https://doi.org/10.1177/2041669516686986

Macknik, S. L., King, M., Randi, J., Robbins, A., Teller, Thompson, J., & Martinez-Conde, S. (2008). Attention and awareness in stage magic: Turning tricks into research. *Nature Reviews Neuroscience*, *9*(11), 871–879. https://doi.org/10.1038/nrn2473

Macknik, S. L., & Martinez-Conde, S. (2009). Real magic: Future studies of magic should be grounded in neuroscience. *Nature Reviews Neuroscience*, *10*(3), 241–241. https://doi.org/10.1038/nrn2473-c2
MAG ID: 2032614985

Mahon, B. Z., Schwarzbach, J., & Caramazza, A. (2010). The Representation of Tools in Left Parietal Cortex Is Independent of Visual Experience. *Psychological Science*, *21*(6), 764–771. https://doi.org/10.1177/0956797610370754

Mandelbaum, E. (2014). Thinking is Believing. *Inquiry*, *57*(1), 55–96. https://doi.org/10.1080/0020174X.2014.858417

Mandelbaum, E., & Quilty-Dunn, J. (2015). Believing without Reason, or: Why Liberals Shouldn't Watch Fox News. *The Harvard Review of Philosophy*, *22*, 42–52. https://doi.org/10.5840/harvardreview2015226

Marx, S., & Einhäuser, W. (2015). Reward modulates perception in binocular rivalry. *Journal of Vision*, *15*(1), 11–11. https://doi.org/10.1167/15.1.11

Meng, M., & Tong, F. (2004). Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision*, *4*(7), 539–551. https://doi.org/10.1167/4.7.2

Menon, V. (2023). 20 years of the default mode network: A review and synthesis. *Neuron*, *111*(16), 2469–2487. https://doi.org/10.1016/j.neuron.2023.04.023

Mitchell, J. F., Stoner, G. R., & Reynolds, J. H. (2004). Object-based attention determines dominance in binocular rivalry. *Nature*, *429*(6990), 410–413. https://doi.org/10.1038/nature02584

Mitterer, H., & de Ruiter, J. P. (2008). Recalibrating Color Categories Using World Knowledge. *Psychological Science*, *19*(7), 629–634. https://doi.org/10.1111/j.1467-9280.2008.02133.x

Modirshanechi, A., Kiani, M. M., & Aghajan, H. (2019). Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks. *Neuroimage*, *196*, 302–317. https://doi.org/10.1016/j.neuroimage.2019.04.028 WOS:000470833800029

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat's visual cortex. *The Journal of Physiology*, *283*(1), 101–120. https://doi.org/10.1113/jphysiol.1978.sp012490

Mruczek, R. E. B., von Loga, I. S., & Kastner, S. (2013). The representation of tool and non-tool object information in the human intraparietal sulcus. *Journal of Neurophysiology*, *109*(12), 2883–2896. https://doi.org/10.1152/jn.00658.2012

Muckli, L. (2010). What are we missing here? Brain imaging evidence for higher cognitive functions in primary visual cortex V1. *International Journal of Imaging Systems and Technology*, *20*(2), 131–139. https://doi.org/10.1002/ima.20236

Mudrik, L., Deouell, L. Y., & Lamy, D. (2011). Scene congruency biases Binocular Rivalry. *Consciousness and Cognition*, *20*(3), 756–767. https://doi.org/10.1016/j.concog.2011.01.001

Mudrik, L., Faivre, N., & Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences*, *18*(9), 488–496. https://doi.org/10.1016/j.tics.2014.04.009

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, *99*(23), 15164–15169. https://doi.org/10.1073/pnas.192579399

Oliva, A., & Schyns, P. G. (2000). Diagnostic Colors Mediate Scene Recognition. *Cognitive Psychology*, *41*(2), 176–210. https://doi.org/10.1006/cogp.1999.0728

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision*, *8*(5), 13.1–16. https://doi.org/10.1167/8.5.13

Ono, M., Hirose, N., & Mori, S. (2022). Tactile information affects alternating visual percepts during binocular rivalry using naturalistic objects. *Cognitive Research: Principles and Implications*, *7*(1), 40. https://doi.org/10.1186/s41235-022-00390-w

Parris, B. A., Kuhn, G., Mizon, G. A., Benattayallah, A., & Hodgson, T. L. (2009). Imaging the impossible: An fMRI study of impossible causal relationships in magic tricks. *NeuroImage*, *45*(3), 1033–1039. https://doi.org/10.1016/j.neuroimage.2008.12.036

Parvizi, J., Jacques, C., Foster, B. L., Withoft, N., Rangarajan, V., Weiner, K. S., & Grill-Spector, K. (2012). Electrical Stimulation of Human Fusiform Face-Selective Regions Distorts Face Perception. *Journal of Neuroscience*, *32*(43), 14915–14920. https://doi.org/10.1523/JNEUROSCI.2609-12.2012

Paulun, V. C., & Fleming, R. W. (2020). Visually inferring elasticity from the motion trajectory of bouncing cubes. *Journal of Vision*, *20*(6), 6. https://doi.org/10.1167/jov.20.6.6

Paulun, V. C., Kawabe, T., Nishida, S., & Fleming, R. W. (2015). Seeing liquids from static snapshots. *Vision Research*, *115*, 163–174. https://doi.org/10.1016/j.visres.2015.01.023

Paulun, V. C., Schmidt, F., van Assen, J. J. R., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of Vision*, *17*(1), 20. https://doi.org/10.1167/17.1.20

Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, *3*(1), 13–26. https://doi.org/10.1038/s44159-023-00254-0

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research*, *41*(24), 3145–3161. https://doi.org/10.1016/S0042-6989(01)00173-0

Plikat, V., Grassi, P. R., Frack, J., & Bartels, A. (2025). Hierarchical surprise signals in naturalistic violation of expectations. *Imaging Neuroscience*, *3*, imag_a_00459. https://doi.org/10.1162/imag_a_00459

Pramod, R. T., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain (P. Kok, F. P. de Lange, P. Kok, & J. Snow, Eds.). *eLife*, *11*, e71736. https://doi.org/10.7554/eLife.71736

Press, C., Kok, P., & Yon, D. (2020). The Perceptual Prediction Paradox. *Trends in Cognitive Sciences*, *24*(1), 13–24. https://doi.org/10.1016/j.tics.2019.11.003

Quiroga, R. Q. (2016). Magic and cognitive neuroscience. *Current Biology*, *26*(10), R390–R394. https://doi.org/10.1016/j.cub.2016.03.061

Radford, C., & Weston, M. (1975). How Can We Be Moved by the Fate of Anna Karenina? *Proceedings of the Aristotelian Society, Supplementary Volumes*, *49*, 67–93.

Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, *38*(1), 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79. https://doi.org/10.1038/4580

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience*, *33*(40), 15978–15988. https://doi.org/10.1523/JNEUROSCI.1580-13.2013

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To See or not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, *8*(5), 368–373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x

Richter, D., Ekman, M., & Lange, F. P. de. (2018). Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. *Journal of Neuroscience*, *38*(34), 7452–7461. https://doi.org/10.1523/JNEUROSCI.3421-17.2018

Richter, D., Kietzmann, T. C., & Lange, F. P. de. (2024). High-level visual prediction errors in early visual cortex. *PLOS Biology*, *22*(11), e3002829. https://doi.org/10.1371/journal.pbio.3002829

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. https://doi.org/10.1038/14819

Roark, C. L., & Holt, L. L. (2022). Long-term priors constrain category learning in the context of short-term statistical regularities. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02114-z

Robieux, L., Thomas, C., & Botella, M. (2025). A Kind of Magic: Social Representations of Magic from Magicians and non-Magicians. *Empirical Studies of the Arts*, 02762374251324913. https://doi.org/10.1177/02762374251324913

Schellekens, W., van Wezel, R. J. A., Petridou, N., Ramsey, N. F., & Raemaekers, M. (2016). Predictive coding for motion stimuli in human early visual cortex. *Brain Structure and Function*, *221*(2), 879–890. https://doi.org/10.1007/s00429-014-0942-2

Schmuckler, M. A. (2001). What Is Ecological Validity? A Dimensional Analysis. *Infancy*, *2*(4), 419–436. https://doi.org/10.1207/S15327078IN0204_02

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain (T. Yeo, M. J. Frank, J. Snow, & J. Gallivan, Eds.). *eLife*, *8*, e46619. https://doi.org/10.7554/eLife.46619

Simon, T. J., Hespos, S. J., & Rochat, P. (1995). Do infants understand simple arithmetic? A replication of Wynn (1992). *Cognitive Development*, *10*(2), 253–269. https://doi.org/10.1016/0885-2014(95)90011-X

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074. https://doi.org/10.1068/p281059

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*(1), 12141. https://doi.org/10.1038/ncomms12141

Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., & Margulies, D. S. (2021). The default mode network in cognition: A topographical perspective. *Nature Reviews Neuroscience*, *22*(8), 503–513. https://doi.org/10.1038/s41583-021-00474-4

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92–97. https://doi.org/10.1016/j.bandc.2015.11.003

Stawarczyk, D., Bezdek, M. A., & Zacks, J. M. (2021). Event Representations and Predictive Processing: The Role of the Midline Default Network Core. *Topics in Cognitive Science*, *13*(1), 164–186. https://doi.org/10.1111/tops.12450

Sterzer, P., & Rees, G. (2008). A Neural Basis for Percept Stabilization in Binocular Rivalry. *Journal of Cognitive Neuroscience*, *20*(3), 389–399. https://doi.org/10.1162/jocn.2008.20039

Tanaka, J. W., & Presnell, L. M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics*, *61*(6), 1140–1153. https://doi.org/10.3758/BF03207619

Teichmann, L., Quek, G. L., Robinson, A. K., Grootswagers, T., Carlson, T. A., & Rich, A. N. (2020). The Influence of Object-Color Knowledge on Emerging Object Representations in the Brain. *Journal of Neuroscience*, *40*(35), 6779–6789. https://doi.org/10.1523/JNEUROSCI.0158-20.2020

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Thomas, C., & Didierjean, A. (2016). Magicians fix your mind: How unlikely solutions block obvious ones. *Cognition*, *154*, 169–173. https://doi.org/10.1016/j.cognition.2016.06.002

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. https://doi.org/10.1038/381520a0

Todorovic, A., Ede, F. van, Maris, E., & Lange, F. P. de. (2011). Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. *Journal of Neuroscience*, *31*(25), 9118–9123. https://doi.org/10.1523/JNEUROSCI.1425-11.2011

Tong, F., Meng, M., & Blake, R. (2006). Neural bases of binocular rivalry. *Trends in Cognitive Sciences*, *10*(11), 502–511. https://doi.org/10.1016/j.tics.2006.09.003

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular Rivalry and Visual Awareness in Human Extrastriate Cortex. *Neuron*, *21*(4), 753–759. https://doi.org/10.1016/S0896-6273(00)80592-9

Tootell, R. B. H., Hadjikhani, N. K., Vanduffel, W., Liu, A. K., Mendola, J. D., Sereno, M. I., & Dale, A. M. (1998). Functional analysis of primary visual cortex (V1) in humans. *Proceedings of the National Academy of Sciences*, *95*(3), 811–817. https://doi.org/10.1073/pnas.95.3.811

Triplett, N. (1900). The Psychology of Conjuring Deceptions. *American Journal of Psychology*, *11*(4), 439–510. https://doi.org/10.2307/1412365
MAG ID: 2323589457

Valenti, J. J., & Firestone, C. (2019). Finding the "odd one out": Memory color effects and the logic of appearance. *Cognition*, *191*, 103934. https://doi.org/10.1016/j.cognition.2019.04.003

Van de Cruys, S., Wagemans, J., & Ekroll, V. (2015). The Put-and-Fetch Ambiguity: How Magicians Exploit the Principle of Exclusive Allocation of Movements to Intentions. *i-Perception*, *6*(2), 86–90. https://doi.org/10.1068/i0719sas

Vandenbroucke, A. R. E., Fahrenfort, J. J., Meuwese, J. D. I., Scholte, H. S., & Lamme, V. A. F. (2016). Prior Knowledge about Objects Determines Neural Color Representation in Human Visual Cortex. *Cerebral Cortex*, *26*(4), 1401–1408. https://doi.org/10.1093/cercor/bhu224

Walton, K. L. (1978). Fearing Fictions. *The Journal of Philosophy*, *75*(1), 5–27. https://doi.org/10.2307/2025831

Wheatstone, C. (1997). XVIII. Contributions to the physiology of vision. —Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, *128*, 371–394. https://doi.org/10.1098/rstl.1838.0019

Windsor, M. (2019). What is the Uncanny? *The British Journal of Aesthetics*, *59*(1), 51–65. https://doi.org/10.1093/aesthj/ayy028

Wiseman, R. J., & Nakano, T. (2016). Blink and you'll miss it: The role of blinking in the perception of magic tricks. *PeerJ*, *4*, e1873. https://doi.org/10.7717/peerj.1873

Witzel, C., Valkova, H., Hansen, T., & Gegenfurtner, K. R. (2011). Object Knowledge Modulates Colour Appearance. *i-Perception*, *2*(1), 13–49. https://doi.org/10.1068/i0396

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, *11*(18), R729–R732. https://doi.org/10.1016/S0960-9822(01)00432-8

Wunderlich, K., Schneider, K. A., & Kastner, S. (2005). Neural correlates of binocular rivalry in the human lateral geniculate nucleus. *Nature Neuroscience*, *8*(11), 1595–1602. https://doi.org/10.1038/nn1554

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749–750. https://doi.org/10.1038/358749a0

Xu, Y. (2005). Revisiting the Role of the Fusiform Face Area in Visual Expertise. *Cerebral Cortex*, *15*(8), 1234–1242. https://doi.org/10.1093/cercor/bhi006

Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: Where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, *22*(3), 181–192. https://doi.org/10.1038/s41583-020-00420-w

Yu, K., & Blake, R. (1992). Do recognizable figures enjoy an advantage in binocular rivalry? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1158–1173. https://doi.org/10.1037/0096-1523.18.4.1158

Zacharia, A. A., Ahuja, N., Kaur, S., Mehta, N., & Sharma, R. (2020). Does valence influence perceptual bias towards incongruence during binocular rivalry? *Cognitive Processing*, *21*(2), 239–251. https://doi.org/10.1007/s10339-020-00957-9

Zaretskaya, N., & Bartels, A. (2013). Perceptual effects of stimulating V5/hMT+ during binocular rivalry are state specific. *Current Biology*, *20*(23), R919–R920. https://doi.org/10.1016/j.cub.2013.09.002

Zaretskaya, N., Thielscher, A., Logothetis, N. K., & Bartels, A. (2010). Disrupting Parietal Function Prolongs Dominance Durations in Binocular Rivalry. *Current Biology*, *20*(23), 2106–2111. https://doi.org/10.1016/j.cub.2010.10.046

Zeki, S., McKeefry, D. J., Bartels, A., & Frackowiak, R. S. J. (1998). Has a new color area been discovered? *Nature Neuroscience*, *1*(5), 335–335. https://doi.org/10.1038/1537

Zeki, S. (2015). Area V5—a microcosm of the visual brain. *Frontiers in Integrative Neuroscience*, *9*.

Zhou, W., Jiang, Y., He, S., & Chen, D. (2010). Olfaction Modulates Visual Perception in Binocular Rivalry. *Current Biology*, *20*(15), 1356–1358. https://doi.org/10.1016/j.cub.2010.05.059

Zhou, W., Zhang, X., Chen, J., Wang, L., & Chen, D. (2012). Nostril-Specific Olfactory Modulation of Visual Perception in Binocular Rivalry. *Journal of Neuroscience*, *32*(48), 17225–17229. https://doi.org/10.1523/JNEUROSCI.2649-12.2012

Publications

Title	Date of publication	Journal	Co-authors
How can we be moved by magic?	2023.11.16	British Journal of Aesthetics	Dr. Pablo Grassi, Prof. Dr. Hong Yu Wong
Hierarchical surprise signals in naturalistic violation of expectations	2025.01.24	Imaging Neuroscience	Dr. Pablo Grassi, Julius Frack, Prof. Dr. Andreas Bartels
Memory Color influences conscious object perception	in review	Neuroscience of Consciousness	Dr. Pablo Grassi, Dr. Michael Bannert, Prof. Dr. Andreas Bartels

Own contribution

1. How can we be moved by magic?:
    a) Conceptualization
    b) Writing: First draft
    c) Writing: Rewriting
2. Hierarchical surprise signals in naturalistic violation of expectations:
    a) Stimulus creation
    b) Coding of experiment
    c) Data collection
    d) Data curation
    e) Data analysis
    f) Visualization
    g) Writing: First draft
    h) Writing: Rewriting
3. Memory Color influences conscious object perception:
    a) Formulating Research Question

b) Coding of experiment

c) Data collection

d) Data curation

e) Data analysis

f) Visualization

g) Writing: First draft

h) Writing: Rewriting

Contribution of co-authors

How can we be moved by magic?

Dr. Pablo Grassi: Original idea, writing - original draft and rewriting.

Prof. Dr. Hong Yu Wong: Supervision and rewriting.

Hierarchical surprise signals in naturalistic violation of expectations

Dr. Pablo Grassi: Conceptualization, formal analysis (supporting), investigation, methodology, supervision, visualization, writing - original draft and writing - review and editing

Julius Frack: Creation of video stimuli

Prof. Dr. Andreas Bartels: Conceptualization, methodology, supervision and writing - review and editing

Memory Color influences conscious object perception

Dr. Pablo Grassi: conceptualization (equal), formal analysis (supporting), investigation (supporting), methodology (equal), visualization (supporting), writing – original draft (equal), writing – review and editing

Dr. Michael Bannert: conceptualization (equal), formal analysis (supporting), investigation (supporting), methodology (equal), visualization (supporting), writing – original draft (supporting), writing – review and editing

Prof. Dr. Andreas Bartels: conceptualization (equal), formal analysis (supporting), methodology (equal), visualization (supporting), supervision, resources, writing – review and editing

How can we be moved by magic?

Pablo R. Grassi, Vincent Plikat and Hong Yu Wong

When engaging with magic, we are moved by seemingly impossible events that contradict what we believe to be possible in the real world. We are surprised, curious, and baffled when we cannot explain how the magic we are witnessing is possible. We generally understand the events to be illusions. But how is it possible to be moved by something we know to be unreal? This problem is related to the paradox of fiction in aesthetics. Here, we introduce the problem in the domain of theatrical magic, discuss possible solutions, and present a tentative account that allows us to accommodate inconsistent, seemingly incompatible beliefs at different representational levels in the viewers' mind.

1. Introduction

When engaging with magic, we are moved by seemingly impossible events. Magic shows us incredible events that contradict background knowledge about what we believe to be possible in the world: things appear, disappear, levitate, and so on. The experience of magic evoked by these events has a unique phenomenology. Recently, it has been proposed that this state should feature as a central element of a science of magic (Rensink and Kuhn 2015) and is an emerging topic of interest in psychology and cognitive science (Lamont 2017; Camí, Gomez-Marin and Martínez 2020; Grassi and Bartels 2021). Yet, it has been largely ignored by philosophers and art critics (with the exception of Leddington 2016).

The following situations illustrate this experience:

> Josephine booked a magic show for her daughter's birthday. In an act, she is asked to inspect the top hat of the magician. She does so: it is an ordinary hat. She gives it back to the magician, who tells a short story about how she can make animals appear, shows the audience that the hat is empty and positions it carefully over a table, all in full view of the audience. The magician goes back a couple of steps, utters some magic words, and asks Josephine to check the hat again. Josephine suspects that the magician is going to make a rabbit appear. She was right. 'No way!' she exclaims, wide-eyed and surprised. She pulls a rabbit from the top hat and shows it around. 'How is this possible? That is a real rabbit. But the rabbit was not there before, right? Or was it?', she asks herself.
>
> 1983. Francisco and his partner are watching TV and see David Copperfield's announcement that he is going to make the Statue of Liberty disappear on live TV. They keep watching. They see the audience sitting at the harbour with the Statue of Liberty in front of them. A helicopter, a radar, and cameras are on set to help convince the viewers. David Copperfield then raises a giant curtain in front of the audience and lowers it to reveal that the statue has disappeared. The viewers only see a circle of

British Journal of Aesthetics https://doi.org/10.1093/aesthj/ayad026

> lights. Searchlights are used to show that the space is empty. The image in the radar representing the statue is no longer visible. Shortly after, David lifts the curtain again and makes the statue, seemingly out of nowhere, reappear. Francisco cries out, surprised, 'Wow!', turns to his partner and asks: 'Did you see that? Is this real?… It is obviously not real… it is just a magic show, but… how?' Francisco is baffled by the inexplicable event.

Josephine and Francisco are surprised by the seemingly impossible events. But they do not believe in 'real magic'. They both know that these were just tricks (although they are ignorant of how the tricks work). Furthermore, they perhaps even knew what was about to happen (because it was easily predicted or was announced in advance). Intriguingly, they enjoy and are moved by magic, beyond their curiosity about how the tricks work, even though they know that they are seeing illusions and predicted what would happen in advance. There is a puzzle about magic here.

The experience of magic involves an event the viewers have a sense of witnessing, but that, at the same time, believe to be impossible. Surprisingly, the viewers of magic are brought into this cognitive conflict although they know the performance to be illusory—just trickery. How is it possible that they are genuinely moved by a performance they believe not to be real? This problem is related to the 'paradox of fictional emotions', which asks how viewers are moved by fiction, if one is only moved by things believed to be real (see Radford and Weston 1975; Walton 1978; Lamarque 1981; Carroll 1990; Stecker 2011; Tullmann and Buckwalter 2014; Friend 2020, among others).[1]

To address this problem of the engagement with magic, we will first present the experience of magic as a set of epistemic emotions (surprise, curiosity, and bafflement) evoked by a sequence of sensory events (i.e. the performance unfolding in time) that violate our prior knowledge about the world (Section 2). We will then introduce the problem of theatrical magic. We consider and reject the following putative explanations: that the viewers of magic are moved because they do not know how a trick is done; or because they entertain a non-assertive thought of something impossible as happening; or because of habituated emotional responses. Instead, we argue that viewers of magic are moved because of the co-existence of the conflicting beliefs that the impossible events are really occurring and that the events are illusions (Section 3).

2. Epistemic Emotions in Magic

Magic is about violating expectations: Francisco knows that the Statue of Liberty cannot disappear and yet sees it disappear, and Josephine knows that objects do not appear in

1 To the best of our knowledge, Jason Leddington's work is the first to discuss magic in context of aesthetics and related problems, such as the paradox of fictional emotions, especially in an unpublished section in an early draft of 'The Experience of Magic' (Leddington 2016). There he also describes a further problem that asks why people actually *enjoy* magic if they are being fooled and tricked (see also Leddington 2017, 2020). This problem is related to the paradox of horror and will not be discussed here.

empty hats and yet observed a rabbit appear in one. They are moved by seemingly impossible events that contradict their prior knowledge about the world. They find themselves in a cognitive conflict between something they know cannot happen and yet see happening before their eyes.

The concept of impossibility plays an important role in understanding magic. Indeed, the experience of magic is often described as: 'the experience of wonder generated by perceiving an apparently impossible phenomenon' (Rensink and Kuhn 2015: 10), 'the response to a seemingly impossible event' (Lamont 2017: 4), and 'an experience as of an impossible event' (Leddington 2016: 254). Accordingly, we follow a common understanding of magic as 'creating illusions of the impossible' (Nelms 1969, cited in Rensink and Kuhn 2015: 1).

But just what sense of impossibility are we concerned with? These events are neither logically, physically, nor technologically impossible. What sense of impossibility is it then? They are events the viewers, given their knowledge and perspective, 1) believe to be impossible and yet 2) believe to be unfolding before their eyes. They are *epistemic impossibilities* (that is, possibilities which are inconsistent with what we know; Edgington 2004). For example, observing a normal-looking die rolling exactly a predicted series of numbers by a magician is perceived as something 'impossible', only when the observer is unaware of the existence of loaded dice. To be clear, some events shown in many magic tricks are not strictly impossible, but just highly improbable (e.g. correctly *naming* a randomly selected card). However, viewers might think of them as 'impossible', because, in context of a magic show, viewers will be compelled to believe that the magician determined the outcome of the allegedly stochastic process (e.g. correctly *predicting* a randomly selected card), and thus creating the illusion that she knows something it should be impossible for her to know—given the information the viewers are presented with during the trick (e.g. the deck was fairly shuffled, the card was freely selected, etc.).[2]

Thus, when talking about 'impossible events' here, what we mean is that they are epistemic impossibilities *in the context*. This does not imply that they are actually impossible events and is also not meant to exclude improbable events that are considered to be impossible within the context of a magic performance (e.g. the correct *naming of a card* presented as an *informed prediction*).

With this in view, we can ask: what kind of responses do these events evoke? We think the responses can be best characterized as threefold. First, in view of incredible occurrences, viewers are *surprised*: they drop their jaw, raise their eyebrows, make exclamations ('*Wow!*'), and so on. Then, viewers show *curiosity* about the nature of the events: they look for solutions, question others, reconstruct the events and so on. But in most cases,

2 That is why magicians often provide false causal explanations to amplify the perceived impossibility of the attempted feats and misdirect their audiences. Many tricks are theatrically introduced as the product of supernatural abilities (e.g. 'I am going to read your mind'), supernatural objects (e.g. 'I will now use this magic wand to …'), or simply accompanied by magic gestures or formulations implying a form of supernatural causality (e.g. 'I will predict your selection', 'Abracadabra!') so as to provide a framing for these as impossible causes for the audience.

they will fail to explain away what they witnessed, which, in turn, evokes *confusion* (or bafflement). These mental states—surprise, curiosity, and confusion—can be referred to as 'epistemic emotions', as they involve affect triggered by situations of cognitive incongruity (Pekrun et al., 2017). Surprise is the mental state evoked by events that contradict our expectations and violate prior beliefs (Berlyne 1960; Barto, Mirolli and Baldassarre 2013). Curiosity is the motivational state that drives us to obtain information to explain the contradictory events by stimulating information-seeking behaviours (Berlyne 1954; Loewenstein 1994). And confusion (or bafflement) is the mental state that occurs when we fail to resolve the conflict (Vogl et al., 2019). Accordingly, magic moves audiences, first, because of the discrepancy between the viewers' expectations and what they observe, second, because they will try to close their gap in knowledge and integrate the new evidence into their understanding of the world, and, finally, because (most of them) will fail to do so. We believe these three epistemic emotions to be central to the aesthetics of magic, akin to what the role fear plays in horror and laughter in comedy.

The cognitive incongruity that evokes these epistemic emotions can be thought of as a situation in which the viewers believe (1) that '*p* cannot happen' (and will therefore not happen) and believe, based on the observation of the performance unfolding in time, (2) that '*p* happened' (Lamont 2017; Kuhn 2019). The first belief ('*p* cannot happen') is a belief based on background knowledge (e.g. Francisco knows that statues do not disappear and is therefore implicitly not expecting the Statue of Liberty to disappear). The second belief ('*p* happened') is a newly acquired empirical belief based on presented sensory evidence during the performance (e.g. Francisco stops seeing the Statue of Liberty, and believes that the Statue disappeared, as it was just there a moment ago, the radar image disappeared, etc.).[3] Arguably, all magic performances aim for this kind of conflict—David Copperfield wants his audience to believe that he made the Statue of Liberty disappear, as much as the magician in Josephine's party wants her to believe that a rabbit appeared in a hat—knowing that the audiences consider these occurrences to be impossible. Hence, we can think of *illusionists* as sequentially presenting strong perceptual evidence using different methods of misdirection to create false beliefs about a situation that contradicts prior knowledge (e.g. giving false explanations, sleight of hand, etc.) (Kuhn et al., 2014; Camí, Gomez-Marin and Martínez 2020; Grassi and Bartels 2021).

