Insight under the learning lens

What spontaneous learning dynamics can teach us about insight

Dissertation

zur Erlangung des Doktorgrades Dr. rer. nat.

an der Universität Hamburg

Fakultät für Psychologie und Bewegungswissenschaften

Institut für Psychologie

vorgelegt von

Anika Theresa Löwe

Hamburg, 2024

Tag der Disputation: 27.01.2025

Prüfungsausschuss

Vorsitz: PD Dr. Patrick Bruns

1. Dissertationsgutachten: Prof. Dr. Nicolas Schuck

2. Dissertationsgutachten: Prof. Dr. Helen Barron

1. Disputationsgutachten: Prof. Dr. Sebastian Gluth

2. Disputationsgutachten: Prof. Dr. Paul Muhle-Karbe

Contents

List of Figures

1 General Introduction

As the story goes, Archimedes (c.287 - c.212 BC) – Greek mathematician, philosopher, mechanic and inventor – was once tasked by the king of Syracuse to determine whether his crown was made out of pure gold. When he stepped into a full bath, causing water to splash out of the tub, Archimedes realised that his body had displaced the water at equal volume. He had thus found a way to calculate the density of the crown, an irregular object. By measuring the crown's mass and dividing it by the volume of displaced water, he could derive its density and determine whether it was made of pure gold. Allegedly, Archimedes thereupon excitedly ran out into the streets, still undressed, exclaiming ”Eureka!” (Greek: ”I have found it!”) (Vitruvius, 1567). Although it seems unlikely that this event would have actually happened, it probably is the first reported instance of an ”Aha!” moment - now often referred to as a Eureka effect.

Epiphanies, Eureka effects, light bulb and Aha! moments all describe the same sort of universally known experience that psychologists cluster under the term *insight*. In modern media, this sudden emergence of an idea is usually depicted by a light bulb – a visualisation that gained popularity in the 1920s through cartoons such as ”Felix the Cat”. The analogy of light as a symbol of knowledge or revelation can be traced back to Plato's ”Allegory of the Cave” (Trans. Grube, 1997) where the metaphorical transition from darkness to light describes the passage from ignorance to understanding.

As I will describe below, insight is often tied to notions of creativity, intelligence and consciousness. However, in this thesis, I take a learning perspective on the phenomenon of insight and will solely focus on behavioural markers expressed through abrupt performance changes – indicators of sudden realisations. Creativity, intelligence and consciousness as well as affective epiphenomena of insight, while interesting in their own right, are not of importance for this approach of investigating insight on a performance level.

Figure 1: Archimedes exclaiming *Eureka*.
Drawing by Pietro Scalvini, engraving by Carlo Orsolini, 1737.

1.1 Early Origins of Insight Research

Around the same time that the light bulb metaphor made an appearance in cartoons such as ”Felix the Cat” (1920), the first empirical investigation of insight took place. In 1917, Wolfgang Köhler – one of the three founding fathers of Gestalt psychology together with Max Wertheimer and Kurt Koffka – published ”The Mentality of Apes” (Köhler, 1925), a work on the intersection of ethology and cognitive psychology, describing chimpanzees’ abilities to solve problems through insight.

Köhler had spent the years of 1913-1917 studying the intelligence of apes in the Anthropoid Station in Tenerife. The behavioural experiments he conducted on several chimpanzees focused on the cognitive abilities involved in problem solving such as manipulating objects to reach a food reward. His observations led him to believe that the intellectual capacity of these higher apes was actually close to humans and he therefore concluded that intelligence correlates with brain development (Köhler, 1925).

The most prominent report of insight in this work describes an experimental situation where a banana had been fastened at a height that was out of reach for the chimpanzees. While they all tried jumping for it, the chimpanzee Sultan, however, ”[...] soon relinquished this attempt, paced restlessly up and down, suddenly stood still in front of the box, seized it, tipped it hastily straight towards the objective, but began to climb upon it at a (horizontal) distance of half a metre, and springing upwards with all his force, tore down the banana.” (Köhler, 1925). Sultan’s insightful technique of using tools in a novel way after having been stuck trying to apply his familiar skills, remains the most seminal example of Köhler’s early insight research.

A few years prior to Köhler’s experiments on Tenerife, the behavioural psychologist Edward Thorndike had published ”Animal Intelligence” (1911), a summary of experiments conducted with dogs and mainly cats who were trapped in Thorndike’s famous puzzle boxes – cages with a hidden escape mechanism that the animals had to discover in order to free themselves. Thorndike was of the belief that animals learn by forming associations between their actions and environment, through trial and error, and do not learn through insight (Thorndike, 1911). Köhler strongly opposed this view, linking the missing insight in Thorndike’s experiments to a lack of visual overview of the problem space. He reasoned that animals could not be expected to show intelligent or insightful solutions if essential elements of the experimental apparatus, such as the puzzle box cage, were invisible to the animal (Köhler, 1925).