The experience of magic can then be thought of as an experience evoked by a profound mismatch between prior beliefs based on what we know and what we are currently witnessing (Fraps 2014; Grassi and Bartels 2021). The surprise induced by magic is *special* as it involves the apparent violation of deeply held beliefs about the general nature of the world by occurrences we directly perceive in the real world (e.g. seeing something

3 We consider this second belief a component of the cognitive incongruity in magic, first, because when we perceive that *p*, we also generally come to consciously believe that *p* and second, because acceptance of the sequentially presented sensory evidence is required to be epistemically moved by it (see Section 3.4). Importantly, the content of this observation-based belief is the inferential product of a series of actual perceptual events (e.g. the Statue of Liberty has disappeared), and not a more generic belief (e.g. believing that objects can disappear).

levitate in front of us and seemingly violating the law of gravity, or seeing an occluded object disappear and seemingly violating object permanence).[4] Thus, it is the higher-level cognitive mismatch between an anomalous sensory-based belief (e.g. believing that a rabbit appeared in a hat *thought to be empty based on the sequence of events witnessed*) and our general understanding of the world (e.g. knowing that solid objects do not appear out of thin air) that is specific to the experience of magic. This allows us to differentiate between mundane surprise induced by violations of simple sensory expectations due to unexpected events—like a startle response evoked by an unexpected insect—and surprise induced by inexplicable events in magic, where we experience things we believed to be impossible.

3. The Paradox of Theatrical Magic

Most audiences know they are being deceived when they go to a magic show and know that they are not observing 'real magic' but illusions. If viewers are asked about a magic show afterwards, most will never assert that they experienced 'actual' supernatural events. Moreover, viewers do not tend to behave as if what they experience were real. When a magician convincingly saws someone in half on stage, viewers generally do not call the police, try to prevent it, or run from the theatre. They understand the situation to be illusory or fictional. And yet, they are surprised, curious, and baffled, and they respond in ways rather different to the emotional responses to magical characters in fictional environments (like Yoda, Gandalf, or Dumbledore). Peter Lamont describes this as follows: 'The experience of magic … is a response to an event that: (a) one is convinced cannot happen; (b) one is convinced does happen; and (c) one understands to be an illusion'. (Lamont 2017: 4).

How is this possible? Background knowledge about the non-actual nature of the events should neutralize any epistemic conflict. When we know that the discord is not real, neither dissonance nor epistemic reactions (surprise, curiosity, confusion) seem to be appropriate. Normally, if someone were to tell us something that contradicts our beliefs, but we are sure that it is wrong or is a lie, we would not epistemically care (even though we might, for example, be offended). If we were to see something that contradicts what we know but realize it is not real, then we would stop worrying. If this is the case, why is the audience of magic *epistemically moved* by a performance they understand to be illusory?

4 Note that 'surprise' in context of information theory refers to the sub-personal improbability of an observation given a hypothesis or model of the world (sometimes called 'surprisal' or 'Shannon information') and not to personally experienced surprise. Here, we use the latter, experiential sense when speaking about surprise. Surprisal may be in discord with personally experienced surprise (Clark 2013, 2018). As Clark writes: 'the percept that, overall, best minimizes surprisal (hence minimizes prediction errors) "for" the brain may well be, for me the agent, some highly surprising and unexpected state of affairs—imagine, for example, the sudden unveiling of a large and doleful elephant elegantly smuggled onto the stage by a professional magician'. At a sensory level, '[t]he sight of the doleful elephant may then emerge as the least surprising (least "surprisaling"!) percept available, given the inputs, the priors and the current weighting on sensory prediction error. Nonetheless, systemic priors did not render that percept very likely in advance, hence (perhaps) the value to the agent of the feeling of surprise' (Clark 2013: 196).

In many respects, this problem is reminiscent of the 'paradox of fictional emotions' (Radford and Weston 1975; Walton 1978). There, the problem is: how is it possible that people experience emotions towards fictional objects, even though they do not believe they exist? In magic, we can describe the problem as follows:

(A) People have a cognitive incongruity when experiencing something seemingly impossible (*cognitive incongruity premise*).
(B) Audiences of magic know that the seemingly impossible events are not real, but illusions (*knowledge premise*).
(C) People do not have a cognitive incongruity when they know that the seemingly impossible events are not real, but illusions (*incompatibility premise*).

Alone, each of the statements sounds plausible and intuitively true, but together they are incompatible. We call this the *paradox of theatrical magic*. Note that the problem of magic is distinctive as it deals with the question of why we are moved by illusions—and not by fiction. While these two notions are not unrelated, they differ in as much as illusions present events as actually happening in the 'real' world, while fiction invites the viewer to engage with events in a fictional world (cf. Leddington 2016).[5]

For example, during the birthday party, Josephine is moved because she witnessed a rabbit appearing in a hat she thought to be empty. And while the occurrence of a rabbit is real and does not involve any visual illusion of sorts (i.e. it is a real rabbit), the occurrence of a rabbit *appearing* in the magician's hat is a *deceptive appearance*, a trick, an illusion.[6] Thus,

5 Please note that we do not consider borderline cases of extraordinary feats that are not strictly illusions (or are not framed as such) here (like eating glass, escaping from a vault, etc.) and focus on more classic cases of magic instead. These cases of real feats are not against our account. Most can be accounted for by reformulating the knowledge premise (B) to 'Audiences of magic *believe* that the seemingly impossible events are not real, but illusions'.

6 When talking about illusions here, we are not referring to perceptual illusions (e.g. 'visual illusions' in a narrow sense, like the Müller-Lyer illusion or the Ebbinghaus illusion). Instead we are employing a more general understanding of illusions as that of *deceptive appearances in the real world*. Accordingly, a magic illusion is a performance that, when successful, compels the viewers to misrepresent the actual states of affairs in the world and create false observation-driven beliefs as of something impossible happening. Clearly, the rabbit Josephine found in the hat is a real rabbit; there is no 'visual illusion' involved in the trick (at least not in a narrow sense). It is rather the sequence of ordinary sensory events (e.g. seeing an empty hat) that generates a set of unproblematic automatic beliefs (e.g. 'the hat is empty'), which are in turn contradicted by actual events (e.g. witnessing a real rabbit in the hat) that constitute the *illusory* event as of a rabbit magically appearing in a hat (cf. Smith, Dignum and Sonenberg 2016). Accordingly, the rabbit is not the illusion, but rather the event of witnessing a rabbit in a hat thought to be empty. These illusory events are more complex than purely 'perceptual' illusions as they require more cognitive processes (i.e. memory and inference). Viewers are epistemically moved by a rabbit in a hat if and only if they had the prior belief that the hat was empty (cf. Grassi and Bartels 2021). These illusions can be thought of as 'cognitive' illusions (Macknik et al., 2008). Indeed, some illusions are based on more perceptual illusions (e.g. the Zig-Zag illusion in which a magician seemingly divides a participant into thirds), while others involve short-term memory (e.g. a magician making a rabbit appear in a hat by presenting a sequence of deceptive events) or cognitive processes (e.g. a magician using cognitive tricks to create the illusion she can read minds), with most tricks involving a combination of these.

the illusion here consists in the deceptive, sequential presentation of strong evidence (i.e. the story, showing the empty hat, etc.) to create the illusion of a solid object appearing from out of nowhere, conjured by uttering some words. Thus, Josephine has a cognitive incongruity and is moved by a seemingly impossible event (A), even though she knows that objects cannot appear out of nowhere, she knows that the magic words used by magicians have no real causal power, and knows that she is witnessing an illusion (B). And yet, it is puzzling that she is epistemically moved by something she knows to be a deceptive appearance (C).

In what follows, we will accept premises (A) and (B) and argue that the paradox arises from accepting (C). Thus, the audience of magic is seen as recognizing that the events are illusions (B) and as genuinely experiencing a cognitive incongruity (A) (see Section 2) and not just make-believe epistemic emotions, as make-believe theories of fiction might argue (see e.g. Walton 1978). Our proposal consists of rejecting (C) by drawing on a nuanced understanding of belief and thereby accepting co-existing conflicting representations when engaging with magic. Before we introduce our proposal below (Section 3.4), we will describe three alternative attempts to understand why we are moved by magic, targeting (C). The first attempt is the intuitive answer that we are moved by magic because even if we know the events to be illusions, we do not know how they work (Section 3.1). The second attempt is to apply a familiar approach to the paradox of fiction to our engagement with magic and argue that we do not need to believe in the existence of things to be moved by them (Section 3.2). Finally, the third attempt refers to an 'emotional' belief of sorts associated with habituated responses to explain why we are moved by magic (Section 3.3).

3.1 Epistemic Reactions to Fictional Puzzles

A first answer to the problem of theatrical magic would note that knowledge that the performance is illusory does not resolve the cognitive conflict. For instance, Josephine *knows that* the rabbit did not magically appear in the hat, but she does not *know how* the rabbit got there. Crucially she has no answer to this question, for if she either knew the workings of the trick beforehand or she thinks she has figured out how the trick works, then the cognitive conflict—and with it the experience of magic—would disappear.

Ignorance about how a trick is done is a necessary condition for the epistemic emotions constitutive of the experience of magic. Accordingly, one may reject the *incompatibility premise* (C) by arguing that the spectators are epistemically moved by the cognitive conflict magic tricks evoke—not because they *believe something impossible to be happening*, but because they have no means of explaining how something impossible *appears to be happening*. As the spectators know and assume their background beliefs to be true, they are epistemically moved by the contradictory sensory occurrences because they fail to find means to explain them.[7] Magic tricks could then be thought of as cognitive problems

7 Leddington's recent account of the experience of magic goes roughly along these lines. Magic resists intelligibility as the viewers not only fail to understand how a trick is performed, but they see no means by which the effects *could* be produced. Accordingly, Leddington describes the viewers' attitude as: 'There *must* be an explanation, but I have no idea how there could be. All possibilities seem to have been exhausted' (Leddington 2016: 261).

that we find pleasure in engaging with, despite us knowing that they are not real, not dissimilar to fictional mysteries. Indeed, some viewers might report something like a 'problem-solving' attitude when engaging with magic ('*I know it is a trick, but I still want to know how it is done*').

However, while this is a possible attitude the viewers might have towards magic performances and thus present a genuine solution to the problem, we believe this is neither a standard attitude of spectators, the experiential state (most) magicians want to bring their viewers into, nor, as we will argue below, an adequate description of the belief-acquisition processes involved in the experience of magic and behavioural responses thereof. For puzzles may elicit curiosity, but less so surprise. In the best case, such an attitude would provide curiosity and bafflement for those viewers who find pleasure in engaging with puzzles. In the worst case, this attitude might very well prevent the viewers from getting caught in cognitive dissonance and so prevent them from being moved by magic altogether, which is why magicians are specifically trained to deal with this kind of attitude in viewers. Most people want first and foremost to be surprised and amazed with magic. In a questionnaire about magic, only 10 per cent of the responders said that what they liked more about magic was trying to find out how the effects were done, while most of the responders (25 per cent) liked the element of surprise the most, and if given the chance, they preferred to watch new tricks than to get the solution to a trick they watched (60 per cent vs 40 per cent; Jay 2016). Often viewers do not want to know how the tricks work and the magic to be 'spoiled'. Josephine and Francisco are not moved because they fail to solve a puzzle; they are moved by events they believed to be impossible and yet believe to be happening.

3.2 Epistemic Reactions to Non-assertive Thoughts

Alternatively, we may reject the *incompatibility premise* (C) by denying that emotional responses require belief in the actuality of the events and/or object of emotions, which is also the most uncontroversial solution to the paradox of fiction. These so-called 'thought theories' claim that emotions can be triggered by non-assertive 'mental representations or thought-contents' (Lamarque 1981: 296) or by 'the content of thoughts entertained' (Carroll 1990: 88). We can fear the thought of Dracula, even if we do not believe that Dracula exists (Carroll 1987). Similarly, one may argue magic moves us because we entertain the thought of impossible and inexplicable events even if we do not (necessarily) believe the events to be actual.

However, while we grant that some emotions may be triggered by mental representations independently of explicitly endorsed existence beliefs (e.g. fearing Dracula), the specific epistemic emotions evoked in magic are different in at least two important respects. First, the specific epistemic emotions evoked in magic cannot be evoked through imaginary, fictional works, even with events that would seem broadly similar. For example, our emotional response to seeing Luke Skywalker levitate something in the movie 'Star Wars' is completely different to that of seeing a magician levitate objects in a magic show, because the latter is targeting our understanding of the workings of the real world,

while everything is possible in fiction.[8] Thus, to be epistemically moved by seemingly impossible events in the way characteristic of the experience of theatrical magic, we need to accept that the relevant events are happening in the real world and within the known constraints of the real world. Clearly, one can still be surprised, curious, or baffled by unexpected changes in the narrative content of fictional works, but this is not the kind of surprise, curiosity, and bafflement involved in appreciating theatrical magic.

Second, and in contrast to other emotions, it seems unlikely, if not impossible, to evoke epistemic emotions that depend on mental states representing the world as being thus-and-so (i.e. expectations and beliefs)—such as surprise, curiosity, or bafflement—by voluntarily entertaining a thought. While one might fear products of the imagination, it would seem that you cannot surprise or baffle yourself. Perhaps epistemic emotions can be triggered by a chain of thoughts where one draws a series of inferences to a surprising or baffling conclusion, but this is not the kind of thought process involved in the experience of magic. Thus, the experience of magic cannot come down to that of imagined non-assertive thoughts of impossible events. Josephine and Francisco are not moved by tokening impossible events like 'a rabbit appeared' or 'the Statue has vanished' in their thoughts but moved by *actual events* that they witness and accept as happening in the real world.

3.3 Epistemic Reactions as Habituated Responses

One may further argue that while Josephine and Francisco might 'intellectually' know that the events are impossible and even recognize them to be illusions, they perhaps cannot help believing that what they saw happened at an 'emotional' level. They could have a 'gut' or 'emotional' belief of sorts that they do not endorse, but that pushes them to respond as if it were true. Then, as they cannot help but respond in this habituated way, one may again reject the *incompatibility premise* (C) to solve the problem.

Leddington (2016) has argued along this line in his recent analysis of the experience of magic (although as a description of the experience of magic and not as a solution to the paradox of theatrical magic, see also Footnote 7). Leddington suggests that the cognitive conflict constitutive of the experience of magic is not a matter of two contradictory 'intellectual' beliefs, but rather between an 'intellectual belief' and an 'emotional belief' of sorts. Leddington characterizes the latter following the notion of 'belief-discordant alief' from Tamar Gendler, as a mental state that is distinct from belief such that it is automatic (i.e. it operates without the intervention of conscious reflection), does not involve endorsement, and consists of an associative link of representational [R], affective [A], and behavioural [B] components (Gendler 2008b, 2008a). Leddington points out that belief-discordant aliefs are akin to the cognitive conflict we encounter in magic, in which we are automatically moved by an event we do not consciously endorse and know to be impossible.

8 Cf. Leddington's draft to 'The Experience of Magic' (Leddington 2016).

However, aliefs are overly restrictive. Some tricks that are dependent on visual appearances may induce alief-like visceral responses, like seeing a hand pierced by a knife—'[R] *A pierced hand!* [A] *Shock!* [B] *Close your eyes!*'—even if we consciously believe that no one is being hurt. But many magic tricks that end in ordinary sensory events do not involve any distinctive habituated response (e.g. a coin in a hand or a ball under a cup). For while we might have some habituated associative responses to coins, balls, and cups, those responses are not what we have in mind when talking about the experience of magic.

But there is a more important problem with this account. Aliefs are supposedly reality-resistant or not 'evidence-sensitive' (Gendler 2008b: 566). That is why, following Gendler, we experience vertigo at heights and disgust in view of faeces-shaped fudges, even if we have good intellectual reason to do otherwise. She suggests a simple principle to differentiate between beliefs and aliefs:

> Beliefs change in response to changes in evidence; aliefs change in response to changes in habit. If new evidence won't cause you to change your behavior in response to an apparent stimulus, then your reaction is due to alief rather than belief.
> (Gendler 2008b: 566)

As we have argued, magic relies on people not *knowing how* the illusion works. Prior knowledge, a strong suspicion, or even a clumsy sleight of hand is enough to destroy a trick. As our responses are (indeed) evidence-sensitive and the audience is curious, reference to aliefs is inadequate to account for the experience of magic. For if the experience of magic were to be accounted for by an alief (i.e. an automatic, habituated, and evidence-insensitive chain of associations), our experience should not change in view of new evidence. But it is clear that it can and does.[9]

3.4 Epistemic Reactions to Illusions

We agree with Leddington in invoking automatic attitudes that are not necessarily endorsed to understand the conflict that magic evokes. But a more nuanced understanding of beliefs that sees them as being (sometimes) inconsistent and automatically accepted might provide a stronger account.

As it happens, we are often in situations involving conflicting beliefs, half believing something, or uncertainty. David Lewis famously describes how he believed that the Nassau Street in Princeton ran roughly east-west, that the railroads nearby ran roughly north-south and that both were roughly parallel (Lewis 1982). After realizing the inconsistency, he corrected his view and believed that both run roughly northeast-southwest.

---

9 Note that Leddington uses a loose understanding of alief, considering only its representational content, and leaves aside the associatively linked affective and behavioural components. Accordingly, arguments against a strong alief state as a four-place relation may be unsuitable to address a loose understanding of alief as a two-place relation. If the representational content of aliefs is propositional and akin to, for example, a perceptual belief (Mandelbaum 2013), then Leddington's account can be made compatible with the notion of the experience of magic as evoked by discordant representations (Grassi and Bartels 2021) and the account proposed here.

Similarly, people believe in natural and supernatural explanations for the same phenomena (Legare and Gelman 2008), confabulate reasons for beliefs opposite to their original opinions (Johansson et al., 2005), hold implicit beliefs, such as racial or gender biases, that are unendorsed and uncorrelated to explicit ones (Greenwald, McGhee and Schwartz 1998), and believe in outright contradictions, such as believing that Princess Diana was murdered while at the same time believing that she faked her own death (Wood, Douglas and Sutton 2012) or providing both psychic (i.e. paranormal) and conjuring ('just a trick') explanations for the same mentalism performance (Lesaffre et al., 2018).

This is possible because belief is more fragmented than we have realized (Borgoni, Kindermann and Onofri 2021). Even if there are norms of belief revision that require the web of beliefs to be consistent, the actual architecture of belief is fragmented. Beliefs need not to be at the forefront of the mind at the same time; they can be graded (i.e. we can believe with more or less confidence that *p*), they can be held at different levels of our mental architecture (e.g. there are more or less abstract beliefs, inferential beliefs, moral beliefs, perceptual beliefs, etc.), and they need not be consciously endorsed (e.g. we can have implicit beliefs). We may thus have inconsistent beliefs, as beliefs need not be held in a single coherent web of belief (Lewis 1982; Egan 2008; Mandelbaum 2014; Borgoni, Kindermann and Onofri 2021).

Then, we may resolve the problem of theatrical magic by accepting that the viewers of magic believe in both: the events to be actually happening, based on the automatic acceptance of perceptual evidence (i.e. the sequence of deceptive events unfolding in time), and the events to be illusions, based on contextual background knowledge, like knowing one is witnessing an illusionist.[10] Viewers need to hold the first belief, because to be *epistemically* surprised, curious, and confused by an event in magic requires that they believe in the actuality of the event. This requires that the observation-based belief that '*p* happened' come into conflict with the higher-level, deeply held background beliefs about this world that deemed it impossible.[11] However, conflictingly, viewers believe the events to be illusory, as most viewers of magic will say so if asked and most people do not behave as if the observed events were real (e.g. they show no complex voluntary behaviour when somebody is convincingly impaled on stage)—this is like the general case of responses to fiction. As the audience is thought to be (at least initially) under the illusion that what they observe is real, we can think of this as an 'illusion approach' to magic, in line with what Noël Carroll calls the 'illusion theory of fiction' (Carroll 1990).

---

10 In line with this, the psychologist Peter Lamont (2017) describes the experience of magic as arising from contradicting beliefs based on different psychological processes (memory and perception) bounded by the 'awareness of the wider context' of knowing one is watching an illusion, so that: 'The experience of magic, then, might be seen as a form of astonishment that is bounded' (Lamont 2017: 6).

11 Conversely, the belief that 'in fiction, *p*' does not come into conflict with deeply held background beliefs about *this world*. It may evoke emotions, but not epistemic emotions related to a conflict with deeply held beliefs about *this world*. That is why we are moved differently by magic as to how we are moved by characters such as Dumbledore, Gandalf, and Yoda. See also Section 3.2.

In the context of the paradox of fiction, 'Illusion' approaches that think of the viewers as somehow believing that fictional events and objects are real are generally rejected because the proposed mechanisms appear to misdescribe the state the consumers of fiction are in (e.g. doubting, momentarily forgetting or suspending the belief that they are engaging with fiction) and, most importantly, because they cannot explain why people do not behave as if the fictional events were real (Radford and Weston 1975; Walton 1978; Carroll 1990). For example, Kendall Walton describes how Charles, who fears a fictional slime from a horror movie:

> has *no* doubts about whether he is in the presence of an actual slime. If he half believed … we would expect him to have *some* inclination to act on his fear in the normal ways. … Even a hesitant belief, a mere suspicion, that the slime is real would induce any normal person seriously to consider calling the police and warning his family. … He is not *uncertain* whether the slime is real; he is perfectly sure that it is not. (Walton 1978: 7)

As the same point applies to magic, one may wonder if it is justified to think that the viewers of magic succumb to the illusions presented. What are the differences? Contrary to the normal consumption of fictional works (e.g. reading a book or watching a movie), magic is completely integrated in the viewers' world so that the presented sensory occurrences that unfold in the 'real world' are arguably automatically accepted.[12] The audience in a magic show is (initially) under the illusion that what they observe is real: when Josephine sees a rabbit in the top hat she believed to be empty, she immediately and effortlessly comes to believe that a rabbit appeared in the top hat; and when Francisco sees that the Statue of Liberty is not where it used to be he immediately and effortlessly comes to believe that the Statue of Liberty disappeared. Knowledge that the perceived events are illusory does not interfere with the initial acceptance of the observations. The acquisition of beliefs based on direct experience is automatic and effortless. In the common case, we take witnessed events at face value and accept the existence of objects upon seeing them. Only after the initial acceptance of the outputs of our automatic sensory and cognitive inferential systems can we consciously and reflectively evaluate them.[13]

Immediate appraisal of the accepted events against background beliefs will evoke surprise (Francisco: '*Wow!*', Josephine: '*No Way!*'), which in turn brings the viewers to reflectively try to reject or re-evaluate the newly acquired information to make it fit background

12 While illusion theories might seemingly have difficulties in explaining how people come to believe in fictional objects and events to allow for genuine emotions in a media such as text (Carroll 1990), it is less obvious that such arguments hold for cases of immersive fictional experiences in real and virtual environments.

13 The outcomes of our perceptual system are accepted by default because the 'unconscious' perceptual inferences already represent the 'real' world as being in a certain way—namely, whatever best explains the restricted incoming sensory information (Helmholtz 1867; Gregory 1980; Clark, 2013; Hohwy 2014). A system that requires prior knowledge to fill in the gaps when inferring hidden world states is already subpersonally committing to an interpretation. In a predictive mind, seeing is believing. Similarly, a more general understanding of belief formation, the Spinozan model, suggests that all novel information is not subject to any evaluation prior to being believed and that rejection requires an additional, retroactive step (Gilbert 1991; Egan 2008; Mandelbaum 2014).

knowledge through curiosity-driven behaviours: we may express hesitancy about what is the case, reflect upon the wider situation to reassure ourselves that what we perceive is an illusion (Francisco: '*It is just a magic show*'), search for explanations (Josephine: '*How is this possible?*'), look for epistemic support from our co-viewers (Francisco: '*Did you see that?!*'), question our senses, reconstruct and double-check our memories and wonder if what we observe is 'really' happening (Francisco: '*Is this real?!*', Josephine: '*But the rabbit was not there before, right? Or was it?*'). We do all of this as if we would be truly illuded by the show, as if we would have temporarily forgotten or are uncertain that what we are perceiving is only an illusion, for we find no means to explain away the sensory events and resolve the conflict of beliefs. In a way, the compelling sensory occurrences in magic shows make the audience unwillingly 'suspend the disbelief' that the events are non-actual.[14] Francisco and Josephine are—to some extent—uncertain and doubtful if the events are in fact just tricks, even though they will most probably still assert so if explicitly asked.

In magic, we accept the existence of unexpected events and objects upon witnessing them, and so it is natural to be moved by them. Contextual knowledge that the events are illusions prevents neither initial acceptance of the observations nor being epistemically moved by them. And what is more, audiences are not constantly thinking that the events are illusions and do not hold this belief with absolute certainty.[15]

Together, it appears that commonly rejected 'illusion' notions in context of the paradox of fiction (e.g. doubting, forgetting, 'suspension of disbelief') are actually feasible mechanisms to explain why we are moved by magic. And while it is a matter of empirical enquiry to pinpoint the exact mechanisms involved, the 'illusion' approach outlined here can still capture our natural responses to magic (surprise, curiosity, bafflement) by highlighting the role of the beliefs we effortlessly acquire when engaging with magic illusions. But, if we think of the viewers of magic as moved by events they somewhat believe to be happening, why do they not behave accordingly?

As it happens, having a belief in the background does not mean the belief is idle, nor does it mean that the belief has been dismissed. Background contextual knowledge that the events are illusions may still allow the viewers to behave appropriately and may come to the forefront of the mind if necessary (e.g. when viewers recall '*it is just a magic show*'). Accordingly, when a magician convincingly saws someone in half, we may be epistemically moved by the events we believe to be happening, but contextual knowledge that we

---

14 Note that 'suspending disbelief' that the events are illusory is different to the idea of suspending disbelief in the impossibility of the events. In the latter case, Josephine could, for instance, be suspending her disbelief that objects cannot appear out of thin air. However, if Josephine suspended her background disbelief in the impossible, she would not be moved by the events. Believing in the impossible would destroy the experience of magic (cf. Leddington 2016; Lamont 2017; Kuhn 2019).

15 Even the deeply held beliefs about what is possible are at stake when observing magic. People often update their beliefs to accommodate the events they experience (e.g. by ascribing magicians supernatural powers). Throughout the history of magic, charlatans have exploited the audience's disposition to believe in the paranormal and presented their feats as supernatural abilities. Simple framing of magic as psychism (Mohr, Koutrakis and Kuhn, 2015; Lesaffre et al., 2018) or verbal suggestion (Wiseman and Greening 2005) are enough to induce paranormal beliefs.

are in a show observing illusions will (generally) block any urge for performing complex deliberate actions, like calling the police. The inconsistent beliefs guide our behaviour differentially, with contextual knowledge in some respect controlling behaviour (e.g. via high-level executive functions). Thus, a more nuanced view on beliefs allows us to account for the behavioural disanalogy present in magic in which the viewers usually do not behave in accord to the belief that the events were real.

It is the interplay of the two incompatible beliefs that helps us to resolve the paradox of theatrical magic and what captures the cognitive dynamics and the experience of magic. On the one hand, we have immediate evidence of what we see (e.g. that the rabbit is in the hat). The beliefs that are automatically generated based on direct experience violate deeply held beliefs about the general nature of the world (hyperpriors; e.g. that objects do not appear out of nowhere) (Fraps 2014; Grassi and Bartels 2021). This conflict elicits the distinctive battery of epistemic emotions of surprise, curiosity, and confusion that are characteristic of the experience of magic. On the other hand, high-level contextual knowledge that the events are not real sets the scene for the magic show and frames the considered epistemic and behavioural situation. We are witnessing a magic show where compelling illusions are presented. The tension between these two incompatible beliefs held at different levels is what gives us the cognitive dynamics and the experience of magic.[16]

In sum, we suggest the audiences hold two incompatible beliefs when engaging with magic: the belief in the actuality of the events based on automatic belief-acquisition processes, together with the contextual belief that it is an illusion residing a different representational level (cf. Lamont 2017). The effortless acquisition of beliefs based on direct experience about the occurrences in a magic show ('*p* happened') will evoke surprise, curiosity, and bafflement when appraised against background knowledge and expectations ('*p* cannot happen'), while higher-level contextual knowledge (e.g. 'the occurrence of *p* is an illusion') will diminish any inclination to act according to the lower-level belief in the actuality of the events and will be, if not sufficiently weakened by the compelling sensory evidence, asserted if asked for.

4. Conclusion

Josephine and Francisco are moved by seemingly impossible events, even though they know them to be tricks. Intriguingly, this contextual knowledge does not prevent them from being epistemically moved by the events (surprised, curious, confused). As common strategies to

16 Note that our proposed solution to the 'paradox of theatrical magic' builds upon and is consistent with recent research in psychology and cognitive science that appeals to the violation of prior knowledge and (implicit) expectations or to a conflict of beliefs to explain our reactions to magic (Smith, Dignum and Sonenberg 2016; Lamont 2017; Kuhn 2019; Grassi and Bartels 2021). In line with these accounts, we understand the experience of magic as arising from a conflict between a prior-knowledge-based expectation and an accepted observation-based belief (cf. Grassi and Bartels 2021, see Section 2) and follow recent developments in cognitive science that understand beliefs as fragmented in order to address the problem (Borgoni, Kindermann and Onofri 2021).

the paradox of fiction fail to adequately account for the problem of theatrical magic, we propose a psychologically informed solution based on the automatic acceptance of the observed events. When engaging with magic, viewers hold distinct world-tracking beliefs such as believing that the events actually happen and believing that they are illusions. This approach accommodates the conflicting beliefs at different levels of our mental architecture. Effortlessly acquired beliefs based on direct evidence evoke, when appraised against background knowledge, the epistemic emotions of surprise, curiosity, and bafflement. In turn, contextual knowledge about the situation will diminish any inclination towards complex behaviour of the sort that the beliefs would elicit if the events were taken to be unquestionably real.

We believe this approach reflects the distinctiveness of the aesthetic of magic and offers a new perspective to the paradox of fiction in cases of immersive and deceptive artworks that present events as occurring in the real world. Magic moves us because of the compelling presentation of events in the real world we believe to be impossible, which remain largely unaffected by us knowing them to be illusions.[17]

Pablo R. Grassi
Department of Psychology, University of Tübingen, Schleichstrasse 4, 72076 Tübingen, Germany
Centre for Integrative Neuroscience, Otfried-Müller-Strasse 25, 72076 Tübingen, Germany
Max-Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, 72076 Tübingen, Germany
pablo.grassi@cin.uni-tuebingen.de

Vincent Plikat
Department of Psychology, University of Tübingen, Schleichstrasse 4, 72076 Tübingen, Germany
Centre for Integrative Neuroscience, Otfried-Müller-Strasse 25, 72076 Tübingen, Germany
Max-Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, 72076 Tübingen, Germany
vincent.plikat@uni-tuebingen.de

Hong Yu Wong
Department of Philosophy, University of Tübingen, Bursagasse 1, 72070 Tübingen, Germany
Centre for Integrative Neuroscience, Otfried-Müller-Strasse 25, 72076 Tübingen, Germany
hong-yu.wong @uni-tuebingen.de

17 We are grateful to Gustav Kuhn, Jordi Camí, Luis M. Martínez, Matías Graffigna, Giulia Martina, Malte Hendrickx, the CIN Philosophy of Neuroscience group at the University of Tübingen, and the reviewers for their constructive comments on earlier drafts, and also to Thomas Fraps for valuable discussion. The writing of this paper was supported by the University of Tübingen (PRG and HYW) and the Barbara-Wengeler-Stiftung (VP). PRG is currently funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project number 465409366).

References

Barto, A., Mirolli, M. and Baldassarre, G. (2013) 'Novelty or Surprise?'. *Frontiers in Psychology*, 4: 1–15. https://doi.org/10.3389/fpsyg.2013.00907.

Berlyne, D. E. (1954) 'A Theory of Human Curiosity'. *British Journal of Psychology*, 45: 180–91.

Berlyne, D. E. (1960) *Conflict, Arousal, and Curiosity*. New York: McGraw-Hill Book Company. https://doi.org/10.1037/11164-000.

Borgoni, C., Kindermann, D. and Onofri, A. eds (2021) *The Fragmented Mind*. 1st edn. Oxford: Oxford University Press.

Camí, J., Gomez-Marin, A. and Martínez, L. M. (2020) 'On the Cognitive Bases of Illusionism', *PeerJ*, 8: 1–31. https://doi.org/10.7717/peerj.9712.

Carroll, N. (1987) 'The Nature of Horror'. *The Journal of Aesthetics and Art Criticism*, 46: 51–60. https://doi.org/10.2307/431308

Carroll, N. (1990) *The Philosophy of Horror, or, Paradoxes of the Heart*. New York: Routledge.

Clark, A. (2013) 'Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science'. *Behavioral and Brain Sciences*, 36: 181–204. https://doi.org/10.1017/S0140525X12000477.

Clark, A. (2018) 'A Nice Surprise? Predictive Processing and the Active Pursuit of Novelty'. *Phenomenology and the Cognitive Sciences*, 17: 521–34. https://doi.org/10.1007/s11097-017-9525-z.

Edgington, D. (2004) 'Two Kinds of Possibility'. *Aristotelian Society Supplementary Volume*, 78: 1–22. https://doi.org/10.1111/j.0309-7013.2004.00112.x.

Egan, A. (2008) 'Seeing and Believing: Perception, Belief Formation and the Divided Mind'. *Philosophical Studies*, 140: 47–63. https://doi.org/10.1007/s11098-008-9225-1.

Fraps, T. (2014) 'Time and Magic — Manipulating Subjective Temporality'. In Arstila, V. and Lloyd, D. (eds.) *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*, pp. 263–86. Cambridge (Mass.): MIT press.

Friend, S. (2020) 'Fiction and Emotion: The Puzzle of Divergent Norms'. *The British Journal of Aesthetics*, 60: 403–18. https://doi.org/10.1093/aesthj/ayaa010.

Gendler, T. S. (2008a) 'Alief and Belief'. *Journal of Philosophy*, 105: 634–63. https://www.jstor.org/stable/20620132.

Gendler, T. S. (2008b) 'Alief in Action (and Reaction)'. *Mind &Language*, 23: 552–85. https://doi.org/10.1111/j.1468-0017.2008.00352.x.

Gilbert, D. T. (1991) 'How Mental Systems Believe'. *American Psychologist*, 46: 107–19. https://doi.org/10.1037/0003-066x.46.2.107.

Grassi, R. and Bartels, A. (2021) 'Magic, Bayes and Wows: A Bayesian Account of Magic Tricks'. *Neuroscience \&Biobehavioral Reviews*, 126: 515–27. https://doi.org/10.1016/j.neubiorev.2021.04.001.

Greenwald, A. G., McGhee, D. E. and Schwartz, J. L. K. (1998) 'Measuring Individual Differences in Implicit Cognition: The Implicit Association Test'. *Journal of Personality and Social Psychology*, 74: 1464–80.

Gregory, R. L. (1980) 'Perceptions as Hypotheses'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290: 181–97.

Helmholtz, H. (1867) *Handbuch der Physiologischen Optik*. Leipzig: Leopold Voss.

Hohwy, J. (2014) *The Predictive Mind*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001.

Jay, J. (2016) 'What Do Audiences Really Think'. *MAGIC Magazine*, 25: 46–55.

Johansson, P. et al. (2005) 'Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task'. *Science*, 310: 116–19. https://doi.org/10.1126/science.1111709.

Kuhn, G. et al. (2014) 'A Psychologically-Based Taxonomy of Misdirection'. *Frontiers in Psychology*, 5: 1–14. https://doi.org/10.3389/fpsyg.2014.01392.

Kuhn, G. (2019) *Experiencing the Impossible: The Science of Magic*. Cambridge, Mass: MIT Press.

Lamarque, P. (1981) 'How Can We Fear and Pity Fictions?'. *British Journal of Aesthetics*, 21: 291–304. https://doi.org/10.1093/bjaesthetics/21.4.291.

Lamont, P. (2017) 'A Particular Kind of Wonder: The Experience of Magic Past and Present'. *Review of General Psychology*, 21: 1–8. https://doi.org/10.1037/gpr0000095.

Leddington, J. (2016) 'The Experience of Magic'. *Journal of Aesthetics and Art Criticism*, 74: 253–64. https://doi.org/10.1111/jaac.12290.

Leddington, J. (2017) 'The Enjoyment of Negative Emotions in the Experience of Magic'. *Behavioral and Brain Sciences*, 40: e369. https://doi.org/10.1017/S0140525X17001777.

Leddington, J. (2020) 'Comic Impossibilities'. *The Journal of Aesthetics and Art Criticism*, 78: 547–58. https://doi.org/10.1111/jaac.12762.

Legare, C. H. and Gelman, S. A. (2008) 'Bewitchment, Biology, or Both: The Co-Existence of Natural and Supernatural Explanatory Frameworks Across Development'. *Cognitive Science*, 32: 607–42. https://doi.org/10.1080/03640210802066766.

Lesaffre, L. et al. (2018) 'Magic Performances – When Explained in Psychic Terms by University Students'. *Frontiers in Psychology*, 9: 1–12. https://doi.org/10.3389/fpsyg.2018.02129.

Lewis, D. (1982) 'Logic for Equivocators'. *Noûs*, 16: 431–41. https://doi.org/10.2307/2216219.

Loewenstein, G. (1994) 'The Psychology of Curiosity: A Review and Reinterpretation'. *Psychological Bulletin*, 116: 75–98. https://doi.org/10.1037/0033-2909.116.1.75.

Macknik, S. L. et al. (2008) 'Attention and Awareness in Stage Magic: Turning Tricks Into Research'. *Nature Reviews Neuroscience*, 9: 871–79. https://doi.org/10.1038/nrn2473.

Mandelbaum, E. (2013) 'Against Alief'. *Philosophical Studies*, 165: 197–211. https://doi.org/10.1007/s11098-012-9930-7.

Mandelbaum, E. (2014) 'Thinking is Believing'. *Inquiry*, 57: 55–96. https://doi.org/10.1080/0020174X.2014.858417.

Mohr, C., Koutrakis, N. and Kuhn, G. (2015) 'Priming Psychic and Conjuring Abilities of a Magic Demonstration Influences Event Interpretation and Random Number Generation Biases'. *Frontiers in Psychology*, 5: 1–8. https://doi.org/10.3389/fpsyg.2014.01542.

Pekrun, R., et al. (2017) 'Measuring Emotions During Epistemic Activities: the Epistemically-Related Emotion Scales'. *Cognition and Emotion*, 31: 1268–76. https://doi.org/10.1080/02699931.2016.1204989.

Radford, C. and Weston, M. (1975) 'How Can We Be Moved by the Fate of Anna Karenina?'. *Aristotelian Society Supplementary*, 49: 67–94. https://doi.org/10.1093/aristoteliansupp/49.1.67.

Rensink, R. A. and Kuhn, G. (2015) 'A Framework for Using Magic to Study the Mind'. *Frontiers in Psychology*, 5: 1–14. https://doi.org/10.3389/fpsyg.2015.01508.

Smith, W., Dignum, F. and Sonenberg, L. (2016) 'The Construction of Impossibility: A Logic-Based Analysis of Conjuring Tricks'. *Frontiers in Psychology*, 7: 1–17. https://doi.org/10.3389/fpsyg.2016.00748.

Stecker, R. (2011) 'Should We Still Care about the Paradox of Fiction?'. *British Journal of Aesthetics*, 51: 295–308. https://doi.org/10.1093/aesthj/ayr019.

Tullmann, K. and Buckwalter, W. (2014) 'Does the Paradox of Fiction Exist?'. *Erkenntnis*, 79: 779–96. https://doi.org/10.1007/s10670-013-9563-z.

Vogl, E. et al. (2019) 'Surprise, Curiosity, and Confusion Promote Knowledge Exploration: Evidence for Robust Effects of Epistemic Emotions'. *Frontiers in Psychology*, 10: 24–74. https://doi.org/10.3389/fpsyg.2019.02474.

Walton, K. L. (1978) 'Fearing Fictions'. *The Journal of Philosophy*, 75: 5–27. https://doi.org/10.2307/2025831.

Wiseman, R. and Greening, E. (2005) 'It's Still Bending: Verbal Suggestion and Alleged Psychokinetic Ability'. *British Journal of Psychology*, 96: 115–27. https://doi.org/10.1348/000712604X15428.

Wood, M. J., Douglas, K. M. and Sutton, R. M. (2012) 'Dead and Alive: Beliefs in Contradictory Conspiracy Theories'. *Social Psychological and Personality Science*, 3: 767–73. https://doi.org/10.1177/1948550611434786.

Imaging Neuroscience
an open access journal

Research Article

Hierarchical surprise signals in naturalistic violation of expectations

Vincent Plikat[a,b,c,*], Pablo R. Grassi[a,b,c,*], Julius Frack[d], Andreas Bartels[a,b,c]

[a]Department of Psychology, University of Tübingen, Tübingen, Germany
[b]Centre for Integrative Neuroscience, Tübingen, Germany
[c]Max-Planck Institute for Biological Cybernetics, Tübingen, Germany
[d]Rocket Magic, Tübingen, Germany

*Equal contribution

Corresponding Authors: Vincent Plikat (vincent.plikat@uni-tuebingen.de), Pablo R. Grassi (pablo.grassi@cin.uni-tuebingen.de), Andreas Bartels (andreas.bartels@uni-tuebingen.de)

ABSTRACT

Surprise responses signal both high-level cognitive alerts that information is missing, and increasingly specific back-propagating error signals that allow updates in processing nodes. Studying surprise is, hence, central for cognitive neuroscience to understand internal world representations and learning. Yet, only few prior studies used naturalistic stimuli targeting our high-level understanding of the world. Here, we use magic tricks in an fMRI experiment to investigate neural responses to violations of core assumptions held by humans about the world. We showed participants naturalistic videos of three types of magic tricks, involving objects appearing, changing color, or disappearing, along with control videos without any violation of expectation. Importantly, the same videos were presented with and without prior knowledge about the tricks' explanation. Results revealed generic responses in frontal and parietal areas, together with responses specific to each of the three trick types in posterior sensory areas. A subset of these regions, the midline areas of the default mode network (DMN), showed surprise activity that depended on prior knowledge. Equally, sensory regions showed sensitivity to prior knowledge, reflected in differing decoding accuracies. These results suggest a hierarchy of surprise signals involving generic processing of violation of expectations in frontal and parietal areas with concurrent surprise signals in sensory regions that are specific to the processed features.

Keywords: surprise, magic, predictive coding, violation of expectation, intuitive physics, fMRI

1. INTRODUCTION

Prior experience and "intuitive" knowledge about the physical world guide our perception and allow for a meaningful interaction with the environment. They set up constraints on our expectations based on what we believe to be possible in the world. For example, prior world-knowledge informs us that objects do not vanish of existence if occluded (object permanency), that objects tend to keep their features (feature constancy), that objects cannot pass through other objects (solidity), that objects do not appear out of the blue, and so forth. Informed expectations based on these intuitive physical priors allow us to quickly make sense of incoming sensory information. Such expectations have been shown to strongly modulate perception (de Lange et al., 2018). For example, a "light-from-above" prior constrains depth perception from shading (Adams et al., 2004), and knowledge that objects cannot occupy the same place at the same time explains our inability to perceive two objects simultaneously in bistable perception (Hohwy et al., 2008).

Accordingly, perception can be understood as an inferential process in which top–down informed expectations

Received: 23 July 2024 Revision: 20 December 2024 Accepted: 25 December 2024 Available Online: 8 January 2025

are matched with incoming sensory information (Friston, 2005, 2010; Lee & Mumford, 2003; Rao & Ballard, 1999). In this "predictive processing" framework, deviations between expectations and incoming data (i.e., prediction errors) are used to update an internal model of the world at different levels of complexity and abstraction. Higher-level priors (like object permanency) represent more abstract aspects of the world and constrain lower-level inferences that represent more immediate features of the world (like a particular object) (Clark, 2013; Hohwy, 2014). In this context, intuitive physical knowledge can be thought of as an internal model representing different aspects of the causal structure of the physical world, not unsimilar to a "physics engine" in a virtual environment (Battaglia et al., 2013).

This prior knowledge of physical principles can be studied using events that seemingly violate them. For example, evidence from developmental psychology using so-called violation of expectation (VOE) paradigms suggest that infants acquire important aspects about the workings of the physical world during the first year of life, such as object permanency and solidity (Hespos et al., 2009; S. Wang, 2004; Wynn, 1992). However, it is largely unclear how surprise-responses relating to intuitive physical principles are represented in the human brain.

Most previous neuroimaging studies investigated responses to lower-level VOE, for example using paradigms involving novel or infrequent (and thus unexpected) stimuli (e.g., Linden, 1999; Stevens et al., 2000; Todorovic et al., 2011; Wessel et al., 2012), omission of expected stimuli (e.g., Todorovic et al., 2011; Wacongne et al., 2011), or changes in stimulus sequences (Downar et al., 2000, 2001; Grundei et al., 2023). These studies have consistently revealed the involvement of lower-level stimulus-specific prediction errors in modality- and feature-specific areas (i.e., visual areas are modulated by visual surprises, auditory cortex by auditory surprises, and so forth, see e.g., Egner et al., 2010; Kok et al., 2012; Todorovic et al., 2011), as well as of higher-level frontal and parietal areas, including the fronto-insular and anterior cingulate cortex (ACC) (as reviewed in Fouragnan et al., 2018; Kim, 2014). Importantly, these frontal and parietal areas are consistently involved in signaling lower-level sensory expectation violations irrespective of sensory modality (Downar et al., 2000; Grundei et al., 2023).

In contrast, only few studies investigated VOE of higher-level physical principles (such as object permanency). These studies used more complex stimuli such as computer-generated animations (Bardi et al., 2017; Liu et al., 2024) or naturalistic videos showing magic tricks (Danek et al., 2015; Parris et al., 2009) and consistently revealed higher-level surprise signals in frontal and parietal areas. This is largely in line with the suggested role of several frontal and parietal areas in the representation of physical concepts (Fischer et al., 2016; Schwettmann et al., 2019). However, and in contrast to studies investigating lower-level VOEs, modality and feature-specific sensory areas have either not been in the focus of these higher-level VOE studies or have hitherto not observed VOE-related activity in them (cf., Liu et al., 2024).

Here, we designed a novel VOE paradigm to uncover the effect of VOE of physical concepts (i.e., "world-model" VOEs) on both: previously identified higher-level frontal and parietal areas, as well as lower-level sensory areas. Our paradigm contained a battery of standardized magic videos that we presented to human participants while measuring fMRI responses to investigate: 1) which regions are generally involved when viewing natural videos that violate physical principles; 2) whether specific types of violations, like appearance of objects, or changes of color, modulate sensory areas known to process the feature concerned; and 3) whether knowledge of the explanation of a given magic trick modulates the observed VOE activity.

To this aim, we created and validated videos for a naturalistic VOE paradigm showing either dedicated magic tricks (to create the illusion of seemingly impossible events to actually occur, cf., Grassi & Bartels, 2021) or matched control actions that involved no violation of physical principles. The VOE videos were designed to evoke surprise responses related to unexpected object appearance, disappearance of objects, and feature change (color-changing objects). They were performed by a professional magician (Julius Frack), and each trick-type was performed using three common objects (balls, playing cards, and pencils). Each trick was presented before and after revealing the method of the tricks. This allowed us to compare responses with and without VOE using identical videos.

Univariate analyses revealed a hierarchy of surprise signals: frontal and parietal areas, including the dorsal ACC and areas of the default mode network (DMN), were involved when perceiving events violating physical principles regardless of the type of trick used, and their activity was positively correlated with subjective surprise. In contrast, posterior sensory areas were modulated specifically by the type of expectation-violation, such as color-processing medial fusiform cortex by color change, and object-selective LOC by the appearance of objects. Controls indicate that their modulation is due to the feature-specific surprise and not due to the feature-change. Multivariate analyses extended the results: information about the specific types of expectation-violations was exclusively encoded in posterior regions, and significantly decodable down to the earliest levels of cortical visual processing (V1–V3). Additionally, decoding

accuracy significantly decreased with prior knowledge, revealing a reduction in surprise signals. Together, our results demonstrate a generic response in frontal and parietal areas to violation of physical principles, along with concurrent representations of specific expected information in early sensory areas.

2. METHODS

2.1. Participants

We performed fMRI on 27 subjects. Three subjects were excluded from data analysis due to excessive movement and/or sleepiness during the scanning sessions. Data from a total of 24 subjects were analyzed (16 female; 8 male; mean age 24.4 ± 4.3 SD years). All subjects had normal or corrected to normal vision, no history of neurological impairments nor contraindication for fMRI. Participants had no expertise as magicians and were naive to the magic tricks used. Participants provided written informed consent prior to the experiment. The study was approved by the ethics committee of the University Clinic Tübingen and was conducted in accord with the Declaration of Helsinki.

2.2. VOE stimuli

2.2.1. Video recordings

A set of 63 videos were created for the VOE study. In accord with previous work (Parris et al., 2009), we created videos for three different conditions: the videos showed either magic tricks (magic condition, 18 videos), similar actions without a magic event (control condition, 18 videos), or unusual actions with the objects used in the magic tricks (unusual condition, 9 videos). We included the unusual condition to investigate neural correlates of surprise in a similar setting while not violating any physical concept (cf., Parris et al., 2009). We further created explanation videos showing how each of the magic tricks was achieved (18 videos).

All videos were performed by professional illusionist Julius Frack in a standardized setting consisting of a black background and a black table (see Fig. 1A for an example). To investigate the effect of specific VOE, we presented magic tricks showing three different violations of physical principles: appearances (A) (i.e., a red object appears), color changes (C) (i.e., a red object changes color to blue), and vanishes (V) (i.e., a red object disappears) (see Fig. 1C). To generalize across objects, we used three easily distinguishable objects: balls, playing cards, and thick pencils. Each object was used equally often in each trick type. Finally, we created two versions of each type of VOE for each object (e.g., two different videos showing a red ball changing to a blue ball using different methods).

2.2.2. Full set of stimuli

Each of these magic tricks had a matched control video that showed the same sequences of actions as the trick, but without VOE (e.g., following the same actions a ball would not change its color). Thus, we had a total of 18 magic videos (3 types of VOEs × 3 objects × 2 methods), with 18 matching control videos and nine unusual videos (3 unusual actions per used object). As the solutions to the magic tricks were revealed in distinct sets throughout the experiment, we used a variety of different methods for each type of VOE. This ensured that participants were only able to infer trick solutions they were intended to understand.

Prior to the fMRI experiment, we performed two psychophysics experiments with a total of 18 subjects (nine subjects in each experiment) to ensure the suitability of our stimuli and to select the magic tricks to be used in the fMRI experiment (see results in Supplementary Section *Behavioral evaluation of stimuli*).

2.2.3. Luminance and durations

Tricks were recorded in a standardized setting under the same lighting conditions. To balance out remaining inequalities, custom MATLAB (MathWorks, Natick, MA) scripts were used to standardize the videos (resolution of 1920 x 1080 with 25 frames per second), which were filmed on different days and had different lengths. We manually applied white-balance using 5 to 10 selected white-pixels for all videos of a same day, matched the luminance and contrast of the videos based their first frame, and shortened the videos to be no longer than 14 s. The final duration of the videos was 12.8 s ± 1.08 s (mean ± SD).

2.2.4. Stimulus presentation

Stimuli were presented using MATLAB 2019b using Psychtoolbox3 (version 3.0.16 http://psychtoolbox.org/) on a Linux computer and back-projected to a translucent screen mounted at the rear of the scanner bore using a ProPixx projector (VPixx Technologies, Saint-Bruno-de-Montarville, Canada) at a frame rate of 144 Hz. Participants viewed the screen (26.1 x 14.7 visual degrees) via a mirror mounted on the 64-channel head coil (Siemens, Erlangen, Germany) at a distance of 105 cm. To center the stimulus presentation, we cropped 160 pixels from the left and right side of the videos that only showed a black background. Accordingly, the shown part of the videos covered 21.8 x 14.7 visual degrees (1600 x 1080 pixels).