Köhler’s main focus lay on forms and visual representation of the problem space – to him, insight solutions necessarily had to reflect the structure of the problem situation. He went as far as only considering a problem solution insightful when behaviour had arisen from a consideration of the structure of a given situation, leading him to define insight as ”the appearance of a complete solution with reference to the whole lay-out of the field” (Köhler, 1925).

Max Wertheimer frames insight as ”productive thinking”. In his posthumously published work under the same title (Wertheimer, 1959), he contrasts two ways of problem solving: *reproductive* thinking as a way of forming associative chains of familiar knowledge and *productive* thinking as insightful problem solving leading to novel, creative solutions (Wertheimer, 1959). Wertheimer also agreed that insightful productive thinking must start from a place of understanding the entire structure of the problem (Wertheimer, 1959). According to the Gestalt school, insight is reached by transitioning from a state of uncertainty to a state of complete comprehension of the solution at once (Köhler, 1925; Wertheimer, 1959).

This juxtaposition between arriving at a problem solution either through analytical thought chains (reproductive thinking) or by restructuring task representations (productive thinking), remains popular even a century later (Weisberg, 2015). Neo-Gestalt psychologists (termed by Robert Weisberg (2015)) still largely continue that spirit today by maintaining a dichotomous view of analytical thinking vs. insight through creative thinking. Insight is attained through overriding information (redistribution theory (Ohlsson, 2011)) or as the result of breakthrough thinking (Perkins, 2000).

Stellan Ohlsson's (1984) information processing angle on insight combines problem solving with a semantic network theory, thereby accounting for the restructuring phenomenon in insight as proposed by the Gestalters. According to the information processing theory of insight, representational changes influence the internal search for the solution, which is restricted by capacity limitations (Ohlsson, 1984). When the goal state appears in the internal problem search space by means of representational change, insight can occur (Ohlsson, 1984). Furthermore, Ohlsson introduced an important concept for the conceptualisation of insight – the impasse (Ohlsson, 1984). He proposes that insight is a breaking out of the impasse – the "mental state in which problem-solving has come to a halt" and "the problem-solver cannot think of any way to proceed" (Ohlsson, 1992). The representational restructuring can take several forms of altering long-term memory pattern activation, such as re-encoding of information, constraint relaxation or elaboration (Ohlsson, 1992). Both re-encoding and elaboration work on restructuring the representation of the *problem* state. Elaboration extends the representation by adding features previously unnoticed, while re-encoding changes the fundamental problem representation (Ohlsson, 1992). Constraint relaxation on the other hand, assumes that mental constraints have been posed on the *goal* state, making the problem unsolvable unless this mental representation is changed (such as in considering 3D solutions instead of a 2D goal space for visual puzzles) (Ohlsson, 1992).

Knoblich et al. (1999) a few years later expanded on this idea in their *Representational Change Theory* where they proposed that impasses are broken by changing the problem representation through either the relaxation of constraints on the solution as described above or through the decomposition of perceptual chunks. The latter describes a process of decomposing chunks – created patterns of features or components that led to the impasse – into component features, enabling alternative combinations to tackle the problem situation (Knoblich et al., 1999). Impasse can be a terminal state if the person is not actually capable of finding the solution, or can lead to partial insight if the impasse is broken out of to resume analytical thinking (Ohlsson, 1992).

1.2 Behavioural Markers of Insight

Although Köhler did not focus on this, the chimpanzee Sultan reached what would be termed an impasse when he was described as having ”paced restlessly up and down, suddenly stood still in front of the box, seized it” and then suddenly started his insightful approach to obtain the banana. While insight research nowadays focuses largely on human problem solving, most behavioural signatures of insight today have been described in Köhler's chimpanzee Sultan.

The most important behavioural and neural markers of insight, affective components accompanying insight as well as insight problems used to measure Aha! moments, are described below.

Impasse Early problem solving research on insight focused heavily on representational restructuring and breaking of the impasse (Ohlsson, 1992; Knoblich et al., 1999; Wertheimer, 1959), defined as the state where the subject is either unable to continue with problem solving or actively chooses not to proceed (Cranford & Moss, 2012). Other researchers however disagree on the notion that impasse constitutes a necessary prerequisite of insight (Stuyck, Aben, Cleeremans, & Van den Bussche, 2021; Fleck & Weisberg, 2013; Danek, Fraps, von Müller, Grothe, & Öllinger, 2014). Stuyck et al. (2021) trace a lack of impasse back to differences in task specifics where for example, compared to classic insight problems, shorter trial durations might be too short for impasse to develop (see Insight Problems for details).