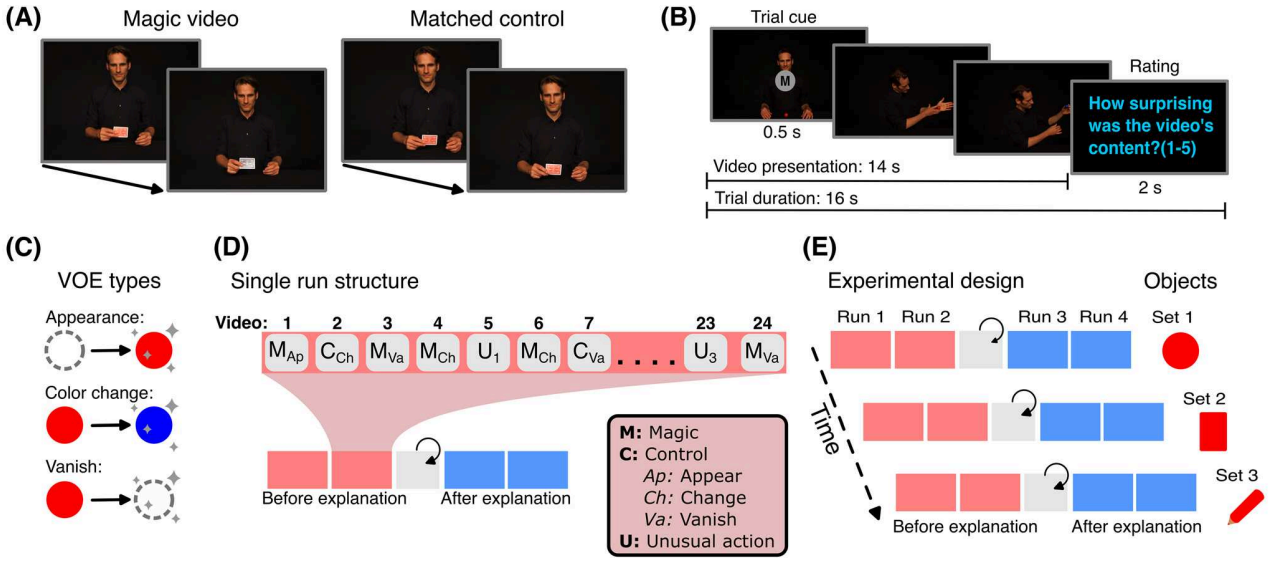


Fig. 1. (A), example of a color-changing card magic trick (left) and its corresponding matched control video (right). (B), shown is the timeline of a single trial. Every video started with a central cue lasting 500 ms, indicating whether the video will show a magic trick (M) or not (X). Each video presentation lasted 14 s; if a video happened to be shorter than 14 s, the last frame of the video was shown until the total duration was 14 s as a static image. After the video, subjects had 2 s to answer how surprising the video's content was. (C), shown are the three different types of VOE used (i.e., "magic events"). We created videos showing magic tricks using three objects (balls, playing cards, and pencils) that showed either an unexpected object appear (Ap), change color (from red to blue) (Ch) or vanish (Va). For each object-VOE combination, we created two tricks (i.e., we had two color-changing card tricks). (D), schematic example of a single experimental set. One set consisted of four runs—two before and two after explanation runs (i.e., pre and post revelation). Between pre- and post-revelation runs, we showed participants videos explaining the magic tricks performed during the corresponding set. Each run showed 24 videos with a pseudo-randomized order, ensuring that the same video was not presented in two consecutive trials (12 magic videos [M], 6 matched controlled videos [C], and 6 unusual actions [U]). (E), complete experimental design showing all three sets. The experiment was divided into three sets, one for each of the objects. Each set was divided into four experimental runs, where each run showed all the videos of an object but in a randomized order. After the second fMRI run, the methods behind the tricks shown in the set were revealed, dividing the runs into *before* and *after* the explanation of the tricks (pre- and post-revelation, respectively).

2.3. Experimental design and procedure

2.3.1. fMRI runs

Each run consisted of a total of 24 video trials: two repetitions of the six unique magic videos of one object (three types of VOE, each type recorded in two versions), resulting in 12 magic trick presentations; the six matching control videos (shown once); and two repetitions of three additional unusual videos, resulting in six unusual video presentations (see Fig. 1D). The resulting 24 video trials were presented in a pseudo-random order that avoided repetitions of identical videos (different randomizations across runs).

2.3.2. Paradigm

The fMRI experiment consisted of three sets with four experimental fMRI runs each. Each set presented videos of only one object (i.e., balls, playing cards or pencils). The presentation order of the object sets was counterbalanced across subjects. After the first two fMRI runs in a set (pre-revelation runs), the method behind each magic trick in the set was revealed by showing the participants each of the magic tricks again together with the matching revelation video. Subjects could watch the videos as often as they wished and were asked to confirm per button press that they understood how each of the tricks was achieved. Thereafter, two more runs showing the same videos were performed (post-revelation runs). Together, each set consisted of two fMRI runs before and two fMRI runs after the explanation of the tricks. A visualization of the experimental design is shown in Figure 1E. We hypothesized that providing the explanation of the tricks would decrease VOE responses because participants would adjust their expectations (e.g., knowing where the seemingly disappearing objects were concealed). Consequently, our experimental design allowed us to compare responses with and without VOE using identical stimuli.

2.3.3. *Individual trials*

Each video trial was presented for 14 s. In case a video was shorter (e.g., 13 s), the last frame of the video was shown for the remaining time (e.g., for 1 s). Participants were informed about this. Moreover, the first 500 ms of each video showed a central cue (1.3 visual degrees), indicating whether the video was going to be a magic (M) or a non-magic (X) video (i.e., control or unusual actions). We included the cue (M or X) at the beginning of the trial to prevent participants from being surprised by contextual or serial effects. Moreover, to reduce predictability of video content we flipped each video on every second presentation of the same video horizontally. After each video presentation, participants were asked to rate from 1 to 5 how surprising the content of the video was (1 = not surprising, 5 = very surprising). They had 2 s to respond, after which the next trial started (total trial duration = 16 s, see Fig. 1B). Behavioral surprise ratings were given by the subjects using a button-box with five keys.

2.3.4. *Pre-scan instructions*

Before scanning, we instructed subjects about the task and design of the experiment (i.e., the set design, meaning of trial cues, explanation videos, etc.). Participants were informed that videos would show either magic tricks, control actions, or unusual actions. Moreover, participants were shown a magic and matching control example video (not used in the actual experiment) to give them an impression about the kind and duration of the videos they were about to see. Further, to prevent participants from watching the videos in a "problem-solving" attitude, we instructed them to passively watch and enjoy the videos without trying to get behind the method of the tricks, as these would be explained during the experiment.

2.4. fMRI data acquisition

fMRI data were acquired in a 3 Tesla Siemens Prisma scanner with a 64-channel head coil (Siemens, Erlangen, Germany). Functional images were acquired using an accelerated T2*-weighted gradient-echo echoplanar imaging (EPI) sequence (multiband factor = 2, repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle (FA) = 75°, 62 slices with an isotropic voxel size of 2 × 2 × 2 mm) using GRAPPA (GRAPPA = 2). Each run consisted of 198 images (total duration = 6 min and 36 s). Moreover, a high-resolution T1-weighted structural scan with whole-brain coverage was performed for each participant (TR = 2000 ms, TE = 3.06 ms, inversion time (TI) = 1100 ms, FA = 9°, 192 slices and an isotropic voxel size of 1 × 1 × 1 mm). The structural scan was measured during the explanation of the tricks in the first set.

2.5. fMRI data preprocessing

Functional MRI images were preprocessed first by removing thermal noise from the magnitude EPI images using NORDIC, a PCA-based algorithm (Vizioli et al., 2021) in MATLAB. Then, we discarded the first five volumes of each run to allow for T1 equilibration effects. Using SPM12, we further performed motion correction (realigned to the first image), slice-time correction (using middle slices as reference), and co-registration to the structural scan. Finally, functional MRI data for whole-brain analyses were normalized to the Montreal Neurological Institute template brain (MNI152) and spatially smoothed with a Gaussian kernel of full width at half-maximum of 6 mm for univariate whole-brain analyses. Region-of-interest (ROI) analyses were performed on unsmoothed data in native space. Moreover, for the generation of subject-specific ROIs, we generated inflated individual brain surfaces using Freesurfer 7.1.1 (Dale et al., 1999) using a dedicated docker container (https://hub.docker.com/r/freesurfer/freesurfer).

2.6. Univariate whole-brain data analysis

To investigate differences in neural activity during high-level VOE in the human brain, we performed whole-brain and ROI univariate analyses. We had three specific aims: 1) to investigate neural responses to violations of physical principles in naturalistic stimuli, comparable to previous reports using magic tricks (cf., Danek et al., 2015; Parris et al., 2009), 2) to investigate differential responses between different VOE (i.e., magic trick types), and 3) to investigate the role of prior knowledge in the perception of events violating physical principles by comparing responses to the very same videos before and after revelation.

For these analyses, we created two event-related general linear models (GLM) using the canonical hemodynamic response function and high-pass filtered data with a cut-off at 128s in SPM12. In the first GLM (aim 1), we modeled responses using only three regressors of interest in each run (magic, control, and unusual stimuli conditions). In a second, extended GLM, we investigated possible differential responses to different forms of violation of expectations (aim 2), using seven regressors of interest that modeled BOLD responses of the three magic types, three matching controls, and unusual actions for each run. As in previous reports, we modeled individual trials within a regressor using discrete event times based on the moment of surprise (cf., Danek et al., 2015; Parris

et al., 2009). The definition of the event times was done by averaging the independent selection of a suitable frame done by two authors (VP and PRG). For magic videos, timing was decided based on when the "magic" (i.e., VOE) happened (mean VOE onset was 7.62 s after video start; range: 4.36–9.64 s). For matching control videos, the corresponding time point was selected, that is, the moment in which one would expect to see a magic event in the magic videos (mean = 7.74 s; range: 4.52–10.92 s). For the videos showing unusual actions, timing was selected based on the onset of the unusual actions (mean = 6.06 s; range: 4.72–7.84 s). Moreover, we included participants' response times, six movement parameters, and a column of ones as nuisance regressors in both GLMs. We additionally computed the same analyses modeling the whole video durations (14 s) and report these in the Supplementary Tables S11–S13 and Fig. S7). We used these models to address our three aims as follows.

Aim 1: to identify brain areas preferentially involved in signaling high-level VOE we compared responses to magic and matching control videos in all six pre-revelation runs (i.e., two runs from each object set) using the pooled regressors (from the first GLM) ($Magic_{pre} > Control_{pre}$). Similarly, we compared responses between the magic and the unusual videos ($Magic_{pre} > Unusual_{pre}$). First-level contrasts were used for second-level random-effect analyses using parametric tests as implemented in SPM12 and with non-parametric permutations tests using the non-parametric mapping toolbox SnPM13 (Nichols & Holmes, 2002). Moreover, we tested in significant clusters whether BOLD responses were positively correlated with subjective surprise ratings. We correlated parameter estimates of each video presentation taken from a separate GLM (see section 2.9 Multivariate pattern analysis MVPA), with the corresponding subjective surprise ratings after accounting for the video condition (partial correlation using the video condition as covariate), averaged the Fisher z-transformed correlation coefficients within a cluster, and tested for significance at group-level using one-sided T-tests against 0.

Aims 1 and 2: to investigate generic and violation-specific responses to high-level VOE before the explanation of the tricks, we used a second-level, 2 (magic, control) x 3 (appear, change, vanish) repeated-measures ANOVA to conduct four conjunction analyses (Friston et al., 2005). To find common responses to all VOE, we contrasted each magic trick type with its corresponding control condition before the explanation of the tricks (i.e., $Magic_{A\_pre} > Control_{A\_pre}$) and used those contrasts to perform a conjunction (across all three VOE types). For each VOE type separately, we further performed a conjunction analysis on the responses of one VOE type against the others. For example, responses specific to an appearing object were investigated by means of the conjunction of the contrasts $Magic_{A\_pre} > Magic_{C\_pre} \cap Magic_{A\_pre} > Magic_{V\_pre}$.

Aim 3: to identify brain areas generally involved in the perception of magic and affected by prior knowledge, we compared responses to magic after providing the explanation of the tricks $Magic_{post} > Control_{post}$ and additionally compared the interaction between the revelation and magic ($Magic_{pre} > Control_{pre}$) > ($Magic_{post} > Control_{post}$). Finally, we compared the same contrasts for each magic type separately using the second, extended GLM, for example, ($Magic_{A\_pre} > Control_{A\_pre}$) > ($Magic_{A\_post} > Control_{A\_post}$). Please note that these contrasts are controlling for potential time confounds.

For all whole-brain analyses, a cluster-forming threshold of $p_{unc} = 0.001$ and cluster size threshold of k = 30 voxels was applied. In tables, we report cluster-wise parametric and non-parametric FWE-corrected results as suggested by Nichols and Holmes (2002). Clusters that survive FWE-correction are highlighted in bold. All non-parametric permutation tests were performed using 5000 permutations. Brain areas in whole-brain analyses were identified using the atlasreader toolbox (Notter et al., 2019), applying the Automated Anatomical Labeling atlas 3 (AAL3) (Rolls et al., 2020).

2.7. Univariate whole-brain control analyses

We performed two control analyses for the whole-brain univariate results. First, it is possible that the conjunction analyses examining differences between the magic trick types are confounded by differences in the visual content of the videos at specific time-points. This is because we are not only comparing videos showing "appearances", "color changes", and "vanishes", but also videos showing "red objects", "blue objects", and "no objects" at a specific time point, respectively (see Fig. 1C). To control for this possible stimulus-driven confound, we compared responses between magic and control videos showing similar visual contents at specific time points. Second, our knowledge-dependent analysis could be confounded by general condition-independent time effects and, as participants were repeatedly presented the video stimuli within a set, by the repetition suppression effect (Grill-Spector et al., 2006; Krekelberg et al., 2006). To address these potential confounders, we performed a mixed-effects model, modeling the change in prior-knowledge, experiment-wise and set-wise temporal decays, to explain the average beta estimates of the magic tricks of each run for each significant cluster of the prior-knowledge dependent contrast ($Magic_{pre} > Control_{pre}$) > ($Magic_{post} > Control_{post}$).

Detailed methods of our control analyses can be found in Supplementary Sections *VOE-specific control analysis* and *Prior knowledge dependent control analysis* respectively.

2.8. ROI definition

For the ROI analyses, we defined 26 hypothesis-driven ROIs, separated into two groups.

First, we defined a set of 16 surprise-related ROIs based on significant responses to magic videos from previous experiments (Danek et al., 2015; Parris et al., 2009). We defined 14 frontal and parietal ROIs by combining labels from a multi-modal parcellation of the human cortex (Glasser et al., 2016) and two subcortical ROIs using the Freesurfer automatic parcellation (Fischl et al., 2002).

Second, we used a probabilistic map of visual fields (Wang et al., 2015) to define 10 visual ROIs: primary visual cortex (V1), secondary visual cortex (V2), V3, V3A, V3B, human V4 (hV4), lateral occipital and ventral complex (LO and VO, respectively), intraparietal sulcus (IPS), and frontal eye-fields (FEF). All ROIs were defined in native space. For a detailed list of ROIs, please see Supplementary Section *Region of interest definition*.

We included these visual ROIs to use in the decoding of the VOE type (i.e., appear, change and vanishing, see below) and to test for differences evoked by the different VOE types and the effect of prior knowledge. For example, areas of the ventral visual cortex are known to be responsive to color (Bartels & Zeki, 2000), while the lateral occipital complex is responsive to objects (Grill-Spector et al., 2001). As predictive coding approaches predict feature-specific prediction errors in functionally specialized regions, we expect unexpected color changes to affect color-responsive ROIs (e.g., hV4, VO and PH), and unexpected object appearances to affect object-responsive areas LO and VO, in line with recent imaging evidence (Jiang et al., 2016; Richter et al., 2018; Stefanics et al., 2019). Early visual areas (V1, V2, V3) were included to investigate possible top-down effects of prior knowledge in lower-level areas, when comparing the exact same videos before and after revelations, while the parietal (IPS) and prefrontal ROIs (FEF) were included due to their involvement in top-down voluntary attention (Corbetta & Shulman, 2002). See Figure 5A for a depiction of all 26 ROIs in an exemplary subject.

2.9. Multivariate pattern analysis (MVPA)

Apart from the univariate analyses testing for net signal differences, we further wanted to investigate which areas of the brain carry pattern information about the different types of VOE (unexpected appearance, feature change, and omission). To do so, we performed a series of multivariate pattern analyses (MVPA) on the 26 hypothesis-driven ROIs and a control ROI.

For the decoding analyses we computed a GLM in which every trial (i.e., video presentation) was modeled as a separate regressor to increase the number of data points for training and testing. All analyses were performed using a shrinkage linear discriminant analysis (LDA) on the de-meaned beta estimates of the individual trials (by the mean over all estimates, i.e., all trials, within each voxel) using the Python (version 3.8.13) package scikit-learn's class LinearDiscriminantAnalysis (Pedregosa, 2011). To examine if any of the ROIs contained information about the different types of VOE (appear, color change, and vanish), we trained and tested our decoder to predict the VOE types following a three-fold cross-validation scheme to ensure generalization across objects. We trained on the data of two objects (i.e., estimates from two sets, 48 trials) and tested on the third object (i.e., estimates from the third set, 24 trials). Significance testing of the decoding accuracies was done using a permutation analysis (1000 permutations) implementing the max statistic correction to correct for multiple comparisons (Nichols & Holmes, 2002). A control ROI (third ventricle) was included in the analysis, which should carry no information and thus reflect chance level.

Decoding analyses were performed separately for data before and after explanations of the magic tricks. Permutation-based corrected significance thresholds were 36.98% and 36.92% before revelation and after revelation, respectively. As both analyses were conducted using estimates based on the very same videos, we hypothesized that any significant difference in decoding accuracies between data before and after revelation would be indicative of decodable prior-knowledge dependent surprise signals. We tested for differences in decoding accuracies using paired t-tests between decoding using pre-revelation data and decoding using post-revelation data, only in those ROIs that showed significant decoding (corrected) using pre-revelation data. Using the same ROIs, we performed the same decoding analyses but using beta estimates from GLMs that modeled the moments 5 s before and after the onset VOE in steps of 2 s (i.e., -5, -3, -1, 1, 3, 5 s relative to VOE onset). We expect decoding accuracies to peak around VOE onset, because differences in signals should be strongest, and decrease the further we move away from VOE onset.

Additionally, as an exploratory approach we performed a similar whole-brain searchlight analysis on unsmoothed data in MNI-space using a sphere to decode the magic types (4 mm radius), separately for data before and after explanation of the tricks using the SearchLight class implemented in nilearn (Abraham et al., 2014). The searchlight whole-brain accuracy maps were spatially

smoothed with a 4 mm Gaussian kernel. A permutation-bootstrap hybrid method (in which each randomly generated accuracy map was also smoothed with a 4 mm Gaussian kernel) was used for significance testing and correction for multiple comparisons (Stelzer et al., 2013) using custom-made Python code.

2.10. Behavioral data

To test for differences in surprise ratings between videos before and after the explanation of the tricks, we performed a 2 (before/after explanation) x 3 (magic, unusual and control videos) repeated-measures ANOVA. We expected to see higher surprise ratings for magic videos compared to control videos and higher ratings for magic videos before compared to after the revelation of the methods. We also wanted to test if videos of magic and of unusual actions led to similarly high surprise ratings. Moreover, we tested for differences in surprise ratings of the magic videos for different objects and VOE types in 2 x 3 repeated-measures ANOVAs (rmANOVA) with the factors revelation (before/after) and object (ball/card/pencil) or VOE type (appear/change/vanish), respectively.

2.11. Eye tracking

Gaze positions were measured using an MR-compatible Eyelink 1000 (SR-Research, Ottawa, Canada) positioned at the rear end of the scanner bore. Eye tracking data were analyzed to test whether gaze positions, number of blinks, and number of saccades differ significantly between conditions. Details about data acquisition, preprocessing, and analyses can be found in Supplementary Section *Eye tracking acquisition and analysis*.

2.12. Inference statistics

Effect sizes for repeated-measures ANOVAs and paired tests are presented as partial eta squared (η^2) and Cohen's *d*, respectively. Sphericity of rmANOVAs was tested using the Mauchly test. If sphericity was violated, degrees of freedom were adjusted using the Greenhouse-Geisser correction, the corresponding ε-correction factor is provided. Normality of data was assessed using the Shapiro-Wilk test (for paired tests and post-hoc tests). In case that data were normally distributed, we performed paired t-tests, otherwise we performed non-parametric Wilcoxon signed-rank tests instead. Correction for multiple comparison was performed using a step-down Holm-Bonferroni correction. Please note that in the post-hoc tests we corrected for each hypothesis separately (i.e., p-values for post-hoc tests of one factor are corrected independent of another factor or an interaction). In general, corrected p-values (p_{corr}) are reported in text, uncorrected p-values (p_{unc}) are reported in the corresponding tables. The threshold for statistical significance was set to 0.05 for all tests.

3. RESULTS

3.1. Behavioral surprise ratings

Behavioral data were first tested for differences in surprise ratings for video condition (magic, control, and unusual videos) and revelation condition (before and after revelation) using a 3 × 2 rmANOVA (see Fig. 2A). In sum, this analysis revealed that magic tricks were perceived as more surprising than the control videos, and that surprise ratings dropped after explanation of the tricks. In detail: the analysis revealed significant main effects for both factors (video condition: $F(2,46) = 57.494$, $p_{unc} < 0.001$, $\eta^2 = 0.478$, $\varepsilon = 0.932$, revelation: $F(1,23) = 103.242$, $p_{unc} < 0.001$, $\eta^2 = 0.211$, $\varepsilon = 1$) and interaction: $F(2,46) = 43.081$, $p_{unc} < 0.001$, $\eta^2 = 0.134$, $\varepsilon = 0.641$). As expected, post-hoc Wilcoxon signed-rank tests revealed that magic videos were more surprising than control and unusual videos, before ($W = 1$, $p_{corr} < 0.001$, Cohen's $d = 4.07$ and $W = 8$, $p_{corr} < 0.001$, Cohen's $d = 2.06$, respectively) and after explanation of the tricks ($W = 5$, $p_{corr} < 0.001$, Cohen's $d = 1.55$ and $W = 26$, $p_{corr} < 0.001$, Cohen's $d = 0.812$, respectively), and that all video conditions were more surprising before compared to after the revelation (all three video conditions: $p_{corr} <= 0.001$). Unusual videos were more surprising than control videos pooled across runs ($W = 42$, $p_{corr} = 0.006$, Cohen's $d = 0.59$), but significantly so only in the first two runs ($W = 26$, $p_{corr} = 0.002$, Cohen's $d = 0.71$) (see Supplementary Table S1 for a detailed report of all post-hoc tests).

Average surprise ratings of magic videos were consistently high before the explanation of the tricks (all group means > 3) and decreased afterward (all group means < 2.5) (see Fig. 2A, B and C). We further performed two rmANOVAs using only ratings from magic videos, to test for possible differences in VOE types and the objects used. Both rmANOVAs included the revelation condition as a factor and showed a significant decrease in surprise rating after the revelation of the tricks (as expected from the previous analysis). We additionally found a significant main effect for the magic type ($F(2,46) = 13.8$, $p_{unc} < 0.001$, $\eta^2 = 0.032$, $\varepsilon = 0.848$) and an interaction of magic type and revelation condition ($F(2,46) = 11.02$, $p_{unc} < 0.001$, $\eta^2 = 0.018$, $\varepsilon = 0.795$) (see Fig. 2B). Post-hoc tests showed that appearances were rated less surprising than color changes and disappearances pre-revelation ($W = 12$, $p_{corr} < 0.001$, Cohen's $d = -0.835$ and $W = 35.5$, $p_{corr} = 0.002$, Cohen's

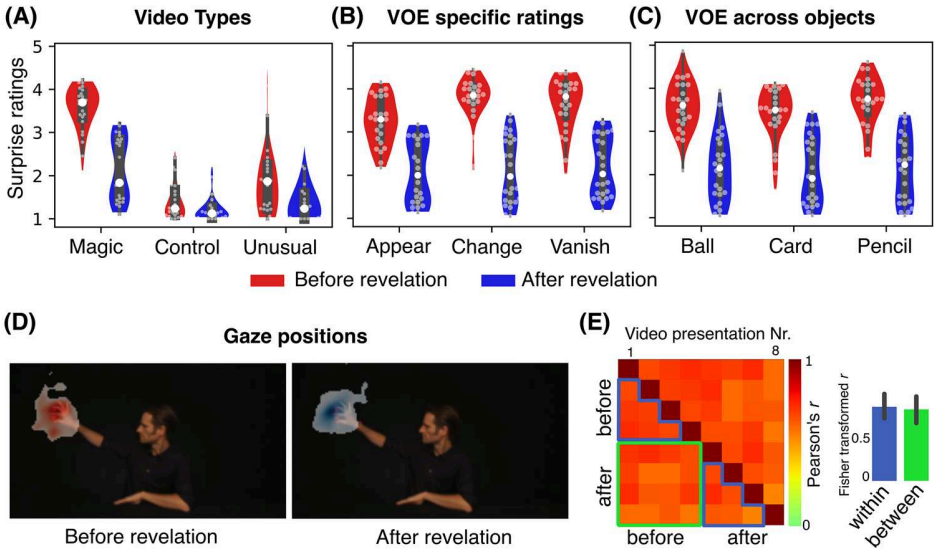


Fig. 2. Shown are behavioral surprise ratings separated for the different video types (A), for magic videos across VOE types (B), and across objects (C). Surprise ratings before the revelations of the magic tricks (red) were consistently higher than after the revelations (blue). (D), Exemplary group gaze positions for the same magic video in the moment a ball appears, before (left) and after (right) the revelation of the trick. (E), Correlation matrix for gaze positions around the moment of magic (-1 s and +2 s) from an exemplary video. To test for differences between gaze paths before and after the revelation of the tricks, we compared (paired t-tests) the average Fisher z-transformed correlations between gaze paths of all combinations of presentations (from presentation Nr. 1 to Nr. 8) *between* pre- and post-presentations (marked green) and those *within* pre- and *within* post-presentations (marked blue), as shown in the bar plot.

d = -0.631, respectively). No difference between color change and vanish magic tricks was observed (see Supplementary Table S2). No main effect for objects nor an interaction of object and revelation condition were found (all *F-values* < 2, all p_{unc} > 0.15 and η^2 < 0.012) (see Fig. 2C).

3.2. Eye-tracking results

Eye-tracking data were tested for systematic differences in gaze traces, saccades, and blinks during viewing of the videos (see also Supplementary Section *Eye-tracking results* and Fig. S2). Gaze traces between pre- and post-revelation runs were similar for all videos (see an example visualization in Fig. 2D) and no significant difference of correlations of gaze traces was observed (all p_{unc} > 0.2) (see Fig. 2E for an example). Moreover, the number of saccades and blinks around the VOE times were similar between experimental conditions and only revealed small differences we deem unlikely to have affected the imaging results.

3.3. Univariate whole-brain analysis

3.3.1. Whole-brain surprise responses

To investigate neural correlates of high-level VOE when viewing seemingly impossible events, we first compared whole-brain responses to the magic and matched control videos before the revelation of the magic tricks ($Magic_{pre}$ > $Control_{pre}$). This contrast revealed several clusters of activity in frontal and parietal cortices, largely in line with previous studies (Danek et al., 2015; Parris et al., 2009) (see Fig. 3A and Table 1). In particular, large clusters of activity were observed in the medial part of Brodmann area 8 (preSMA), the dorsal and ventral anterior cingulate cortex (dACC and vACC), and the posterior parietal cortex (PPC, especially the superior parietal lobe and the precuneus). Moreover, lowering the cluster size threshold to k = 10 revealed subcortical areas such as left caudate nucleus (k = 19), in line with previous studies (Danek et al., 2015; Parris et al., 2009) and bilateral Thalamus (left k = 22, right k = 13). No lower-level sensory area was differentially modulated in view of the unexpected events. Partial correlations between beta estimates of each magic or control video presentation and corresponding surprise ratings (using video condition as covariate) showed significant positive correlations in the dACC/preSMA, parietal cortex (including PPC and bilateral anterior IPS) and left and right SFG (all T-values > 1.78, all p-values < 0.044).

When comparing whole-brain responses to the magic and the unusual videos before the revelation of the magic tricks ($Magic_{pre}$ > $Unusual_{pre}$), similar frontal (dACC) and parietal areas (PPC) were active (see Supplementary

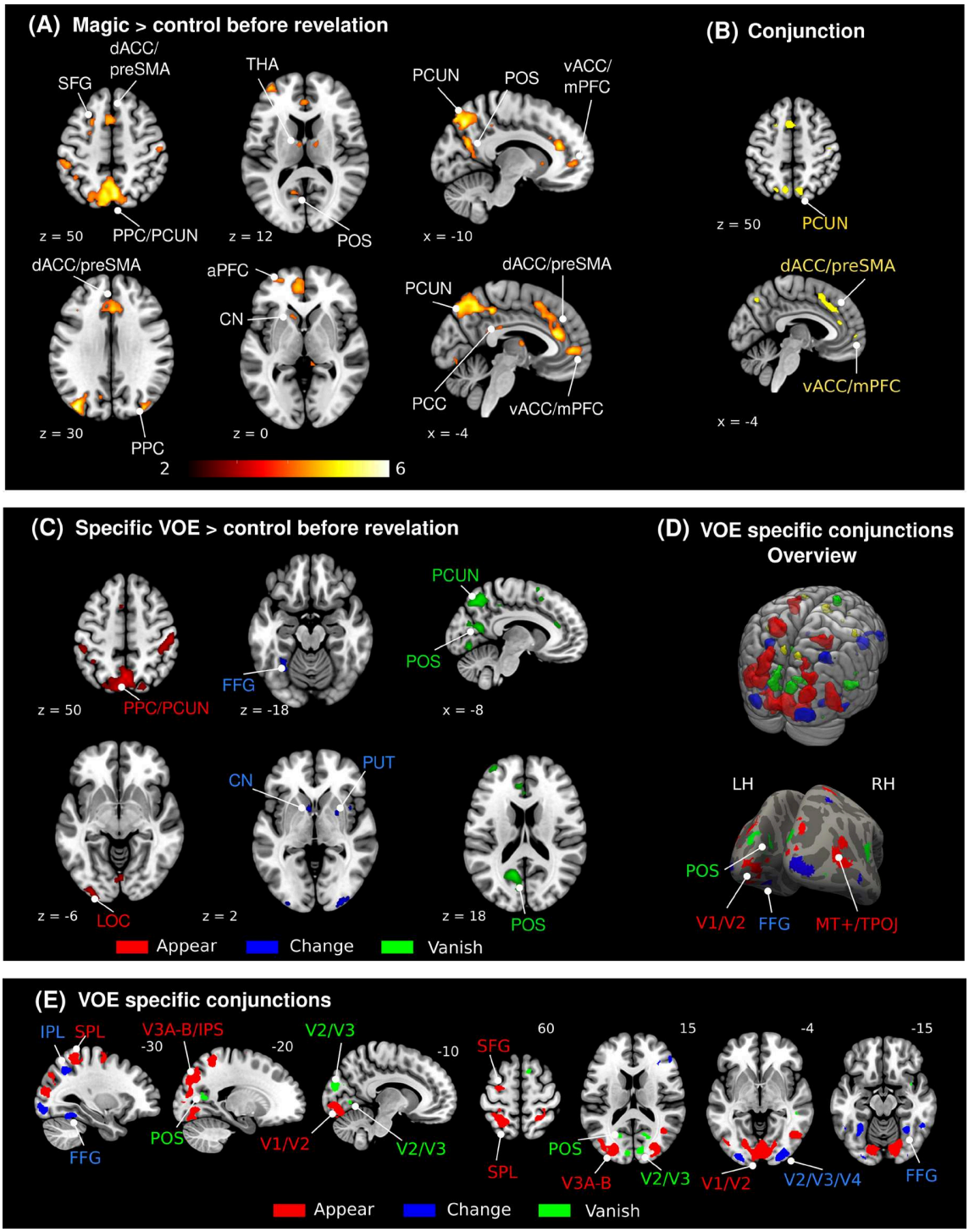


Fig. 3. (A), shown are active regions during viewing magic tricks compared to matched control videos before the participants knew how the tricks were performed (thresholded at $p_{unc} < 0.001$ and k =30, uncorrected). (B), shown are results from the conjunction analyses testing for generic activity (left, yellow) (at $p_{unc} < 0.005$). Only the dorsal ACC and preSMA and the precuneus (PCUN) revealed generic responses to high-level VOE at $p_{unc} < 0.001$. (C), VOE-specific activity revealing several posterior visual areas responsive to appearances (red), color changes (blue), and objects disappearing (green). (D), overview of all conjunctions results in the MNI152 template volume (upper) and of the VOE-specific responses projected on an average surface (fsaverage) (lower). (E), shown are VOE-specific conjunction results. ACC: anterior cingulate cortex; dACC: dorsal ACC; vACC: ventral ACC; CN: caudate nucleus; aPFC: anterior prefrontal cortex; FFG: fusiform gyrus; IPL: interior parietal lobe; IPS: intraparietal sulcus; MT+: motion area MT; mPFC: ventral prefrontal cortex; PCC: posterior cingulate cortex; PCUN: precuneus; POS: parieto-occipital sulcus; PPC: posterior parietal cortex; SFG: superior frontal gyrus; SMA: supplementary motor area; SPL: superior parietal lobe; THA: thalamus; TPOJ: temporo-parietal-occipital junction. LH: left hemisphere; RH: right hemisphere.

Table 1. Significant clusters of activity from the whole-brain contrast comparing responses between magic videos and matched controls ($Magic_{pre} > Control_{pre}$, thresholded at $p \leq 0.001$ and $k = 30$, uncorrected) and the conjunction analysis testing for common areas involved in the processing of violation of expectations (thresholded at $p_{unc} \leq 0.001$ and $k = 30$, uncorrected).

Brain region	AAL atlas labels	p(FWE) permutation	p(FWE) parametric	p(unc)	k	T	x	y	z
$Magic_{pre} > Control_{pre}$									
Posterior parietal cortex	**Occipital_Mid_L**	**0.0004**	**<0.001**	<0.0001	1871	7.16	-34	-82	34
	Precuneus_R	-	-	-	-	6.87	4	-68	52
	Precuneus_L	-	-	-	-	6.48	-10	-68	44
dACC	**Cingulate_Ant_L**	**0.0048**	**<0.001-**	0.0005	886	6.46	-6	30	20
	Cingulate_Ant_R	-	-	-	-	6.05	4	32	24
	Cingulate_Mid_R	-	-	-	-	5.51	2	24	32
L. Parieto-occipital sulcus	**Cuneus_L**	0.0594	**0.016**	0.0082	162	6.18	-14	-62	22
	Calcarine_L	-	-	-	-	4.21	-12	-60	14
R. Intraparietal sulcus	**Occipital_Mid_R**	0.1752	0.230	0.0284	74	5.53	36	-80	32
L. anterior PFC	**Frontal_Sup_2_L**	0.1982	0.288	0.0332	67	5.41	-26	58	4
L. anterior intraparietal	**Postcentral_L**	0.0548	**0.012**	0.0074	172	5.2	-46	-36	54
	Parietal_Inf_L	-	-	-	-	4.47	-42	-38	44
vACC/mPFC	**Cingulate_Ant_L**	0.0548	**0.012**	0.0074	172	5.13	-4	48	2
L. Superior frontal gyrus	**Frontal_Sup_2_L**	0.0522	**0.010**	0.007	179	5.02	-22	12	58
	Frontal_Sup_2_L	-	-	-	-	3.7	-22	0	52
L. anterior middle frontal gyrus	**Frontal_Mid_2_L**	0.1722	0.223	0.028	75	5.01	-34	54	14
L. Postcentral gyrus	**Postcentral_R**	0.1134	0.091	0.0173	103	4.71	52	-16	54
	Postcentral_R	-	-	-	-	4.07	52	-18	46
	Postcentral_R	-	-	-	-	3.72	46	-22	42
R. Superior frontal gyrus	**Frontal_Sup_2_R**	0.2394	0.395	0.0417	57	4.61	28	10	62
L. PCC	**no_label**	0.2832	0.503	0.0511	49	4.48	-2	-30	26
Conjunction of $Magic_{pre} > Control_{pre}$ per VOE type									
dACC/preSMA	**Supp_Motor_Area _L**	n.a.	0.722	0.096	38	4.02	-4	16	50
Precuneus									
	Precuneus_R	n.a.	0.722	0.096	38	3.94	8	-68	48

p(FWE) permutation shows permutation-based cluster statistics. p(FWE) parametric shows parametric cluster statistic. Clusters that survive FWE-correction are highlighted in bold. k = cluster size, T = t statistic at peak voxel, x, y, z = peak voxel MNI coordinates [mm]. dAAC: dorsal anterior cingulate cortex; PFC: prefrontal cortex; vACC: ventral ACC; mPFC: medial PFC; PCC: posterior cingulate cortex; DLPFC: dorso-lateral prefrontal cortex; preSMA: supplementary motor area.

Fig. S3 and Tables S3–S6). Additionally, the right anterior insula, as well as wide-spread occipital and parietal sensory areas (e.g., V1–V3, LO1–2 and IPS) showed increased activity during the presentation of magic videos compared to the videos showing unusual actions. However, a similar pattern of results was also observed when comparing control videos to unusual action videos (*Control* > *Unusual*) (see Supplementary Fig. S3A–B and Tables S3–S6), suggesting that the responses observed may not be solely driven by surprise, but by other factors instead (e.g., differences in visual content).

Inverting the magic related contrasts (i.e., $Control_{pre}$ > $Magic_{pre}$ and $Unusual_{pre}$ > $Magic_{pre}$) revealed, among others, engagement of temporal sulcus (STS) extending into the temporo-parietal junction (TPJ) and the ventro-medial prefrontal cortex (vmPFC) (see Supplementary Fig. S3C), which are areas known to be involved in social cognition (Deen et al., 2015; Hiser & Koenigs, 2018; Lahnakoski et al., 2012; Saxe & Kanwisher, 2003).

3.3.2. Generic surprise responses

To specifically test for generic surprise responses in the brain showing a significant involvement of all three types of VOE, we performed a conjunction analysis combining all different VOE types before the revelation of the tricks ($Magic_{A\_pre}$ > $Control_{A\_pre}$ ∩ $Magic_{C\_pre}$ > $Control_{C\_pre}$ ∩ $Magic_{V\_pre}$ > $Control_{V\_pre}$). We found clusters in the dorsal anterior cingulate cortex (dACC) bilaterally and right posterior parietal cortex (precuneus) revealing generic responses (see Fig. 3B, left and Table 1). Also, using a more liberal threshold of $p_{unc} \leq 0.005$ (and k = 10), we further observed generic activity in the vACC/mPFC and the left precuneus.

3.3.3. Specific surprise responses

After establishing what areas are generally involved in the processing of violation of expectations (i.e., commonly active in seemingly impossible appearances, disappearances, and color changes), we looked for VOE type-specific differential activity in the brain (e.g., areas responsive to something unexpected appearing but not disappearing or changing color). Beyond the systematic generic activation of frontal and parietal areas described above, all three types of VOE evoked responses in posterior visual areas (see Fig. 3C and Supplementary Table S8). Activations specific to each type of VOE were largely confined to posterior sensory areas: appearances induced an increase in activity in V1 and the intraparietal sulcus, color changes in areas of the inferior temporal cortex and ventromedial visual cortex and vanishes in the parieto-occipital sulcus. To better visualize this diverse modulation of sensory areas by the different VOE, we tested which areas were significantly more activated by one type of VOE than by the other two using a conjunction test (e.g., test for appear-specific responses: $Magic_{A\_pre} > Magic_{C\_pre} \cap Magic_{A\_pre} > Magic_{V\_pre}$) (see Fig. 3D–E and Supplementary Table S7). Activity related to objects appearing was observed in more medial parts of early visual areas (peak coordinates: x = 30, y = -88, z = 22), in higher-level visual areas of the lateral and middle occipito-temporal cortex (peak coordinates: x = 46, y = -66, z = 2) and bilateral intraparietal sulcus (peak coordinates: x = 30, y = -88, z = 22 and x = -22, y = -78, z = 44), while activity related to color changes was observed specifically in more posterior parts of secondary visual areas (V2/V3/V4, peak coordinates: x = ±30, y = -96, z = -6), bilaterally in the IPS (peak coordinates: left IPS x = -28, y = -60, z = 4 and right IPS x = 32, y = -54, z = 44) and in color-responsive ventral areas of the fusiform gyrus (FFG, peak coordinates: x = -30, y = -50, z = -16 and x = 32, y = -54, z = -14). Finally, vanishing objects evoked significant responses in the anterior parts of the calcarine sulcus, close to the parieto-occipital sulcus (peak coordinates: x = -20, y = -64, z = 12 and x = 18, y = -76, z = 8) and in the right superior temporal sulcus (STS, peak coordinates: x = 54, y = -38, z = 6).

3.3.3.1. Control analysis for visual content. As these VOE type-specific patterns of activity are located predominantly in visual processing areas, they are potentially related to general visual differences between the conditions tested. The compared videos are not only showing “appearances”, “color changes”, and “vanishes”, but also “red objects”, “blue objects”, or “no objects”, respectively.

To test if these VOE type-specific responses are confounded by different visual input, we performed a control analysis comparing neural responses between 1) videos showing red objects, either as the product of a magic trick (appearances) or as a control to other tricks (vanishes) ($Magic_{A\_pre}$ > $Control_{V\_pre}$) and 2) between videos showing no object, either as the product of a magic trick (vanishes) or as a control to the other tricks (appearances) ($Magic_{V\_pre}$ > $Control_{A\_pre}$). Differential responses in posterior visual areas to these control contrasts were similar to the results of the corresponding conjunction analyses (for appearances and vanishes), suggesting that violation-specific responses observed in posterior visual areas are unlikely to be driven merely by visual content (see overlays in Supplementary Fig. S5). However, if these signals in visual areas reflect specific prediction errors based on different VOE, we would additionally expect them to be modulated by prior

knowledge and to show decreased responses after the explanation of the tricks.

3.3.4. *Prior-knowledge dependent whole-brain responses*

To investigate the effect of prior knowledge on brain responses, we provided participants with the methods behind the magic tricks. Surprisingly, neural activity after explanation of the tricks ($Magic_{post}$ > $Control_{post}$) was similar to that before the explanation of the tricks (see Fig. 4A and Supplementary Table S10). Thus, areas related to the processing of surprising events remained significantly active in view of VOE also after the explanation of the tricks (i.e., dACC, caudate nucleus, anterior insula). The interaction contrast comparing surprise responses before and after providing the explanations ($Magic_{pre}$ > $Control_{pre}$) > ($Magic_{post}$ > $Control_{post}$) revealed only a small number of areas showing prior-knowledge dependent modulations (see Fig. 4B–C and Table 2). We observed higher activation in a large cluster of the medial prefrontal cortex (mPFC) and ventral ACC (peak coordinates: x = 2, y = 46, z = -10) and right posterior cingulate cortex (PCC) with FWE-correction (peak coordinates: x = 6, y = -46, z = 8). These regions, hence, decreased magic-related activity after the revelation. These areas overlap with and constitute a subset of the activity observed with $Magic_{pre}$ > $Control_{pre}$ before the revelation of the tricks.

Interestingly, the observed decrease of activity in the vACC/mPFC and PCC after revelation of the tricks coincides with the midline core areas of the default mode network (DMN), while parietal areas of the dorsal attention network (DAN) showed increased activity. This indicates that our findings are unlikely a result of a decrease in attention, as the DAN is known to direct top-down attention (Corbetta & Shulman, 2002). A visualization of these patterns of activity, together with the DMN and DAN is shown in Figure 4C.

3.3.4.1. No prior-knowledge dependency of trick-specific modulations. While prior-knowledge driven signals overlapped with neural activations to unspecific VOE ($Magic_{pre}$ > $Control_{pre}$), none of the previously observed trick-specific visual areas showed modulation as a factor of prior knowledge. Further, the contrasts investigating the effect of prior knowledge for each VOE type separately revealed similar response patterns to those reported in

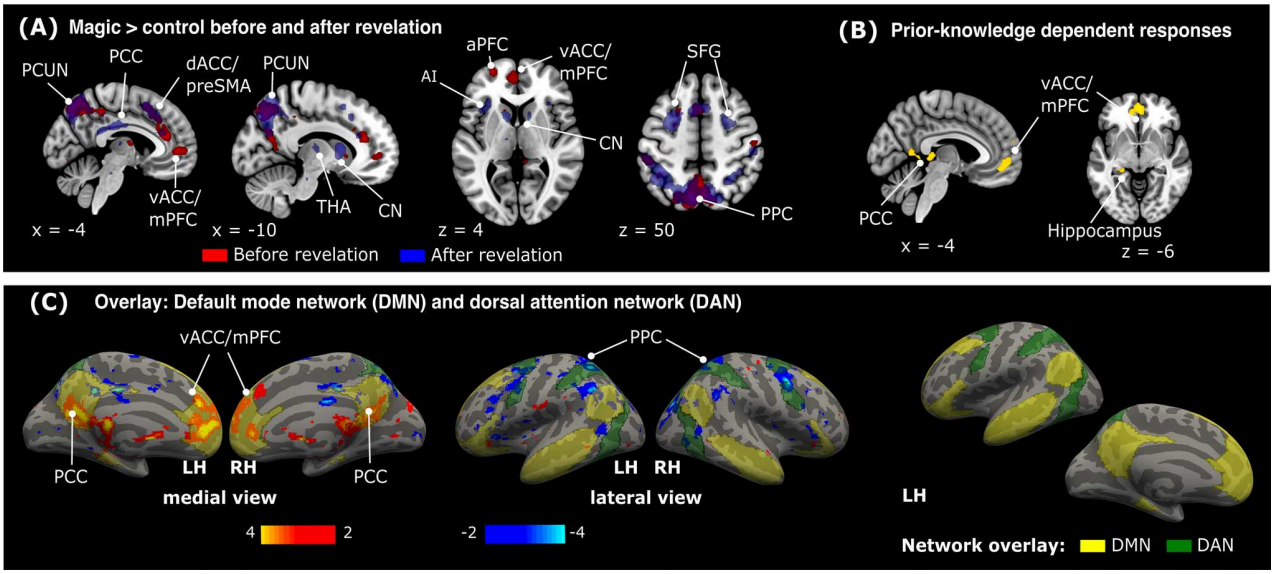


Fig. 4. Prior knowledge dependent modulation of brain responses. (A), overlay of *Magic* > *Control* before (red) and after (blue) explanation of the magic tricks. Both contrasts reveal a consistent activation of surprise related areas, such as the dorsal ACC (dACC), caudate nucleus (CN), and posterior parietal cortex (PPC). (B), the difference of both contrasts (i.e., the interaction between video condition and revelation) revealed only few areas that were significantly more active before the explanation of the tricks: the posterior cingulate cortex (PCC), the ventral ACC/medial prefrontal cortex (vACC/mPFC), and left Hippocampus. Threshold at $p_{unc} < 0.001$ and k = 30. (C), shown are the interaction testing for prior-knowledge dependent responses (thresholded from $t = 2$ to $t = 4$, red to yellow and $t = -2$ to $t = -4$, dark blue to light blue) and the default mode network (DMN) and the dorsal attention network (DAN) as transparent overlay projected on to an average brain surface (fsaverage). The network overlays are provided by Yeo et al. (2011). AI: anterior insula; aPFC: anterior prefrontal cortex; LH: left hemisphere; PCUN: precuneus; preSMA: supplementary motor area; RH: right hemisphere; SFG: superior frontal gyrus; THA: thalamus.

Table 2. Results of contrasts comparing neural responses to VOEs before and after explanation of the tricks (thresholded at $p_{unc} < 0.001$ and k = 30, uncorrected).

Brain region	AAL atlas labels	p(FWE) permutation	P(FWE) parametric	p(unc)	k	T	x	y	z
vACC/mPFC	**Frontal_Med_Orb_R**	**0.0068**	**<0.001**	0.0007	538	5.72	2	46	-10
	Frontal_Med_Orb_L	-	-	-	-	4.96	-10	42	-6
R. Cuneus (V3)	**Cuneus_R**	0.25	0.466	0.0385	52	5.23	8	-86	34
L. POS	**Precuneus_L**	0.1082	0.100	0.014	101	4.99	-8	-56	8
L. PCC	**Calcarine_L**	-	-	-	-	4.31	-6	-48	4
R. PCC	**Precuneus_R**	**0.0424**	**0.009**	0.005	185	4.91	6	-46	8
	no_label	-	-	-	-	4.59	-4	-30	6
	no_label	-	-	-	-	4.56	8	-36	4

p(FWE) permutation shows permutation-based cluster statistics. p(FWE) parametric shows parametric cluster statistic. Clusters that survive FWE-correction are highlighted in bold. k = cluster size, T = t statistic at peak voxel, x, y, z = peak voxel MNI coordinates [mm]. ACC: anterior cingulate cortex; vACC: ventral ACC; mPFC: medial prefrontal cortex; POS: parieto-occipital sulcus; PCC: posterior cingulate cortex; dACC: dorsal ACC.

the main contrasts (see Supplementary Fig. S4). This indicates that visual areas responsive to a specific VOE (e.g., an object changing color) were not modulated by prior knowledge.

3.3.4.2. Controlling for time effects in net responses. To rule out possible time confounds in results comparing responses before and after revelation of the tricks, we inspected how the individual values changed across time as a control analysis by means of a mixed-effects model. Consistent with responses being driven by prior-knowledge, a pre-post predictor was significant in all clusters (all $p_{corr} < 0.05$). However, in one cluster (right Cuneus, x = 8, y = -86, z = 34; k = 52) a significant amount of variance was also explained by the time-decay regressor modeling the time of the experiment, showing that the activity in that cluster had a contribution of time. Yet, the fact that in all clusters the pre-post regressor remained significant despite inclusion of the time-regressor shows that knowledge-dependent effects were true (see Supplementary Fig. S6 and Table S9).

3.4. Univariate ROI analysis

To complement the whole-brain analysis, we defined a set of 26 hypothesis-driven regions-of-interest (ROIs) from which we extracted and analyzed parameter estimates and tested them with paired t-tests (correcting for the number of ROIs). Overall, our ROI-based analyses confirmed our above whole-brain findings. None of the lower-level visual cortices showed a significant increase of neural responses during the magic compared to the matched control condition before the explanation of the tricks (all $p_{unc} > 0.24$). In contrast, significant responses to magic were observed in the visual parietal ROI IPS ($t(23) = 2.746$, $p_{unc} = 0.012$, Cohen's $d = 0.53$), as well as in several higher-level surprise-related ROIs, especially in adACC ($t(23) = 5.32$, $p_{corr} = 0.001$, *Cohen's d* = 0.917), as well as in vACC, pvACC, BA6, BA46, and caudate nucleus (all $p_{unc} < 0.05$, see detailed results in Supplementary Tables S14–S17). Only two ACC ROIs (adACC and vACC) and 8BM (directly superior to the adACC) showed a decrease of activity after the revelation of the tricks ($t(23) = 3.43$, $p_{unc} = 0.002$, Cohen's $d = 0.41$, $W = 79$, $p_{unc} = 0.042$, Cohen's $d = 0.42$ and $t(23) = 2.54$, $p_{unc} = 0.01$, Cohen's $d = 0.41$, respectively). None of the visual ROIs showed a similar response pattern. In contrast, VOE-specific responses were largely constrained to visual ROIs (see Supplementary Section *Univariate ROI results*).

3.5. Multivariate pattern analysis

Our univariate analyses revealed generic responses (modulated by prior knowledge) in frontal and parietal areas, while showing trick-specific responses in posterior sensory areas (unaffected by prior knowledge). However, differences in specific VOE and prior-knowledge modulations might also be reflected in activation patterns and not only in net signal differences. Accordingly, we complemented our univariate analysis by a multivariate pattern analysis to investigate if specific information about VOE types were present in the activity patterns of our surprise-related frontal and parietal ROIs (which showed only generic responses to magic).

3.5.1. Visual ROIs

Decoding of the specific VOE types across objects was possible in all posterior visual ROIs (V1, V2, V3, hV4, V3A, V3B, LO, VO, IPS) before the explanation of the magic tricks (all corrected p-values < 0.001; corrected using a permutation maximal statistic, Nichols & Holmes, 2002). After revelation of the method behind the magic tricks, decoding accuracies significantly dropped in most ROIs (V1, V2, V3, LO, and IPS) (all $p_{corr} <= 0.05$), being below

threshold in LO and IPS (see Fig. 5B and Supplementary Table S20). Additionally, we found uncorrected significant differences in hV4 and V3B (p_{unc} = 0.05 and p_{unc} = 0.016, respectively). Moreover, the peak of decoding accuracies was around the moment of VOE and reduced after the explanation of the tricks (see Fig. 5C).

3.5.2. Surprise related ROIs

In contrast, no surprise-related ROI (nor the control ROI) showed significant above-chance decoding accuracies between VOE types using the permutation-max statistic for correction, except for the PH ROI (p_{corr} < 0.001), an area located in the inferior temporal sulcus (temporo-occipital division), using data before revelation of the tricks. Uncorrected significant above chance decoding (chance level = 33%) was observed in IFJ, AI, pdACC, BA6, BA8, and BA46 using data before revelation. Differences in decoding accuracies before and after revelation were observed only in PH (p_{corr} = 0.005) (see Fig. 5B and Supplementary Table S20) and in BA8 (corrected for the number of ROIs that could *not* significantly decode the VOE type using pre-revelation data, i.e., 16 ROIs) (see Fig. 5B and Supplementary Table S21).

Therefore, while frontal and parietal areas showed a generic and surprise-dependent involvement in processing VOE in the univariate analyses, they carried no or only weak information (i.e., in BA8) as to what exactly happened. In contrast, information about specific VOE was observed in all posterior sensory areas across the visual hierarchy.

3.5.3. Searchlight analysis

Consistent with the ROI analysis, results of the whole-brain searchlight analysis revealed that only posterior visual areas of the brain could significantly decode the magic type in both conditions (using a permutation-bootstrap hybrid correction method, Stelzer et al., 2013). As shown in Figure 5D, significant decoding was possible in large areas of the visual cortex, including most of the occipital cortex and small parts of the temporal and parietal cortex. Crucially, decoding accuracies before the explanation of the magic tricks were significant in more voxels and larger clusters compared to after explanation. Significant decoding before the explanation of the tricks extends to parts of temporal and parietal cortex, whereas significant decoding after revelation is largely restricted to posterior visual areas. In sum, and in contrast to the univariate results that showed no net modulation as a function of prior knowledge (and surprise) in visual areas, the differences in decoding accuracies using data before and after explanation of the tricks suggest that visual areas are, indeed, sensitive to changes in knowledge and encode specific expectations.

3.5.4. Controlling for time effects in pattern activity

Arguably, suprathreshold decoding of magic effects in posterior sensory areas may reflect general stimulus differences in the moment of magic between the different magic trick types independent of the object used, as appear videos systematically showed red objects, color changing videos blue objects, and vanishing videos no object. However, the observed significant differences in decoding accuracy before and after revelation suggest surprise-dependent modulations of activity patterns, as these differences are present in view of the very same videos. Yet, these differences could reflect time- and/or design-related confounds, such as a general decrease of attention and alertness over time. However, since we did not find any significant changes in univariate comparisons in posterior visual areas and we observed a general increase in parts of the dorsal attention network, we believe that our results are not confounded by time and/or design related factors. Nonetheless, we performed control analyses comparing decoding accuracy of objects present or absent in control videos in the pre versus post revelation phase. These analyses showed no modulation of time in control videos (detailed results can be found in the Supplementary Section *MVPA control analyses*).