Further, Kounios and Beeman (2014) criticise that the view of impasse being critical for insight ignores instances of ”(a) when the solution suddenly intrudes on a person's awareness when he or she is not focusing on any solution strategy, (b) when an insight pointing to a solution occurs while a person is actively engaged in analytic processing but has not yet reached an impasse, and (c) when a person has a spontaneous realisation that does not relate to any explicitly posed problem”.

Impasse is thus a rather task-specific occurrence that is not necessarily essential for insight.

Incubation Incubation, first described by Wallas (1926), is a state in the insight sequence where the problem is temporarily set aside in order to rest or focus on something else. According to Wallas, humans undergo four different stages when facing a problem: preparation, incubation, illumination, and verification (Wallas, 1926). Illumination, which can be regarded as insight here, describes the spontaneous manifestation of the problem solution in conscious thought (Wallas, 1926; Hélie & Sun, 2010). Both verification and preparation involve deliberative thinking processes, whereby the verification stage seeks to confirm the validity of the solution (Wallas, 1926).

Hélie and Sun built on Wallas' stages with their explicit-implicit interaction (EII) theory that offers a detailed process-based model of incubation and insight (Hélie & Sun, 2010). At the core of this interactive theory, information flows between an explicit processing level, indicated in the deliberative preparation and verification stages, and an implicit processing level, which is employed during incubation. Insight results from transferring the solution from the implicit to the explicit processing level (Hélie & Sun, 2010).

Experimentally, giving subjects an incubation period to rest has been found to lead to increased insight after incubation, compared to consecutive solution attempts (Brodt, Pöhlchen, Täumer, Gais, & Schönauer, 2018) or a different problem-unrelated task (Craig, Ottaway, & Dewar, 2018). However, subjects performing an undemanding task, as opposed to a demanding one, reported increased mind-wandering and were subsequently more likely to have insight into a hidden task regularity (Tan, Zou, Chen, & Luo, 2015). In support of this, a meta-analytic review of incubation and problem solving found a stronger effect of tasks with low cognitive demands, particularly for verbal insight problems (Sio & Ormerod, 2009). Longer incubation periods were found to lead to larger effects and tasks requiring divergent thinking generally profited more from incubation periods than visual or verbal ones (Sio & Ormerod, 2009).

Sleep A special role is furthermore ascribed to sleep as a particular kind of incubation phase that leads to restructuring of representations and therefore insight (Wagner, Gais, Haider, Verleger, & Born, 2004; Lacaux et al., 2021) by restructuring memories (Cowan et al., 2020). The evidence for this, however, is inconclusive.

The famous study by Wagner et al. (2004) reported a benefit of a full night's sleep for insight with more than twice as many subjects discovering a hidden rule after sleeping. Similar results were presented by Lacaux et al. (2021) who found a beneficial effect of particularly N1 sleep on hidden rule insight after a daytime nap sleep intervention. Other investigations though did either find no benefits of sleep or no difference between wake rest and sleep (Cordi & Rasch, 2021; Schönauer et al., 2018; Brodt et al., 2018).

A potential explanation for these diverging findings might be differential effects of different sleep phases. While some studies found specifically N1 sleep, the first sleep phase of NREM sleep, to be beneficial for insight (Lacaux et al., 2021; Vickrey & Lerner, 2023), another study suggests that slow wave sleep in contrast might have a particularly important role for generating insight (Verleger, Rose, Wagner, Yordanova, & Kolev, 2013).

In order to understand the relationship of sleep as an incubator for insight, it will thus be important to not not only scrutinise different sleep phases, but also understand the different neural signatures underlying those states and their relation to representational restructuring leading to insight. The relationship between sleep and insight is discussed in detail in Chapter 4.

Suddenness The hallmark signature of insight seems to be suddenness (Bowden, Jung-Beeman, Fleck, & Kounios, 2005; Stuyck et al., 2021). A sudden arising of the insight solutions can be observed in virtually all investigations of insight, in humans as well as animals. High suddenness of the solution comprehension and high solution confidence are both linked to an increased likelihood of having experienced insight (Stuyck et al., 2021).