4. DISCUSSION

In this fMRI study, we used naturalistic video stimuli showing magic tricks and matched control actions to investigate responses to violation of expectations (VOE) of deeply held beliefs about the physical world. We used three distinct magic types (object appearance, object disappearance, and feature-change) that were presented with and without prior knowledge about the underlying deceptive methods (i.e., sleights-of-hand). Each magic type was presented using three distinct objects to allow for object-invariant classification of magic types. We looked for 1) generic prediction error responses to perceived violation of physical principles, 2) specific responses to the different magic types, and 3) effects of the viewers' prior knowledge on prediction error processing, for both generic and specific responses.

Our results revealed a hierarchy of surprise signals. First, we observed generic effects of world-model VOE (i.e., common to all magic types) in several clusters of the prefrontal and parietal cortex (such as the dorsal and ventral ACC and the posterior parietal cortex) and a correlation of their responses with subjective surprise ratings. Then, differential activity specific to the different

types of magic was evoked predominately in posterior visual areas of the occipital and parietal cortex. These specific prediction error signals were evident in the univariate analyses and in decoding of the magic types, both of which were largely confined to posterior areas across the visual hierarchy. Finally, following explanation of the tricks, responses were largely unaffected by participants' knowledge and only decreased in select parts of the network showing generic effects of VOE (midline areas of the default mode network). While net activity in visual areas was not significantly modulated by the prior knowledge, decoding of VOE type-specific signals was sensitive to changes in the participants knowledge, showing decreased decoding when participants knew the tricks. These results suggest that higher-level predictive information affects even the earliest levels of cortical visual processing (V1–V3).

4.1. Generic responses to violation of expectations

Witnessing magic events that violate intuitive physical principles evoked activity in a large network of frontal and parietal (dACC, vACC/mPFC, and posterior parietal cortex) and subcortical (caudate nucleus) areas, with no involvement of lower-level sensory areas. This pattern of activity is consistent with that observed in a recent large meta-analysis of surprising events (Fouragnan et al., 2018), and with previous experiments investigating surprise responses using naturalistic videos, such as magic tricks (Danek et al., 2015), computer-generated animations (Bardi et al., 2017), or learned sequences of movements (Schiffer & Schubotz, 2011). Accordingly, our results add to prior evidence showing the key role of the dACC in processing incongruent information (Alexander & Brown, 2011, 2019) and of the caudate nucleus in signaling unexpected and rewarding events (Schultz et al., 1997; Wittmann et al., 2008; Zink et al., 2003).

Most importantly, our results suggest that higher-level VOE in view of seemingly impossible events are processed similarly to breaches of lower-level expectations, such as the presence of infrequent stimuli, unlikely events (Kim, 2014), or changes in the stimulus sequences (Downar et al., 2000; Grundei et al., 2023). This suggests the existence of a dedicated network including frontal and parietal areas that correlate with subjective surprise signals, that signal the detection of incongruent information in the human brain (irrespective of sensory modality and abstraction level).

4.2. Specific responses to violation of expectations

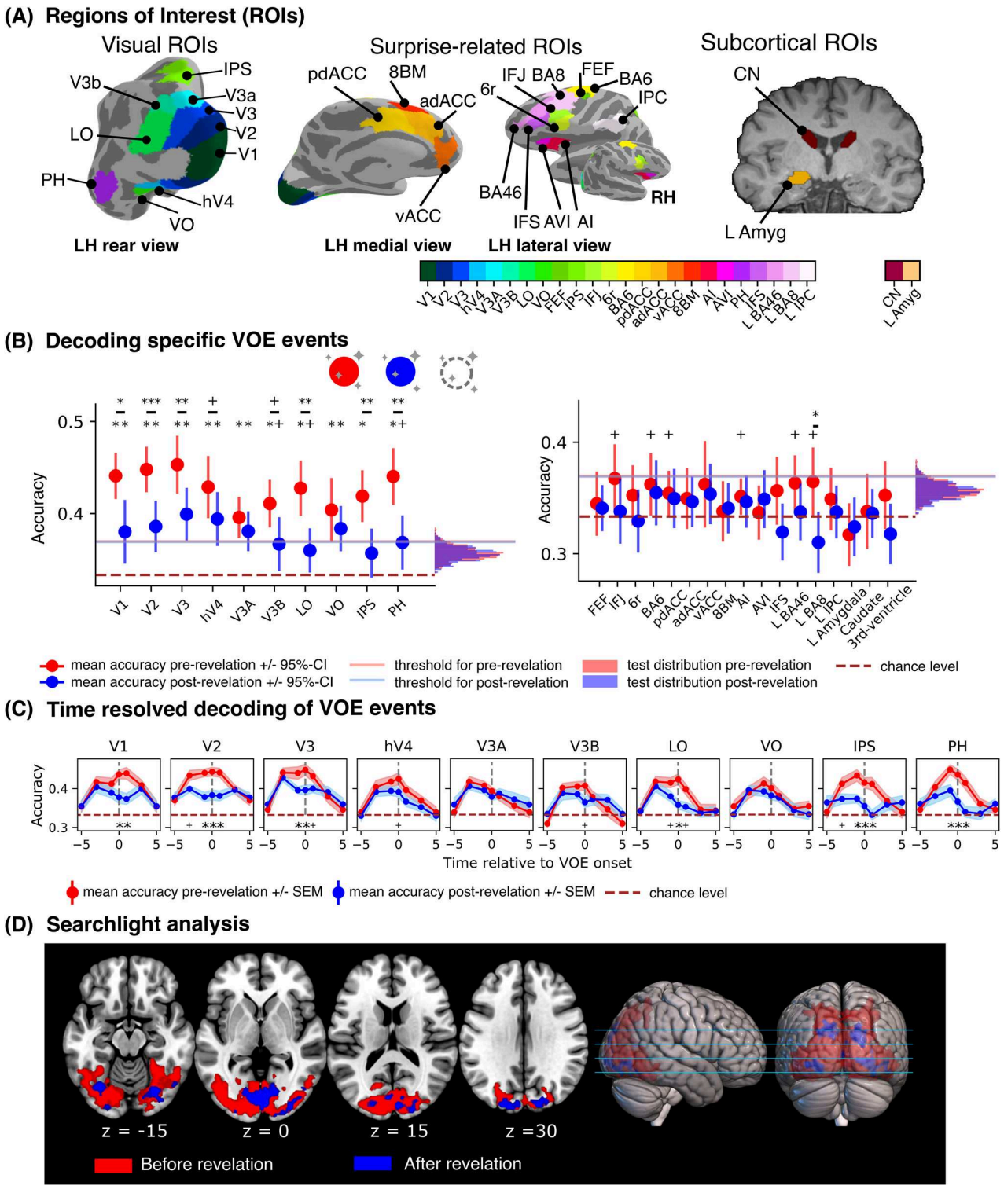
Complementing the generic frontal and parietal involvement in processing naturalistic violations of physical principles, we further looked into the specific effects of the different types of VOE (appear, change, vanish). Specific responses evoked by the different VOE in net activity and multivariate activation patterns (i.e., allowing for a distinction between trick-types) were observed predominately in posterior sensory areas. In contrast, frontal areas revealed no or only weak specific VOE responses. Since this divergent pattern of net activity was only observable when looking into the individual VOE, it is possible that previous studies failed to report the involvement of sensory areas because of pooling responses to different types of VOE (Danek et al., 2015; Liu et al., 2024; Parris et al., 2009).

Previous results using dynamically occluded stimuli report the neural representation of occluded objects in

Fig. 5. Results of decoding analyses (A), regions of interest (ROIs) used in the experiment. Visual ROIs are shown left, all of which (except for the inferior temporal area PH) were defined using a probabilistic map of visual areas (Wang et al., 2015). Surprise-related ROIs presented on the right panel were defined based on previous results (Danek et al., 2015; Parris et al., 2009) using a multimodal parcellation atlas (Glasser et al., 2016), except for frontal eye field (FEF), which was also defined using the probabilistic atlas from Wang et al. (2015) (see Supplementary Section *Surprise-related region of Interest definition* for more information). The two subcortical ROIs, caudate nucleus, and left amygdala were defined using the Freesurfer automatic parcellation (Fischl et al., 2002). (B), shown are the decoding accuracies for decoding the VOE types over objects in a three-fold cross-decoding approach in our theory-driven ROIs (left: ROIs that significantly decoded the VOE type using pre-revelation data, right: ROIs that did *not* significantly decode the VOE type using pre-revelation data). Decoding was performed with data before (red) and after (blue) revelation. All statistics were corrected for multiple tests by using the max-statistic correction across all ROIs (Nichols & Holmes, 2002). Decoding accuracies pre and post revelation were compared using paired t-tests. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, Holm-Bonferroni corrected; $^{+}p < 0.05$, uncorrected. (C), temporal development of decoding accuracies relative to VOE onset. Decoding was performed using beta estimates that modelled moments 5 s before to 5 s after VOE onset in steps of 2 s (-5, -3, -1, +1, +3 +5 s with respect to the VOE onset at 0 s). Decoding accuracies pre-revelation peak around VOE onset and are significantly higher than decoding post-revelation. All statistics were corrected for the number of time steps. *$p < 0.05$, Holm-Bonferroni corrected; $^{+}p < 0.05$, uncorrected. (D), whole-brain searchlight decoding results. We can significantly decode VOE types in the majority of visual cortex before revelation and less so after revelation (correcting using a permutation-bootstrap hybrid method, Stelzer et al., 2013).

posterior visual areas (Erlikhman & Caplovitz, 2017; Hulme & Zeki, 2007; Olson et al., 2004) and in neurons of the inferotemporal cortex of macaque monkeys (Puneeth & Arun, 2016) at different levels of complexity (e.g., occluded faces selectively engaged the fusiform face areas, Hulme & Zeki, 2007). The observed differential activity in visual areas using naturalistic stimuli in the present study are likely VOE type-specific surprise signals when violating said representations.

The following reasons support this interpretation: First, net responses were not driven by differences in visual content (because they were evident across distinct objects, and additional control contrasts ruled content-driven responses out). Second, decoding did not work in



the absence of VOE when we used similar sensory occurrences using data from the matched control videos. Third, prior knowledge significantly reduced decoding accuracies. And finally, most observed responses occurred in functionally specialized regions of the visual cortex. For example, the unexpected appearance of objects evoked activity in the object-responsive LOC and the perception of unexpected colors evoked activity in the ventral color areas of the fusiform gyrus. Both of these are compatible with prior evidence showing increased activity in the LOC upon perceiving unexpected objects (Richter et al., 2018) and in color areas when viewing unexpected colors (Jiang et al., 2016; Stefanics et al., 2019).

Hence, our results suggest that memory-based expectations related to higher-level principles affect visual processing already at the earliest cortical visual processing areas: they encode information about the presence, absence, and features of objects. This is consistent with recent neuroimaging evidence suggesting that responses in early visual areas can be modulated by high-level visual surprise signals (Richter et al., 2024).

4.3. Hierarchical prediction errors in naturalistic perception

The observation of surprise-related information in posterior visual areas is in line with a variety of higher-level memory-based signals that have been reported in visual areas, such as memory color (Bannert & Bartels, 2013), scene context (Muckli et al., 2015), scene segmentation (Grassi et al., 2017, 2018; Scholte et al., 2008), expected visual stimuli (Ekman et al., 2017), and working memory (Harrison & Tong, 2009). Importantly, these signals have been interpreted as evidence of recurrent predictive signals from higher-level areas, as they encoded information that is not thought to originate from V1 (such as memory color, 3D, or Gestalt and scene information). Consistent with this, further studies located corresponding signals in superficial and/or deeper layers of the cortex using laminar fMRI (Aitken et al., 2020; Lawrence et al., 2018; Muckli et al., 2015) or electrophysiological measurements in monkeys (e.g., Papale et al., 2022; Self et al., 2013).

Together, the current results fall in line with predictive coding theories and extend them to VOE regarding higher-level world models (Friston, 2005; Lee & Mumford, 2003; Rao & Ballard, 1999). We show a clear dissociation: generic responses to VOE in higher-level frontal and parietal areas and segregated surprise responses in functionally specialized lower-level sensory areas (involved in the processing of the expected information). This reflects the hierarchical structure of our internal world model (Clark, 2013; Hohwy, 2014), with frontal and parietal areas involved in representing more abstract aspects of the world (such as object permanency), while sensory areas represent lower-level inferences about the immediate and detailed features (such as color and shapes). Accordingly, the observed surprise-related responses in lower-level sensory areas can be thought of as the product of a mismatch between top-down predictions ("a red ball") fed back to lower-level areas to be compared with incoming sensory evidence ("a blue ball") presumably via feedback connections.

4.4. Knowledge-dependent modulations

To further investigate these hierarchical VOE responses, we probed how prior knowledge affected them. We hypothesized that providing the participants with knowledge about the mechanics of the trick for each video would avert VOE: why should we be surprised when viewing a disappearing ball when we know how and that the magician is actually hiding it behind his hands?

Surprisingly, while participants' subjective surprise ratings were significantly reduced after providing them with the explanation of the tricks, net brain responses were almost indistinguishable: areas involved in the processing of unexpected events, such as the dACC, anterior insula, and caudate nucleus (Fouragnan et al., 2018) were systematically active when observing the magic videos even after participants had rational explanations for the tricks. As neural responses were reminiscent to those signaling surprise, it suggests that repeated viewing of explainable events did not prevent VOE.

This intriguing observation likely reflects that people can be moved by things they know to be unreal, such as fictions (i.e., the "paradox of fiction", see Radford & Weston, 1975) or magic illusions (i.e., the "paradox of theatrical magic", see Grassi et al., 2024). This is akin to how we still perceive visual illusions, even if we know how they work. For example, when a magician convincingly saws someone in half on stage, the audience is genuinely moved by the illusion (i.e., surprised), but do not attempt to prevent it nor call the police (because they know it is unreal). Our results suggest that the compelling perceptual illusions that magic provides are initially appraised as surprising, even with existing prior-information (cf., Grassi et al., 2024).

In turn, the only areas whose activity decreased following the explanation of the tricks were two midline core areas of the default mode network (DMN), the ventral ACC/mPFC, and the posterior cingulate cortex. Crucially, our control analysis showed that these modulations could not be explained by the repetition suppression (Grill-Spector et al., 2006; Krekelberg et al., 2006) and condition-independent time effects. The modulation of midline core areas of the DMN by prior knowledge is

consistent with recent reports, showing their involvement in processing surprising events in movies (Brandman et al., 2021), jokes (Jääskeläinen et al., 2016) and structured events unfolding in time (Baldassano et al., 2018; Regev et al., 2013; Simony et al., 2016). Based on these findings, it has been suggested that the DMN is not to be understood exclusively as an "intrinsic" network (as originally proposed, cf., Raichle, 2015), but as a dynamic "sense-making" network involved in the creation of rich models of events by integrating incoming information with prior knowledge as they unfold over time instead (Stawarczyk et al., 2021; Yeshurun et al., 2021).

Here, we show that areas of the "sense-making" network may be sensitive to the rational explanation of magic tricks, whereas all other identified surprise-related regions continued to be sensitive to the VOE (even once the tricks were understood). The decreased involvement of areas of the DMN, together with an increase of activity in frontal and parietal areas of the dorsal attention network (DAN), is consistent with a reduction in prediction error (surprise) signals related to narrative understanding (engaging the DMN) and an increase of top-down attention after explanation of the tricks (engaging the DAN).

4.5. Further considerations

Finally, in addition to the violation of intuitive physics and related surprise signals, further cognitive processes should be considered for the interpretation of our results. First, as all our stimuli include a human performer, our results could include neural responses related to action understanding and social cognition. Yet, notably, areas commonly related to social cognition, such as the STS and the TPJ (Deen et al., 2015; Haxby et al., 2000; Lahnakoski et al., 2012; Saxe & Kanwisher, 2003), were not involved in the perception of magic tricks. Instead, these areas related to social cognition showed an increased involvement when perceiving VOE based on unusual actions (i.e., $Unusual_{pre} > Magic_{pre}$, see also $Control_{pre} > Magic_{pre}$). This is consistent with previous studies revealing an involvement of the STS in the perception of social and psychological VOE, such as irrational actions (Brass et al., 2007; Jastorff et al., 2011; Marsh et al., 2014; Shultz et al., 2011; Vander Wyk et al., 2009), further supporting the notion of functionally-specialized surprise signals.

Moreover, magic tricks, and surprising events in general, capture our attention (Horstmann, 2015) and induce curiosity and information-seeking behaviors (Danek et al., 2014; Lau et al., 2020). Hence, we cannot rule out the involvement of these additional cognitive processes during the perception of seemingly impossible events. However, attentional modulation appears not to drive our effects observed in sensory regions, as they revealed no net BOLD modulation between pre- and post revelation, in contrast to the typically strong net attentional modulation observable in these regions (Jehee et al., 2011; Somers et al., 1999; Tootell et al., 1998). Finally, please note that with concern to surprise-related signal in higher-level regions, differences between pre- and post-revelation activity are unlikely to be related to a set-wise decay in attention, as we see an increase of activity in areas of the dorsal attention network (DAN) post-revelation.

4.6. Conclusion

We used a naturalistic paradigm to violate deeply held beliefs of our physical world, involving three types of expectation violations (object appearance, color change, and object disappearance). Our results show a hierarchy of surprise signals: generic responses to unexpected events in frontal and parietal areas, and responses specific to the type of VOE in distinct functionally specialized sensory areas. Our results suggest that world-model VOE are processed similarly to other surprising events in dedicated areas of the prefrontal and parietal cortex and striatum, and that core midline areas of the default-mode network decrease their involvement once rational understanding is established. Most importantly, we show that early and functionally specialized areas of the visual cortex encode memory-based predictions about the presence, absence, and features of objects.

DATA AND CODE AVAILABILITY

Code (for preprocessing, analysis, and visualisations) and preprocessed data for group analyses can be found at https://osf.io/kn2af/?view_only=067b698a4567441e93b01518a88860a0. Raw MRI data can be provided upon reasonable request.

AUTHOR CONTRIBUTIONS

Vincent Plikat: conceptualization, formal analysis (lead), investigation (equal), visualisation, writing—original draft, and writing—review and editing. Pablo R. Grassi: conceptualization (lead), formal analysis (supporting), investigation (equal), methodology, supervision (equal), visualization, writing—original draft, and writing—review and editing. Julius Frack: resources. Andreas Bartels: conceptualization, methodology, supervision (equal), and writing—review and editing.

DECLARATION OF COMPETING INTEREST

We declare no conflict of interests.

ACKNOWLEDGMENTS

This work was supported by the Barbara-Wengeler-Foundation, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 465409366, and by the Max Planck Society. We further acknowledge support from the Open Access Publication Fund of the University of Tübingen.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: https://doi.org/10.1162/imag_a_00459.

REFERENCES

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14. https://doi.org/10.3389/fninf.2014.00014

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the "light-from-above" prior. *Nature Neuroscience*, *7*(10), 1057–1058. https://doi.org/10.1038/nn1312

Aitken, F., Menelaou, G., Warrington, O., Koolschijn, R. S., Corbin, N., Callaghan, M. F., & Kok, P. (2020). Prior expectations evoke stimulus-specific activity in the deep layers of the primary visual cortex. *PLoS Biology*, *18*(12), e3001023. https://doi.org/10.1371/journal.pbio.3001023

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. https://doi.org/10.1038/nn.2921

Alexander, W. H., & Brown, J. W. (2019). The role of the anterior cingulate cortex in prediction error and signaling surprise. *Topics in Cognitive Science*, *11*(1), 119–135. https://doi.org/10.1111/tops.12307

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

Bannert, M. M., & Bartels, A. (2013). Decoding the yellow of a gray banana. *Current Biology*, *23*(22), 2268–2272. https://doi.org/10.1016/j.cub.2013.09.016

Bardi, L., Desmet, C., Nijhof, A., Wiersema, J. R., & Brass, M. (2017). Brain activation for spontaneous and explicit false belief tasks overlaps: New fMRI evidence on belief processing and violation of expectation. *Social Cognitive and Affective Neuroscience*, *12*(3), 391–400. https://doi.org/10.1093/scan/nsw143

Bartels, A., & Zeki, S. (2000). The architecture of the colour centre in the human visual brain: New results and a review. *European Journal of Neuroscience*, *12*(1), 172–193. https://doi.org/10.1046/j.1460-9568.2000.00905.x

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

Brandman, T., Malach, R., & Simony, E. (2021). The surprising role of the default mode network in naturalistic perception. *Communications Biology*, *4*(1), 79. https://doi.org/10.1038/s42003-020-01602-z

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: Inferential processes versus action simulation. *Current Biology*, *17*(24), 2117–2121. https://doi.org/10.1016/j.cub.2007.11.057

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Danek, A. H., Fraps, T., von Müller, A., Grothe, B., & Öllinger, M. (2014). Working wonders? Investigating insight with magic tricks. *Cognition*, *130*(2), 174–185. https://doi.org/10.1016/j.cognition.2013.11.003

Danek, A. H., Öllinger, M., Fraps, T., Grothe, B., & Flanagin, V. L. (2015). An fMRI investigation of expectation violation in magic tricks. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.00084

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience*, *3*(3), 277–283. https://doi.org/10.1038/72991

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2001). The effect of task relevance on the cortical response to changes in visual and auditory stimuli: An event-related fMRI study. *NeuroImage*, *14*(6), 1256–1267. https://doi.org/10.1006/nimg.2001.0946

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, *30*(49), 16601–16608. https://doi.org/10.1523/JNEUROSCI.2770-10.2010

Ekman, M., Kok, P., & De Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, *8*, 1–9. https://doi.org/10.1038/ncomms15276

Erlikhman, G., & Caplovitz, G. P. (2017). Decoding information about dynamically occluded objects in visual cortex. *NeuroImage*, *146*, 778–788. https://doi.org/10.1016/j.neuroimage.2016.09.024

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(34), E5072–E5081. https://doi.org/10.1073/pnas.1610344113

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*(7), 2887–2906. https://doi.org/10.1002/hbm.24047

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. J., Penny, W. D., & Glaser, D. E. (2005). Conjunction revisited. *NeuroImage*, *25*(3), 661–667. https://doi.org/10.1016/j.neuroimage.2005.01.013

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Grassi, P. R., & Bartels, A. (2021). Magic, Bayes and wows: A Bayesian account of magic tricks. *Neuroscience & Biobehavioral Reviews*, *126*, 515–527. https://doi.org/10.1016/j.neubiorev.2021.04.001

Grassi, P. R., Plikat, V., & Wong, H. Y. (2024). How can we be moved by magic? *British Journal of Aesthetics*, *64*(2), 187–204. https://doi.org/10.1093/aesthj/ayad026

Grassi, P. R., Zaretskaya, N., & Bartels, A. (2017). Scene segmentation in early visual cortex during suppression of ventral stream regions. *NeuroImage*, *146*, 71–80. https://doi.org/10.1016/j.neuroimage.2016.11.024

Grassi, P. R., Zaretskaya, N., & Bartels, A. (2018). A generic mechanism for perceptual organization in the parietal cortex. *The Journal of Neuroscience*, *38*(32), 7158–7169. https://doi.org/10.1523/JNEUROSCI.0436-18.2018

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10–11), 1409–1422. https://doi.org/10.1016/S0042-6989(01)00073-6

Grundei, M., Schmidt, T. T., & Blankenburg, F. (2023). A multimodal cortical network of sensory expectation violation revealed by fMRI. *Human Brain Mapping*, *44*(17), 5871–5891. https://doi.org/10.1002/hbm.26482

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hespos, S. J., Ferry, A. L., & Rips, L. J. (2009). Five-month-old infants have different expectations for solids and liquids. *Psychological Science*, *20*(5), 603–611. https://doi.org/10.1111/j.1467-9280.2009.02331.x

Hiser, J., & Koenigs, M. (2018). The multifaceted role of the ventromedial prefrontal cortex in emotion, decision making, social cognition, and psychopathology. *Biological Psychiatry*, *83*(8), 638–647. https://doi.org/10.1016/j.biopsych.2017.10.030

Hohwy, J. (2014). *The predictive mind* (Vol. First edit). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Horstmann, G. (2015). The surprise–attention link: A review. *Annals of the New York Academy of Sciences*, *1339*(1), 106–115. https://doi.org/10.1111/nyas.12679

Hulme, O. J., & Zeki, S. (2007). The sightless view: Neural correlates of occluded objects. *Cerebral Cortex*, *17*(5), 1197–1205. https://doi.org/10.1093/cercor/bhl031

Jääskeläinen, I. P., Pajula, J., Tohka, J., Lee, H.-J., Kuo, W.-J., & Lin, F.-H. (2016). Brain hemodynamic activity during viewing and re-viewing of comedy movies explained by experienced humor. *Scientific Reports*, *6*(1), 27741. https://doi.org/10.1038/srep27741

Jastorff, J., Clavagnier, S., Gergely, G., & Orban, G. A. (2011). Neural mechanisms of understanding rational actions: Middle temporal gyrus activation by contextual violation. *Cerebral Cortex*, *21*(2), 318–329. https://doi.org/10.1093/cercor/bhq098

Jehee, J. F. M., Brady, D. K., & Tong, F. (2011). Attention improves encoding of task-relevant features in the human visual cortex. *The Journal of Neuroscience*, *31*(22), 8210–8219. https://doi.org/10.1523/JNEUROSCI.6153-09.2011

Jiang, J., Summerfield, C., & Egner, T. (2016). Visual prediction error spreads across object features in human visual cortex. *The Journal of Neuroscience*, *36*(50), 12746–12763. https://doi.org/10.1523/JNEUROSCI.1546-16.2016

Kim, H. (2014). Involvement of the dorsal and ventral attention networks in oddball stimulus processing: A meta-analysis. *Human Brain Mapping*, *35*(5), 2265–2284. https://doi.org/10.1002/hbm.22326

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Krekelberg, B., Boynton, G. M., & Van Wezel, R. J. A. (2006). Adaptation: From single cells to BOLD signals. *Trends in Neurosciences*, *29*(5), 250–256. https://doi.org/10.1016/j.tins.2006.02.008

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., & Nummenmaa, L. (2012). Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Frontiers in Human Neuroscience*, *6*, 233. https://doi.org/10.3389/fnhum.2012.00233

Lau, J. K. L., Ozono, H., Kuratomi, K., Komiya, A., & Murayama, K. (2020). Shared striatal activity in decisions to satisfy curiosity and hunger at the risk of electric shocks. *Nature Human Behaviour*, *4*(5), 531–543. https://doi.org/10.1038/s41562-020-0848-3

Lawrence, S. J. D., van Mourik, T., Kok, P., Koopmans, P. J., Norris, D. G., & de Lange, F. P. (2018). Laminar organization of working memory signals in human visual cortex. *Current Biology*, *28*(21), 3435.e4–3440.e4. https://doi.org/10.1016/j.cub.2018.08.043

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434 https://doi.org/10.1364/josaa.20.001434

Linden, D. E. J. (1999). The functional neuroanatomy of target detection: An fMRI study of visual and auditory oddball tasks. *Cerebral Cortex*, *9*(8), 815–823. https://doi.org/10.1093/cercor/9.8.815

Liu, S., Lydic, K., Mei, L., & Saxe, R. (2024). Violations of physical and psychological expectations in the human

adult brain. *Imaging Neuroscience*, *2*, 1–25. https://doi.org/10.1162/imag_a_00068
Marsh, L. E., Mullett, T. L., Ropar, D., & Hamilton, A. F. D. C. (2014). Responses to irrational actions in action observation and mentalising networks of the human brain. *NeuroImage*, *103*, 81–90. https://doi.org/10.1016/j.neuroimage.2014.09.020
Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., Goebel, R., & Yacoub, E. (2015). Contextual feedback to superficial layers of V1. *Current Biology*, *25*(20), 2690–2695. https://doi.org/10.1016/j.cub.2015.08.057
Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058
Notter, M., Gale, D., Herholz, P., Markello, R., Notter-Bielser, M.-L., & Whitaker, K. (2019). AtlasReader: A Python package to generate coordinate tables, region labels, and informative figures from statistical MRI images. *Journal of Open Source Software*, *4*(34), 1257. https://doi.org/10.21105/joss.01257
Olson, I. R., Gatenby, J. C., Leung, H.-C., Skudlarski, P., & Gore, J. C. (2004). Neuronal representation of occluded objects in the human brain. *Neuropsychologia*, *42*(1), 95–104. https://doi.org/10.1016/S0028-3932(03)00151-9
Papale, P., Wang, F., Morgan, A. T., Chen, X., Gilhuis, A., Petro, L. S., Muckli, L., Roelfsema, P. R., & Self, M. W. (2022). Feedback brings scene information to the representation of occluded image regions in area V1 of monkeys and humans. *bioRxiv*. https://doi.org/10.1101/2022.11.21.517305
Parris, B. A., Kuhn, G., Mizon, G. A., Benattayallah, A., & Hodgson, T. L. (2009). Imaging the impossible: An fMRI study of impossible causal relationships in magic tricks. *NeuroImage*, *45*(3), 1033–1039. https://doi.org/10.1016/j.neuroimage.2008.12.036
Pedregosa, F. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490
Puneeth, N. C., & Arun, S. P. (2016). A neural substrate for object permanence in monkey inferotemporal cortex. *Scientific Reports*, *6*(1), 30808. https://doi.org/10.1038/srep30808
Radford, C., & Weston, M. (1975). How can we be moved by the fate of Anna Karenina? *Aristotelian Society Supplementary Volume*, *49*(1), 67–94. https://doi.org/10.1093/aristoteliansupp/49.1.67
Raichle, M. E. (2015). The Brain's default mode network. *Annual Review of Neuroscience*, *38*(1), 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030
Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580
Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, *33*(40), 15978–15988. https://doi.org/10.1523/JNEUROSCI.1580-13.2013
Richter, D., Ekman, M., & de Lange, F. P. (2018). Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *The Journal of Neuroscience*, *38*(34), 7452–7461. https://doi.org/10.1523/JNEUROSCI.3421-17.2018
Richter, D., Kietzmann, T. C., & De Lange, F. P. (2024). High-level visual prediction errors in early visual cortex. *PLoS Biology*, *22*(11), e3002829. https://doi.org/10.1371/journal.pbio.3002829
Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *NeuroImage*, *206*, 116189. https://doi.org/10.1016/j.neuroimage.2019.116189
Saxe, R., & Kanwisher, N. (2003). Role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1
Schiffer, A.-M., & Schubotz, R. I. (2011). Caudate nucleus signals for breaches of expectation in a movement observation paradigm. *Frontiers in Human Neuroscience*, *5*, 38. https://doi.org/10.3389/fnhum.2011.00038
Scholte, H. S., Jolij, J., Fahrenfort, J. J., & Lamme, V. A. F. (2008). Feedforward and recurrent processing in scene segmentation: Electroencephalography and functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *20*(11), 2097–2109. https://doi.org/10.1162/jocn.2008.20142
Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593
Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain. *eLife*, *8*, e46619. https://doi.org/10.7554/eLife.46619
Self, M. W., van Kerkoerle, T., Supèr, H., & Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area V1 in figure-ground segregation. *Current Biology*, *23*(21), 2121–2129. https://doi.org/10.1016/j.cub.2013.09.013
Shultz, S., Lee, S. M., Pelphrey, K., & McCarthy, G. (2011). The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions. *Social Cognitive and Affective Neuroscience*, *6*(5), 602–611. https://doi.org/10.1093/scan/nsq087
Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*(1), 12141. https://doi.org/10.1038/ncomms12141
Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. H. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(4), 1663–1668. https://doi.org/10.1073/pnas.96.4.1663
Stawarczyk, D., Bezdek, M. A., & Zacks, J. M. (2021). Event representations and predictive processing: The role of the midline default network core. *Topics in Cognitive Science*, *13*(1), 164–186. https://doi.org/10.1111/tops.12450
Stefanics, G., Stephan, K. E., & Heinzle, J. (2019). Feature-specific prediction errors for visual mismatch. *NeuroImage*, *196*, 142–151. https://doi.org/10.1016/j.neuroimage.2019.04.020
Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69–82. https://doi.org/10.1016/j.neuroimage.2012.09.063
Stevens, A. A., Skudlarski, P., Gatenby, J. C., & Gore, J. C. (2000). Event-related fMRI of auditory and visual oddball tasks. *Magnetic Resonance Imaging*, *18*(5), 495–502. https://doi.org/10.1016/S0730-725X(00)00128-4
Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: An MEG study. *Journal of Neuroscience*, *31*(25), 9118–9123. https://doi.org/10.1523/JNEUROSCI.1425-11.2011
Tootell, R. B., Hadjikhani, N., Hall, E. K., Marrett, S., Vanduffel, W., Vaughan, J. T., & Dale, A. M. (1998).

The retinotopy of visual spatial attention. *Neuron*, *21*(6), 1409–1422. https://doi.org/10.1016/S0896-6273(00)80659-5

Vander Wyk, B. C., Hudac, C. M., Carter, E. J., Sobel, D. M., & Pelphrey, K. A. (2009). Action understanding in the superior temporal sulcus region. *Psychological Science*, *20*(6), 771–777. https://doi.org/10.1111/j.1467-9280.2009.02359.x

Vizioli, L., Moeller, S., Dowdle, L., Akçakaya, M., De Martino, F., Yacoub, E., & Uğurbil, K. (2021). Lowering the thermal noise barrier in functional brain mapping with magnetic resonance imaging. *Nature Communications*, *12*(1), 5181. https://doi.org/10.1038/s41467-021-25431-8

Wacongne, C., Labyt, E., Van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20754–20759. https://doi.org/10.1073/pnas.1117807108

Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, *25*(10), 3911–3931. https://doi.org/10.1093/cercor/bhu277

Wang, S. (2004). Young infants' reasoning about hidden objects: Evidence from violation-of-expectation tasks with test trials only. *Cognition*, *93*(3), 167–198. https://doi.org/10.1016/j.cognition.2003.09.012

Wessel, J. R., Danielmeier, C., Morton, J. B., & Ullsperger, M. (2012). Surprise and error: Common neuronal architecture for the processing of errors and novelty. *Journal of Neuroscience*, *32*(22), 7528–7537. https://doi.org/10.1523/JNEUROSCI.6352-11.2012

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, *58*(6), 967–973. https://doi.org/10.1016/j.neuron.2008.04.027

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749–750. https://doi.org/10.1038/358749a0

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zollei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*, 1125–1165. https://doi.org/10.1152/jn.00338.2011

Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: Where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, *22*(3), 181–192. https://doi.org/10.1038/s41583-020-00420-w

Zink, C. F., Pagnoni, G., Martin, M. E., Dhamala, M., & Berns, G. S. (2003). Human striatal response to salient nonrewarding stimuli. *The Journal of Neuroscience*, *23*(22), 8092–8097. https://doi.org/10.1523/JNEUROSCI.23-22-08092.2003

Memory color influences conscious object perception

Vincent Plikat[1,2,3], Pablo R. Grassi[1,2,3], Michael M. Bannert[1,2,3], Andreas Bartels[1,2,3]
[1] Vision and Cognition Lab, Department of Psychology, University of Tübingen, Tübingen, Germany
[2] Centre for Integrative Neuroscience, Tübingen, Germany
[3] Max-Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract:
Can knowledge influence perception? A central case suggesting it can, is evidence showing that knowledge about a color-diagnostic object's typical color can influence its appearance. For example, a grey banana is allegedly perceived with a tint of yellow. However, methodological and conceptual considerations, leave it unclear whether the purported "memory-color" effect actually reflects changes in perception or changes in judgment and responses instead. We therefore combine memory-color with binocular rivalry to test if top-down influences affect the color an object is perceived in. We showed 24 participants familiar objects in their typical and opponent color and asked for concurrent reports of the perceived color. Consistent with Bayesian models of rivalry, we observed that conscious perception of identical spectral color pairs was biased towards the typical color of the presented object. Our results suggest that prior knowledge aids interpretation of ambiguous stimuli and biases conscious perception towards the most plausible interpretation.

Keywords: Multistable perception, color vision, memory color, predictive coding, expectation, top-down effects, cognitive penetrability, ambiguous perception

Research Transparency Statement
Conflicts of interest: The authors declare no conflicts of interest. Funding: This work was supported by the Barbara-Wengeler Foundation, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): project number 465409366 (to PRG) and SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, project 9, project number: 276693517. Artificial intelligence: No artificial intelligence was used to generate stimuli, text or figures. Ethics: The study was approved by the ethics committee of the Department of Psychology at the University of Tübingen. Preregistration: The experiments were not preregistered. Materials: Experimental code will be made publicly available upon publication. Data: all raw data will be made publicly available upon publication. Analysis scripts: all analysis scripts will be made publicly available upon publication.

Introduction

Knowledge is said to directly influence perceptual processes (de Lange et al., 2018). For example, when sensory input is noisy and ambiguous, memory-based expectations can be used to aid interpretation and facilitate processing (e.g., when using context to make sense of a spoken word in a noisy environment). Conversely, when sensory input is salient or deviates from expectations, selective attention boosts its processing, prioritizing the unexpected information for the refinement of future predictions.

A prominent example of these influences of cognition in perception is the *memory-color effect* (MCE), where prior knowledge about an object's typical color may influence perception (Hering, 1920). Through experience, we learn that certain objects – termed color-diagnostic – generally have a particular typical color: grass is green, and bananas are yellow. These object-color associations can bias perception of grayscale objects towards their associated color (Hansen et al., 2006; Olkkonen et al., 2008; Witzel et al., 2011; Kimura et al., 2013). In turn, color-diagnostic objects in their incongruent color (e.g., a blue banana) are unexpected and capture attention, e.g., enhancing performance in tasks like change detection (Cutler et al., 2024) or interfering with object recognition (Teichmann et al., 2020). However, the role that memory color plays in determining what color enters conscious perception in the first place is unclear. This is particularly relevant given that color vision is, by its very nature, subjective and therefore must be studied with tools that dissociate subjective experience from sensory input (Kim et al., 2020).

Moreover, the significance of the MCE extends beyond the field of color vision. It is often regarded as an influential paradigm of cognitive penetrability of perception, i.e., of top-down effects on perception in general. However, it has recently been challenged that the MCE reflects a change in *perception* itself, arguing instead that it influences only (accompanying) cognitive processes like *attention, judgment,* or *recognition* (Firestone and Scholl, 2016; Valenti and Firestone, 2019; Block, 2023), but not how stimuli perceptually appear per se. By this account, when participants adjust a banana's color to make it appear gray and overshoot towards blue chromaticities (Hansen et al., 2006), it is not because they perceive the grey banana with a tint of yellow. Instead, prior knowledge of the banana's typical yellow color causes subjects to only bias the judgement of their unaffected percepts towards the opponent color (i.e., blue), without changes in visual appearance per se (Valenti and Firestone, 2019). As it happens, participants are perfectly capable of telling that, for example, an objectively gray banana is the same color as a gray patch and not a yellow patch (Valenti and Firestone, 2019). This suggests that the MCE might not be perceptual after all.

To address both questions, we combined MCE with binocular rivalry (BR) and tested if memory-based color associations could modulate spontaneous and automatic *conscious perception*. In BR, different images are presented to the two eyes, leading to spontaneous and effortless perceptual alternations despite constant stimulation, allowing to dissociate sensory input from conscious perception (Blake and Logothetis, 2002). We presented participants color-diagnostic objects (e.g., a banana) in a congruent color to one eye (yellow) and in an incongruent color (blue) to the other and asked them to report what they concurrently perceive.

We employed BR to: first, isolate the effect of memory-based color associations on spontaneous and subjective color perception, independent of the spectral properties

of the stimulus input and second, to minimize the confounding effect of attention or judgment on the MCE. We hypothesized that a color's probability to enter visual consciousness would differ depending on whether or not it is congruent or incongruent with object-based expectation. We were neutral with respect to the direction of this difference: on the one hand, previous work found dominance of expected input (Denison et al., 2011; Attarha and Moore, 2015) typically interpreted in a Bayesian predictive coding framework as reflecting the higher prior probability of the expectation-based inputs (Hohwy et al., 2008; Brascamp et al., 2018). However, on the other hand further work found dominance of the unexpected input (Mudrik et al., 2011) potentially due to attentional capture and preferential processing. To elucidate which direction would apply to memory color was the third aim of our study.

Material and Methods

Participants

We measured 24 subjects (10 male, 13 female, one other, 22 right-handed, age 23.2 $\pm$ 3.8) this allowed for counterbalanced conditions and the detections of effect sizes of d = 0.6 with a power of $1 - \beta = 0.8$ at $\alpha = 0.05$ for paired t-tests. All subjects had normal or corrected-to-normal vision, and no color vision impairment as assessed with the Ishihara test procedure. Ten of the subjects were not naive to the research question at hand, one of the authors (VP) and nine cognitive science Bachelor students, for which the experiment was part of a university seminar. All subjects provided written consent prior to the experiments. The study was approved by the ethics committee of the Department of Psychology at the University of Tübingen.

Apparatus

Stimuli were presented on a linearized monitor at a distance of 70 cm using MATLAB R2023b with Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007). Subjects viewed the stimuli through a mirror stereoscope and provided their responses using a dedicated keyboard with three keys.

Experimental rationale, design and procedure

In this study, we tested the influence of prior knowledge about the typical color of common (color-diagnostic objects) on conscious color perception during BR. To this end, we created a BR experiment in which we simultaneously presented color-diagnostic objects in their congruent color (e.g.., a yellow banana) to one eye and in an incongruent color (e.g., a blue banana) to the other eye. The diagnostic colors selected from the perceptually uniform CIELab color space were red, green, yellow and blue, the four cardinal colors. The two pairs of opposing colors, red–green and yellow–blue, were always presented together during rivalry. Importantly, to control low-level chromatic properties, we used the same hues when presenting a color on a congruent or incongruent object, e.g., the congruent red on the tomato was the same as the incongruent red on the lettuce, and vice versa for the paired green hue (see section *Stimuli* for details).

Moreover, to test for general color biases in BR, we conducted a control experiment showing abstract gratings with the same colors from the colored objects of the color-diagnostic experiment ("gratings control"). This served as a baseline estimate for the conscious perception of the same color pairs in the complete absence of object information. To further control for object-color interactions, we performed a second control experiment with objects that do *not* have a color association, such as mugs ("non-diagnostic control").

The study was separated into two sessions. In the first session, subjects would perform either the main color-diagnostic experiment or the non-diagnostic control experiment. In the second session they would perform the gratings control experiment and the experiment that was not performed in the first session. This was counterbalanced over subjects. To familiarize participants to BR and ensure that subjects experienced perceptual alternations, they were shown gray images of basketballs and pumpkins for 120 s (not used in any of the later experiments). Moreover, participants performed a heterochromatic flicker task prior to each experiment to match the stimuli for subjective luminance for each stimulus separately.

Each experiment was divided into two parts: a sustained-rivalry and an onset rivalry part. The sustained-rivalry part consisted of multiple runs (Memory-color experiment: six runs, four trials each; non-diagnostic control: four runs, six trials each; gratings control: two runs, four trials each), where each trial lasted 90 s with an intertrial interval (ITI) of 10 s. The onset rivalry part consisted of a single run with several trials (Memory-color experiment and non-diagnostic control: 96 trials; gratings control: 64 trials), where each trial lasted for at most 5 s with an ITI of 1 s. In the onset rivalry part, a trial ended as soon as the subject reported a dominant percept. At the end of each experiment subjects were asked to fill out a questionnaire which asked for a subjective rating of their conscious perception during rivalry.

In the memory-color experiment, subjects were instructed to press one key during perception of the object in its congruent color and another key during perception of the object in its incongruent color. In the control experiments, subjects were instructed to press one key if they saw the object in red or blue and another in green or yellow. During piecemeal and mixed percepts, subjects were instructed not to press any key. Key assignments were counterbalanced across subjects.

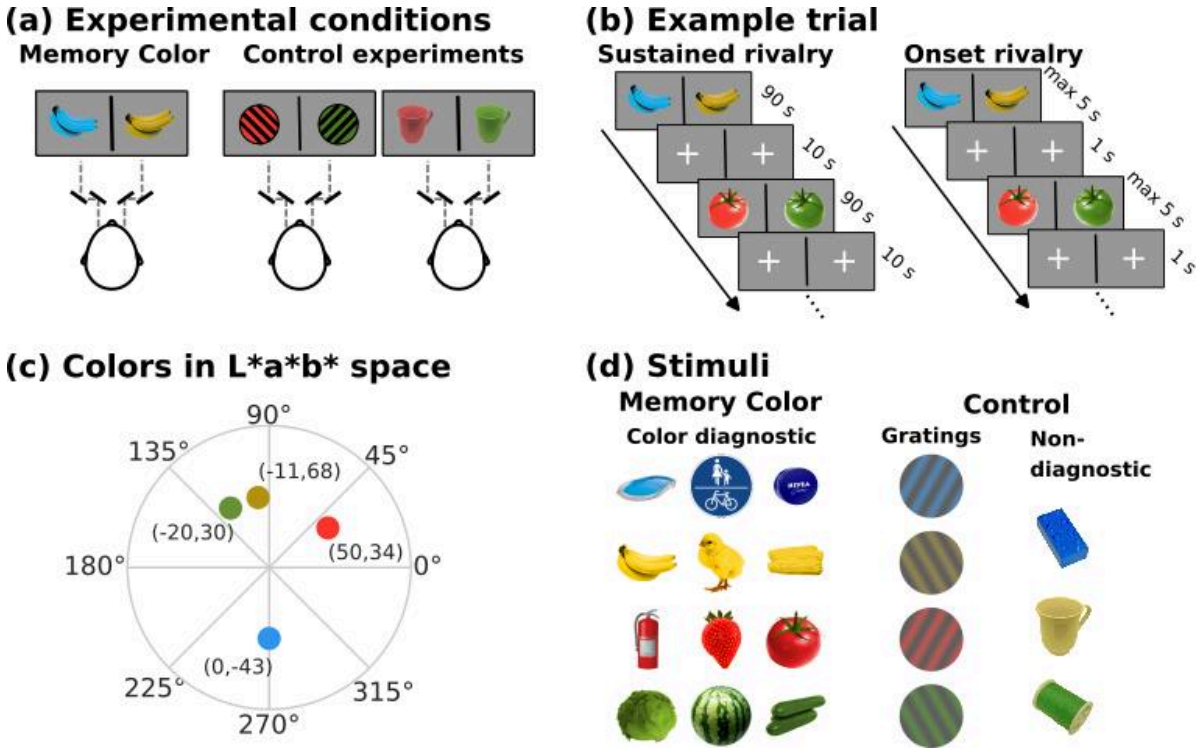


Fig. 1. Experimental design. (a) Overview of the experimental BR conditions. The memory color experiment presented diagnostic objects in their congruent and incongruent color (such as a yellow and blue banana) to each eye separately. In the control experiments, we used the same colors as in the memory-color experiments but on neutral stimuli, such as colored gratings (first control) or non-diagnostic objects (second control). **(b)** Example trials of the sustained-rivalry (left) and onset rivalry (right) part of the memory-color experiment. During sustained rivalry, color-diagnostic objects were shown in congruent color to one eye and incongruent color to the other eye for 90 s with 10 s inter-trial interval. During onset rivalry, the same objects were shown either until subjects reported a percept or 5 s passed. **(c)** Mean hue angles for each color used in L*a*b* space with their corresponding a* and b* values shown in parentheses. **(d)** Stimuli used in each BR experiments. For a detailed description of how these stimuli were preprocessed for the experiments see also Supplementary Section “Stimulus generation”.

Stimuli generation and display

Memory-color experiment

In our main experiment we used a set of twelve color-diagnostic objects, three per cardinal color. All objects were taken from Teichmann et al., (2020), except for a green watermelon and all blue objects that are public domain images taken from the internet (see Fig. 1d for all objects). The red, green and yellow stimuli used have been shown to invoke an MCE in a recognition task (Teichmann et al., 2020). Stimuli were created such that confounding factors due to low-level visual differences, such as luminance or contrast, were minimized. In short, we extracted one hue for each of the four colors of the selected objects. We then projected these hues on equiluminant grayscale versions of the object images in order to create congruently and incongruently colored objects (see Supplementary section "Stimulus generation" and Fig. S1 in the supplementary material for details). This procedure also ensured that incongruent hues on some objects corresponded exactly to congruent hues of other objects.

The stimuli (1.5 - 2 visual degrees in size) were presented on a gray square of 3 visual degrees inside a Mondrian mask of 10 visual degrees (to aid binocular fusion). To increase mutually exclusive rivalry between images, individual stimuli were either rotated 10° to one another or flipped along the vertical axis (in case of vertical stimuli, such as a fire extinguisher).

Gratings control experiment

In a first control experiment we tested for general color biases using colored gratings using the hues selected for the memory-color experiment. Gratings were shown rotated by +30 degrees on the one and -30 degrees on the other eye. Stimulus size was 2.5 visual degrees, and the spatial frequency was 1.6 cycles/visual degree.

Non-color-diagnostic control experiment

In a second control experiment we used objects with no color association (Brodeur et al., 2010) – a mug, a lego block and a thread spool – to control for object-color interactions independent of memory-color. Control stimuli were created the same way as the stimuli from our memory-color experiment, with the exception that we used the hues from the memory-color experiment to color equiluminant grayscaled images of non-diagnostic objects. The three objects were presented in all colors and had the same dimensions as the objects in the memory-color experiment.

Post-experiment questionnaire

After each experiment subjects were asked to fill out a short questionnaire. Each questionnaire asked for a subjective assessment of perceptual dominance during the sustained-rivalry and onset rivalry part. In the second control experiment using non-color-diagnostic objects subjects were additionally asked to indicate whether they associate any color with the objects used. As expected, most participants did not report any color associations for the three objects used in the second control experiment (Thread spool: 86%, Mug: 90%, Lego block: 62%). Please note that we have questionnaire data from 20 subjects for all questions and that one subject failed to answer one of the questions (“Did you have the impression you perceived more time the diagnostic color or the non-diagnostic color?“).

Data processing

For all data analyses, we removed mixed percepts from the sustained-rivalry data as well as all missed trials from the onset rivalry data (i.e., trials in which subjects did not press any button within the time window of 5 s). Further, we assessed if data from any subject was unsuitable for data analysis. Exclusion criteria were an eye dominance of more than 70%, a median dominance duration shorter than 0.5 s and mixed percept durations of more than 70%. No subject had to be excluded. Finally, as the initial phase of rivalry is more susceptible to attentional biases (Mitchell et al., 2004; Chong et al., 2005; Alpers and Gerdes, 2007), and might differ from sustained rivalry (Carter and Cavanagh, 2007), we removed the first percept for the assessment of predominance ratio and median dominance duration. Please note that in this process some trials were excluded entirely, since for some subjects there were no switches in dominance at all in some trials (eight trials in four subjects, mostly perception of congruent red or yellow – see supplementary

table S1 for details). Analyses of the data with or without the first percept did not affect the significance of the results.

Data analysis

To investigate the MCE on conscious perception during BR, we extracted three measures from our data: *onset dominance, predominance and median dominance duration*.

Onset dominance ratio: First, we were interested in the MCE on the very first percept after rivalry onset. For each color pair we selected an arbitrary reference color: red as a reference color for the red-green and yellow as a reference color for the yellow-blue color pairs. We then calculated the onset dominance ratio (ODR) as the following ratio:

$$ODR = (N_r - N_o) / (N_r + N_o)$$

where N_r and N_o are the number of times reference and other color were reported as first percepts, respectively.

Predominance ratio: Further, we were interested in how memory color affects the dominance *duration* of rivaling images. This measure thus quantifies the sustained biasing effect of object knowledge on conscious color perception. The predominance ratio (PR) was calculated as:

$$PR = (d_r - d_o) / (d_r + d_o)$$

where d_r and d_o are total durations of reference or other color across all trials.

Median dominance duration: Finally, the predominance ratio does not provide any information about how long each percept lasts. Hence, we compared median dominance durations (MDD) for congruent and incongruent colors within objects (e.g., dominance duration for a yellow banana vs a blue banana), as well as across objects (e.g., dominance for a yellow banana vs a yellow nivea tin). We further calculated the difference in MDD for each color-pair (i.e., red-green and yellow-blue), separately per condition.

Statistical inference

To assess the influence of an MCE in the main memory-color experiment, we compared each of the measures (ODR, PR and MDD difference) where the reference color was the *congruent* color compared to where the reference color was the *incongruent* color, resulting in a 2 x 2 repeated-measures ANOVA (rmANOVA) with congruency (congruent/incongruent) and reference color (red/yellow) as factors. Please note that this approach controls for potential color biases because objects were presented in the same color pairs irrespective of whether the positive or the negative color was its associated memory color. Differences in MDD within and across objects were tested using paired t-tests. To test for general color biases in context of non-diagnostic objects and gratings, we performed one sample t-tests on each of the measures using data from the control experiments using colored gratings and non-diagnostic objects. As these stimuli have no congruency, we only tested measures against zero. All t-tests were two-sided.

Results

Color-diagnostic objects

Onset dominance ratio

First, we tested whether memory-based object-color associations influence the first percept at rivalry onset. The rmANOVA of the ODR revealed a significant main effect for congruency, $F(1,23) = 16.14$, $p = .001$, $eta^2 = 0.097$, and a significant interaction of congruency and reference color, $F(1,23) = 13.5$, $p = .001$, $eta^2 = 0.041$, but no significant main effect for the reference color, $F(1,23) = 0.19$, $p = .669$, $eta^2 = 0.002$. Post-hoc t-tests showed that the congruent color was reported more frequently, $t(23) = 4.02$, $p = .001$, Cohen's $d_{avg} = 0.8$, 95% CI [0.13, 0.4]. This main effect of congruency was driven by the yellow-blue color pairs, $t(23) = 5.24$, $p < .001$, Cohen's $d_{avg} = 1$, 95% CI [0.26, 0.59], but not the red-green pairs, $t(23) = 1.28$, $p = .213$, Cohen's $d_{avg} = 0.25$, 95% CI [-0.06, 0.26] (see Fig. 2a).

Predominance ratio

The rmANOVA of the PR during sustained rivalry revealed a significant main effect for congruency, $F(1,23) = 9.61$, $p = .005$, $eta^2 = 0.091$, but no significant main effect for reference color, $F(1,23) = 0.95$, $p = .34$, $eta^2 = 0.013$, nor a significant interaction, $F(1,23) = 0.17$, $p = .688$, $eta^2 = 0.001$. A post-hoc t-test revealed that objects were perceived longer in their congruent color than in their incongruent color, $t(23) = 3.1$, $p = .005$, Cohen's $d_{avg} = 0.88$, 95% CI [0.05, 0.25] (see Fig. 2b).

Medium Dominance Duration Difference

Analyses of ODR and PR revealed that objects were perceived more often in their congruent color at rivalry onset and that objects were perceived for a longer time during sustained rivalry. Similarly, MDD difference values differed with respect to congruency, $F(1,23) = 6.6$, $p = .017$, $eta^2 = 0.071$. There was no significant main effect for the reference color, $F(1,23) = 2.19$, $p = .152$, $eta^2 = 0.025$, nor for the interaction $F(1,23) = 0.09$, $p = .764$, $eta^2 = 0.001$. A post hoc t-test showed that dominance durations for objects in their congruent color were longer than in their incongruent color, $t(23) = 2.57$, $p = .017$, Cohen's $d_{avg} = 0.74$, 95% CI [0.09, 0.8] (see Fig. 2c). Comparison of MDD within objects (e.g., yellow vs blue banana) revealed a significant difference for objects associated with red only, $t(23) = 2.57$, $p = .017$, Cohen's $d_{avg} = 0.39$, 95% CI [0.05, 0.45] (see Fig. 2d). Detailed results for post-hoc comparisons can be found in supplementary Table S2.

Together, we find effects of congruency across all three measures of binocular rivalry that we analyzed.

Control experiments

We further examined the same BR metrics in the absence of object knowledge but using identical color stimuli. In a first control experiment we used colored gratings. Here, we observed that perceptual report of red gratings was significantly more frequent in all measures compared to green gratings (ODR: $t(23) = 9.54$, $p < .001$, Cohen's $d_{avg} = 1.95$; PR: $t(23) = 4.14$, $p < .001$, Cohen's $d_{avg} = 0.84$; MDD $t(23) = 4.41$, $p < .001$, Cohen's d =

0.9). In contrast, we did not observe any significant bias towards blue or yellow gratings (ODR: t(23) = 1.2, p = .242, Cohen's d_{avg} = 0.24; PR: t(23) = 1.05, p = .305, Cohen's d_{avg} = 0.21; MDD: t(23) = 1.52, p = .143, Cohen's d_{avg} = 0.31, see supplementary Fig. S2). Analyses of a second control experiment using non-diagnostic objects yielded qualitatively similar results. For details see supplementary results section *non-diagnostic objects* and supplementary Fig. S2. Hence, both control experiments revealed a general bias towards red perception. Note that this bias cannot explain the preference for congruent colors observed in the memory-color experiment, which was mainly driven by yellow and green objects. If anything, the color bias would have worked against the main findings.

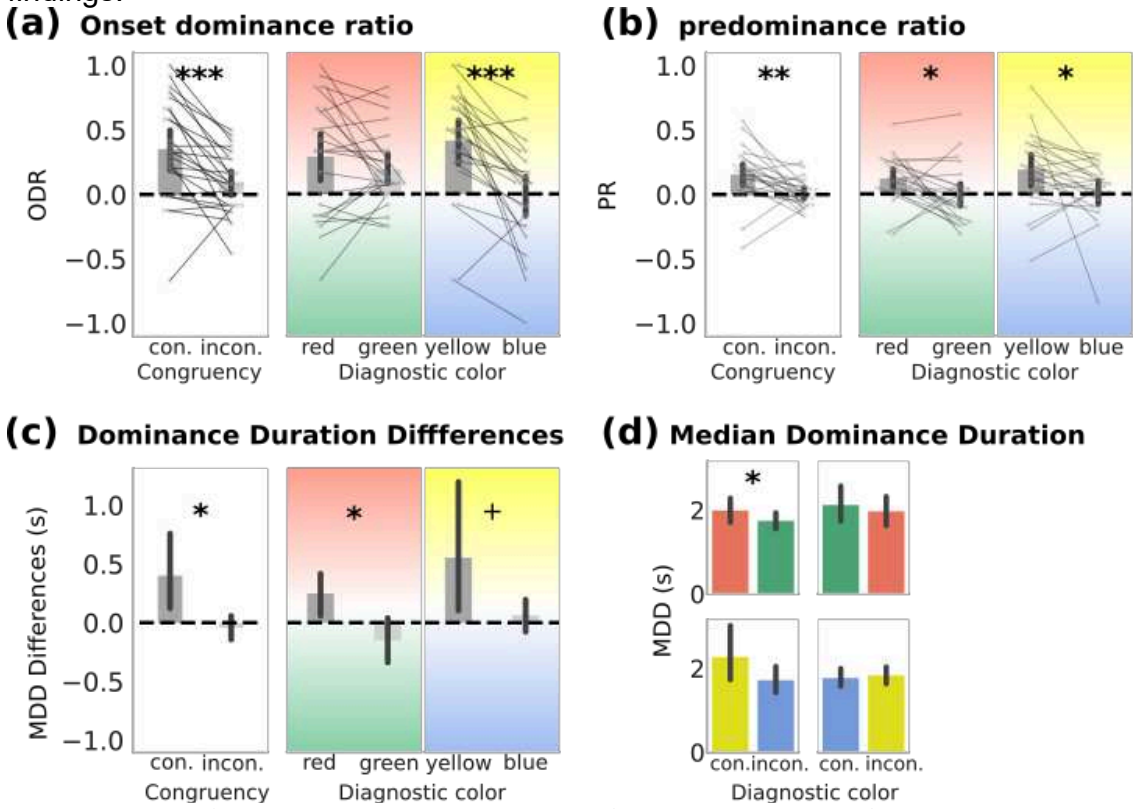


Fig. 2. Results of the memory-color experiment. Shown are the results from paired comparisons of **(a)** onset dominance ratio (ODR), **(b)** predominance ratio (PR) and **(c)** difference in median dominance duration (MDD) averaged over all colors (left panels) and separately for each color pair (middle and right panels). **(d)** MDDs compared within objects (e.g., dominance for a red strawberry vs a green strawberry). All error bars depict 95% confidence intervals. + p < .1; *, p < .05; **, p < .01; ***, p < .001.

Post-experiment questionnaire

Results of the post-experimental questionnaire reflected behavioral responses in the experiment. Subjects reported to have perceived objects more in their diagnostic color, t(18) = 2.96, p = .008, Cohen's d_{avg} = 0.68. This was the case especially for red, t(19) = 4.47, p < .001, Cohen's d_{avg} = 1, and yellow objects, t(19) = 3.48, p = .002, Cohen's d_{avg} = 0.76. The ratings further reflected the general red bias found in the behavioral data with a strong bias towards red in both control experiments (gratings: t(19) = 3.24, p = .004, Cohen's d_{avg} = 0.72; non-diagnostic objects: t(19) = 11.38, p < .001, Cohen's d_{avg} = 2.54).

No bias towards yellow or blue was found in neither control (grating: $t(19) = 1.28$, $p = .217$, Cohen's $d_{avg} = 0.29$; non-diagnostic objects: $t(19) = 0.43$, $p = .673$, Cohen's $d_{avg} = 0.1$).

Discussion

In this study we tested whether memory-based object-color associations can influence conscious perception during BR. To this end, we presented dichoptically color-diagnostic objects in their associated color to one eye and in an incongruent color to the other eye. We observed an MCE in all measures such that an object's typical color determined which of two rivaling colors entered visual consciousness and for how long it remained dominant.

Our results suggest that memory-based associations affect spontaneous and automatic *conscious perception*. We observed that during ambiguous visual input a congruent object-color association is favored relative to an incongruent association in conscious perception. Our control experiments show that the results cannot be explained by low-level features or general color biases.

But is *memory* actually affecting *perception*?

It has recently been argued that the visual system is "encapsulated" from higher-level cognition, and that (all) prior studies that report cases of top-down effects in perception – including memory color – can actually be explained by (or are susceptible to) a series of methodological and conceptual pitfalls (Firestone and Scholl, 2016). The pitfalls often stem from conflating effects on perception with effects on *judgement* or *attentional processes*. For example, it has been suggested that memory color influences *judgments* and *responses* about color, but not the actual visual *appearance* (Valenti and Firestone, 2019) – particularly in studies in which participants are asked to adjust the color of color-diagnostic objects to gray (as in Hansen et al., 2006; Olkkonen et al., 2008; Witzel et al., 2011). In contrast, our experiment used BR, where participants simply reported the color that spontaneously and automatically entered their consciousness and is thus less susceptible to explanations based on post-perceptual inferences. Unlike adjustment or comparison tasks, BR offers a more direct measure of perceptual awareness, reducing the scope for judgment-based confounds. Moreover, while memory color could involuntarily bias attention toward congruent colors, incongruent objects are known to capture attention (Cutler et al., 2024), and attended stimuli typically dominate in BR (Dieter and Tadin, 2011; Paffen and Alais, 2011) – in particular the first percept (Ooi and He, 1999; Mitchell et al., 2004; Chong et al., 2005; Alpers and Gerdes, 2007). Yet, we observed a clear preference for congruent colors relative to incongruent colors in all measures for sustained and onset rivalry, suggesting that memory shapes perception directly rather than indirectly via attention. Given that our MCE resists explanation by judgment or attention, we conclude that our results provide evidence of a top-down (memory) effect in perception, challenging the notion of an encapsulated visual system, unless one includes color-object associations as part of the visual system (cf. Block 2023).

Enhancement of expected stimuli and Bayesian inference

The observed MCE means that colors are more likely to enter visual consciousness if they match object-based expectations compared to when they do not. This observation is consistent with several studies that report a preference in BR for

congruent and expected stimuli, such as upright compared to inverted faces (Engel, 1956), predicted motion (Denison et al., 2011; Attarha and Moore, 2015), known and newly learned cross-modal associations (Conrad et al., 2010, 2013; Chen et al., 2011; Lunghi et al., 2014; Einhäuser et al., 2017; Piazza et al., 2018), naturalistic stimuli (Baker and Graf, 2009), familiar stimuli (Yu and Blake, 1992) and percepts matching preceding mental imagery (Pearson et al., 2008). Beyond rivalry, this enhancement of congruent or expected stimuli also extends to other ambiguous paradigms, such as bistable structure-from-motion stimuli (cf. Gilroy and Blake, 2004; Sterzer et al., 2008).

These convergent findings support the notion of perception as a Bayesian inferential process, where prior information (such as newly learned associations or visual history) is combined with sensory information to determine the most probable cause of the sensory inputs (Lee and Mumford, 2003; Kersten et al., 2004; Friston, 2005). In our study, objects in their congruent colors likely provide higher prior probabilities compared to incongruent colors, biasing perception in their favor. This aligns well with Bayesian models of BR (Hohwy et al., 2008) and perception in general (Clark, 2013; Hohwy, 2014; de Lange et al., 2018), suggesting that the visual system resolves ambiguity by favoring interpretations consistent with expectations.

However, this presents a paradox: novel and surprising stimuli are theoretically more informative for the visual system and convey new information for learning (Press et al., 2020), which is likely why color-diagnostic objects in incongruent colors capture attention (Cutler et al., 2024). As it happens, some BR experiments report enhanced dominance for incongruent objects in a scene (Mudrik et al., 2011) and unexpected images in a sequence (Denison et al., 2016), seemingly at odds with our findings and the Bayesian preference for priors. As novel and surprising stimuli are known to capture attention (Theeuwes, 1992, 2010; Itti and Baldi, 2005; Brockmole and Henderson, 2008; Underwood et al., 2008; LaPointe and Milliken, 2016), these conflicting results could be related to differences in attentional capture and allocation of attentional resources between experiments (Dieter and Tadin, 2011). Moreover, Mudrik et al. (2011) and Denison et al. (2016) used complex naturalistic stimuli, in contrast to the simpler stimuli used in most studies showing an advantage in perception for the expected stimuli (e.g., Pearson et al., 2008; Denison et al., 2011; Attarha and Moore, 2015). Thus, the direction of perceptual benefit – favoring expectation or novelty – may additionally depend on task demands, stimulus complexity and processing stage. For example, Mudrik et al., (2011) observed a preference for the unexpected/incongruent percept when scene perception aided ambiguous object perception (favoring novelty), whereas we observe a preference for expected/congruent percept when object perception aided ambiguous feature perception (favoring expectation). These differences are likely related to where exactly BR is resolved in the brain (see also Lawler and Silver, 2023), and how predictive signals of different complexity interact with attentional processes therein (Dieter and Tadin, 2011).

How memory color biases binocular rivalry

In BR, competition can occur at multiple levels of the visual pathway (Blake and Logothetis, 2002; Tong et al., 2006). For example, neuroimaging shows eye-specific rivalry in V1 (Haynes et al., 2005; Qian et al., 2023) and in ventral color (Kim et al., 2020) and face and place areas (Tong et al., 1998). Moreover, neuroimaging (Zaretskaya et al., 2010; Weilnhammer et al., 2013, 2021), electrophysiology (Panagiotaropoulos et al.,

2012; Kapoor et al., 2022) and TMS experiments (Carmel et al., 2010; Kanai et al., 2010; Zaretskaya et al., 2010; Weilnhammer et al., 2021; Schauer et al., 2024) have presented compelling evidence for the role of frontal and parietal areas in resolving and modulation perceptual ambiguity (see Brascamp et al., 2018 for a review). Rivalry activity in visual areas may thus reflect modulatory feedback signals from higher frontal and parietal areas (Maier et al., 2008; Grassi et al., 2016a; De Jong et al., 2020; Qian et al., 2023). This is compatible with "predictive coding" models implementing Bayesian inference, which see top-down memory-based predictive signals being fed-back to lower-level sensory cortices to be compared with bottom-up input (Mumford, 1992; Rao and Ballard, 1999; Lee and Mumford, 2003; Friston, 2005) and neuroimaging evidence using ambiguous stimuli (Murray et al., 2002; De-Wit et al., 2012; Zaretskaya et al., 2013; Grassi et al., 2016b, 2017, 2018).

Our behavioral observation of a memory-color effect enhancing input based on prior knowledge in BR is in line with existing evidence suggesting top-down modulation of rivalry signals in visual cortices. Moreover, our results are supported by further neuroimaging reports of neural information in visual areas signaling memory-color (Bannert and Bartels, 2013; Vandenbroucke et al., 2016) and surprise based on unexpected naturalistic color changes (i.e., color changing magic tricks; Plikat et al., 2025). Hence, our results show that in cases of ambiguous visual stimulation, prior knowledge influences which stimulus enters conscious perception, consistent with the existence of top-down signals in the visual cortex.

Conclusion

Here, we combined memory-color with BR to show that when confronted with ambiguous color stimuli, prior knowledge about an object's typical color influences conscious perception. We observed a significant influence of memory-color in BR, favoring expected colors when viewing color-diagnostic objects, consistent with Bayesian models of BR. These results cannot be explained by low-level differences or general color biases, and are difficult to attribute to residual differences in attention or judgement. Instead, we argue that the effects reflect a genuine top-down effect of knowledge in perception.

Acknowledgment

We thank the participants of the Experimental cognitive science course at the University of Tübingen (winter term 2023/2024) for their help in conducting this experiment.

References

Alpers GW, Gerdes ABM (2007) Here is looking at you: Emotional faces predominate in binocular rivalry. Emotion 7:495–506.

Attarha M, Moore CM (2015) Onset rivalry: factors that succeed and fail to bias selection. Atten Percept Psychophys 77:520–535.

Baker DH, Graf EW (2009) Natural images dominate in binocular rivalry. Proc Natl Acad Sci USA 106:5436–5441.

Bannert MM, Bartels A (2013) Decoding the yellow of a gray banana. Current Biology 23:2268–2272.

Blake R, Logothetis NK (2002) Visual competition. Nature reviews Neuroscience 3:13–21.

Block N (2023) The Border Between Seeing and Thinking, 1st ed. Oxford University PressNew York.

Brainard DH (1997) The Psychophysics Toolbox. Spatial Vision 10:433–436.

Brascamp J, Sterzer P, Blake R, Knapen T (2018) Multistable Perception and the Role of Frontoparietal Cortex in Perceptual Inference. Annual Review of Psychology 69:77–103.

Brockmole JR, Henderson JM (2008) Prioritizing new objects for eye fixation in real-world scenes: Effects of object–scene consistency. Visual Cognition 16:375–390.

Carmel D, Walsh V, Lavie N, Rees G (2010) Right parietal TMS shortens dominance durations in binocular rivalry. Current biology 20:R799-800.

Carter O, Cavanagh P (2007) Onset Rivalry: Brief Presentation Isolates an Early Independent Phase of Perceptual Competition He S, ed. PLoS ONE 2:e343.

Chen Y-C, Yeh S-L, Spence C (2011) Crossmodal Semantic Constraints on Visual Perception of Binocular Rivalry. i-Perception 2:900–900.

Chong SC, Tadin D, Blake R (2005) Endogenous attention prolongs dominance durations in binocular rivalry. Journal of Vision 5:6.

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36:181–204.

Conrad V, Bartels A, Noppeney U (2010) Audiovisual interactions in binocular rivalry. 10:1–15.

Conrad V, Kleiner M, Bartels A, Hartcher O'Brien J, Bülthoff HH, Noppeney U (2013) Naturalistic Stimulus Structure Determines the Integration of Audiovisual Looming Signals in Binocular Rivalry Geng JJ, ed. PLoS ONE 8:e70710.

Cutler A, Rivest J, Cavanagh P (2024) The role of memory color in visual attention. Atten Percept Psychophys 86:28–35.

De Jong MC, Vansteensel MJ, van Ee R, Leijten FSS, Ramsey NF, Dijkerman HC, Dumoulin SO, Knapen T (2020) Intracranial Recordings Reveal Unique Shape and Timing of Responses in Human Visual Cortex during Illusory Visual Events. Current Biology 30:3089-3100.e4.

de Lange FP, Heilbron M, Kok P (2018) How Do Expectations Shape Perception? Trends in Cognitive Sciences 22:764–779.

Denison RN, Piazza EA, Silver MA (2011) Predictive Context Influences Perceptual Selection during Binocular Rivalry. Front Hum Neurosci 5.

Denison RN, Sheynin J, Silver MA (2016) Perceptual suppression of predicted natural images. Journal of Vision 16:6.

De-Wit LH, Kubilius J, Wagemans J, de Beeck HPO (2012) Bistable Gestalts reduce activity in the whole of V1, not just the retinotopically predicted parts. Journal of vision 12:1–14.

Dieter KC, Tadin D (2011) Understanding Attentional Modulation of Binocular Rivalry: A Framework Based on Biased Competition. Front Hum Neurosci 5.

Einhäuser W, Methfessel P, Bendixen A (2017) Newly acquired audio-visual associations bias perception in binocular rivalry. Vision Research 133:121–129.

Engel E (1956) The Role of Content in Binocular Resolution. The American Journal of Psychology 69:87.

Firestone C, Scholl BJ (2016) Cognition does not affect perception: Evaluating the evidence for "top-down" effects. Behav Brain Sci 39:e229.

Friston K (2005) A theory of cortical responses. Philosophical transactions of the Royal Society of London Series B, Biological sciences 360:815–836.

Gilroy LA, Blake R (2004) Physics embedded in visual perception of three-dimensional shape from motion. Nat Neurosci 7:921–922.

Grassi PR, Schauer G, Dwarakanath A (2016a) The role of the occipital cortex in resolving perceptual ambiguity. Journal of Neuroscience 36:10508–10509.

Grassi PR, Zaretskaya N, Bartels A (2016b) Parietal cortex mediates perceptual Gestalt grouping independent of stimulus size. NeuroImage 133:367–377.

Grassi PR, Zaretskaya N, Bartels A (2017) Scene segmentation in early visual cortex during suppression of ventral stream regions. NeuroImage 146:71–80.

Grassi PR, Zaretskaya N, Bartels A (2018) A Generic Mechanism for Perceptual Organization in the Parietal Cortex. The Journal of Neuroscience 38:7158–7169.

Hansen T, Olkkonen M, Walter S, Gegenfurtner KR (2006) Memory modulates color appearance. Nat Neurosci 9:1367–1368.

Haynes J-D, Deichmann R, Rees G (2005) Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. Nature 438:496–499.

Hering E (1920) Grundzüge der Lehre Vom Lichtsinn. Berlin, Heidelberg: Springer Berlin / Heidelberg.

Hohwy J (2014) The Predictive Mind. Oxford: OUP Oxford.

Hohwy J, Roepstorff A, Friston K (2008) Predictive coding explains binocular rivalry: an epistemological review. Cognition 108:687–701.

Itti L, Baldi P (2005) Bayesian surprise attracts human attention. Advances in Neural Information Processing Systems:547–554.

Kanai R, Bahrami B, Rees G (2010) Human parietal cortex structure predicts individual differences in perceptual rivalry. Current Biology 20:1626–1630.

Kapoor V, Dwarakanath A, Safavi S, Werner J, Besserve M, Panagiotaropoulos TI, Logothetis NK (2022) Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. Nat Commun 13:1535.

Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annual review of psychology 55:271–304.

Kim I, Hong SW, Shevell SK, Shim WM (2020) Neural representations of perceptual color experience in the human ventral visual pathway. Proc Natl Acad Sci USA 117:13145–13150.

Kimura A, Wada Y, Masuda T, Goto S, Tsuzuki D, Hibino H, Cai D, Dan I (2013) Memory Color Effect Induced by Familiarity of Brand Logos Bruce A, ed. PLoS ONE 8:e68474.

Kleiner M, Brainar, David, Pelli, Denis (2007) Kleiner, Mario, David Brainard, and Denis Pelli. “What’s new in Psychtoolbox-3?.” (2007): 14. In, pp 14. Perception.

LaPointe MRP, Milliken B (2016) Semantically incongruent objects attract eye gaze when viewing scenes for change. Visual Cognition 24:63–77.

Lawler EA, Silver MA (2023) Enhanced perceptual selection of predicted stimulus orientations following statistical learning. Journal of Vision 23:3.

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America A 20:1434–1448.

Lunghi C, Morrone MC, Alais D (2014) Auditory and Tactile Signals Combine to Influence Vision during Binocular Rivalry. J Neurosci 34:784–792.

Maier A, Wilke M, Aura C, Zhu C, Ye FQ, Leopold DA (2008) Divergence of fMRI and neural signals in V1 during perceptual suppression in the awake monkey. Nature Neuroscience 11:1193–1200.

Mitchell JF, Stoner GR, Reynolds JH (2004) Object-based attention determines dominance in binocular rivalry. Nature 429:410–413.

Mudrik L, Deouell LY, Lamy D (2011) Scene congruency biases Binocular Rivalry. Consciousness and Cognition 20:756–767.

Mumford D (1992) On the computational architecture of the neocortex - II. The role of the corticocortical loops. Biological Cybernetics 66:241–251.

Murray SO, Kersten D, Olshausen BA, Schrater P, Woods DL (2002) Shape perception reduces activity in human primary visual cortex. Proceedings of the National Academy of Sciences of the United States of America 99:15164–15169.

Olkkonen M, Hansen T, Gegenfurtner KR (2008) Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. Journal of Vision 8:13.

Ooi TL, He ZJ (1999) Binocular rivalry and visual awareness: The role of attention. Perception 28:551–574.

Paffen CLE, Alais D (2011) Attentional Modulation of Binocular Rivalry. Frontiers in Human Neuroscience 5:1–10.

Panagiotaropoulos TI, Deco G, Kapoor V, Logothetis NK (2012) Neuronal Discharges and Gamma Oscillations Explicitly Reflect Visual Consciousness in the Lateral Prefrontal Cortex. Neuron 74:924–935.

Pearson J, Clifford CWG, Tong F (2008) The Functional Impact of Mental Imagery on Conscious Perception. Current Biology 18:982–986.

Piazza EA, Denison RN, Silver MA (2018) Recent cross-modal statistical learning influences visual perceptual selection. Journal of Vision 18:1.

Plikat V, Grassi PR, Frack J, Bartels A (2025) Hierarchical surprise signals in naturalistic violation of expectations. Imaging Neuroscience 3:imag_a_00459.

Press C, Kok P, Yon D (2020) The Perceptual Prediction Paradox. Trends in Cognitive Sciences 24:13–24.

Qian C, Chen Z, De Hollander G, Knapen T, Zhang Z, He S, Zhang P (2023) Hierarchical and fine-scale mechanisms of binocular rivalry for conscious perception.

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2:79–87.

Schauer G, Grassi PR, Gharabaghi A, Bartels A (2024) Parietal theta burst TMS does not modulate bistable perception. Neuroscience of Consciousness 2024:niae009.

Sterzer P, Frith C, Petrovic P (2008) Believing is seeing: expectations alter visual awareness. Current Biology 18:697–698.

Teichmann L, Quek GL, Robinson AK, Grootswagers T, Carlson TA, Rich AN (2020) The Influence of Object-Color Knowledge on Emerging Object Representations in the Brain. J Neurosci 40:6779–6789.

Theeuwes J (1992) Perceptual selectivity for color and form. Perception & Psychophysics 51:599–606.

Theeuwes J (2010) Top–down and bottom–up control of visual selection. Acta Psychologica 135:77–99.

Tong F, Meng M, Blake R (2006) Neural bases of binocular rivalry. Trends in cognitive sciences 10:502–511.

Tong F, Nakayama K, Vaughan JT, Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate cortex. Neuron 21:753–759.

Underwood G, Templeman E, Lamming L, Foulsham T (2008) Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. Consciousness and Cognition 17:159–170.

Valenti JJ, Firestone C (2019) Finding the “odd one out”: Memory color effects and the logic of appearance. Cognition 191:103934.

Vandenbroucke ARE, Fahrenfort JJ, Meuwese JDI, Scholte HS, Lamme VAF (2016) Prior Knowledge about Objects Determines Neural Color Representation in Human Visual Cortex. Cereb Cortex 26:1401–1408.

Weilnhammer V, Fritsch M, Chikermane M, Eckert A-L, Kanthak K, Stuke H, Kaminski J, Sterzer P (2021) An active role of inferior frontal cortex in conscious experience. Current Biology 31:2868-2880.e8.

Weilnhammer VA, Ludwig K, Hesselmann G, Sterzer P (2013) Frontoparietal cortex mediates perceptual transitions in bistable perception. Journal of Neuroscience 33:16009–16015.

Witzel C, Valkova H, Hansen T, Gegenfurtner KR (2011) Object Knowledge Modulates Colour Appearance. i-Perception 2:13–49.

Yu K, Blake R (1992) Do recognizable figures enjoy an advantage in binocular rivalry? Journal of Experimental Psychology: Human Perception and Performance 18:1158–1173.

Zaretskaya N, Anstis S, Bartels A (2013) Parietal Cortex Mediates Conscious Perception of Illusory Gestalt. Journal of Neuroscience 33:523–531.

Zaretskaya N, Thielscher A, Logothetis NK, Bartels A (2010) Disrupting parietal function prolongs dominance durations in binocular rivalry. Current Biology 20:2106–2111.