Metcalfe and Wiebe (1987) introduced a concept of warmth rating to track subject's metacognitive performance monitoring. While solving insight problems, participants had to rate how close (i.e. *warm*) they were to the correct solution. On algebraic non-insight problems, subjects' warmth ratings could accurately predict performance, showing that these tasks were solved incrementally (Metcalfe & Wiebe, 1987). The same was not true for insight problems where solutions appeared in a ”sudden flash of illumination” and participants were unable to indicate their progress through warmth ratings (Metcalfe & Wiebe, 1987).

Gick and Lockhart (1996) reason that finding the correct representation of a problem that leads to the solution and an easy application thereof evokes a sensation of sudden appearance. They further relate this to an affective component of insight by proposing that surprise about the fact that the solution representation is different from initial attempts yields a surprising affective component of either positive (delight) or negative (chagrin) nature (Gick & Lockhart, 1996). According to their hypothesis it is thus not necessarily the cognitive insight components that are sudden, but the sudden and surprising affective components accompanying insight (Gick & Lockhart, 1996).

Affective Components Besides aforementioned feelings of solution certainty and either positive or negative surprise (Gick & Lockhart, 1996), insight is usually accompanied by a feeling of relief or pleasure (Danek et al., 2014; Kounios & Beeman, 2015). This primacy of positive affect is reflected in participants' self-reports when solving insight problems (Danek et al., 2014). The three predominant emotional states associated with insight and Aha! experiences have been identified as happiness, ease and certainty (Shen, Yuan, Liu, & Luo, 2016).

Danek et al. (2014) introduce tension release as an additional physiological aspect of the Aha! insight experience, suggesting a build up of tension during unsuccessful problem solving and a rapid decline thereof upon unexpectedly finding the solution (Danek et al., 2014). Insight thus appears to us as a sudden solution comprehension, usually without conscious access to the steps leading up to the solution, triggering surprise and bursts of positive emotions (Sprugnoli et al., 2017).

Insight Problems

Riddles and Divergent Thinking Classic insight tasks introduced by the Gestalt psychologists are puzzle-like divergent thinking tasks that often involve a misleading component which has to be overcome. Duncker's (1945) famous candle task for example has the following setup: ”On the table lie, among many other objects, a few tacks and three crucial objects: three little pasteboard boxes [...].”. The task is to fasten candles to the wall with no further instructions given. Duncker (1945) describes the solution as follows: ”with a tack piece, the three boxes are fastened to the door, each to serve as platform for a candle. In the setting a.p., the three boxes were filled with experimental material: in one there were several thin little candles, tacks in another, and matches in die third.”.

Another early example of a classic insight task is the 9-dot problem (Maier, 1930) (Fig. 2A). Subjects are presented nine dots arranged in a 3x3 grid with the instructions that ”four lines must be drawn in such a manner that all dots will be passed through. The pencil must not be taken from the paper and no line should be retraced.” (Maier, 1930). Only if subjects detach themselves from the idea that the solution will have the form of a symmetrical shape, they can solve the problem.

The 8-coin problem has a similar functionality with the task being to ”alter an array of x coins by moving y coins only, to create a final array in which each coin touches exactly z other” (Ormerod, MacGregor, & Chronicle, 2002) (Fig. 2B). The primary constraint people face when solving this task is thinking in two-dimensional terms – only when they have the insight of considering three-dimensional moves, can they solve the task (Ormerod et al., 2002).

More recently, these divergent thinking based insight tasks have been getting more creative and even magic tricks have been employed to investigate participants' ability to gain insight into the underlying logic (Danek et al., 2014; Hedne, Norman, & Metcalfe, 2016). The general idea with these riddle like problems is that in order to solve them, participants cannot rely on previous experience with familiar strategies (Tulver, Kaup, Laukkonen, & Aru, 2023).

These classic insight problem pose a few experimental limitations however. For one, their high difficulty and complicated underlying generative processes yield only a small percentage of insight solvers, while they further also only lead to one potential insight event since they cannot be retested and lastly, insight is the only way to solve them as analytical problem solving won't bring about the solution (Sprugnoli et al., 2017). Researchers have thus come up with new types of insight problems, exploring different cognitive and perceptual realms as well as developing tasks involving multiple insight trials and possibilities to also solve problems analytically.

Verbal Insight The most commonly used insight problems today are two instances of verbal insight tasks: the remote associates task (RAT) and compound remote associates (CRA). The RAT was first created by Mednick (1968) to measure creative thinking before insight researchers started utilising it (Beeman & Bowden, 2000). Inspired by the RAT, Bowden et al. (2003) later developed the CRA. In both problems, participants are tasked to find a matching word related to three stimuli words. In the CRA the solution word must form a compound word with the presented stimuli (i.e. “crab, pine, sauce”,

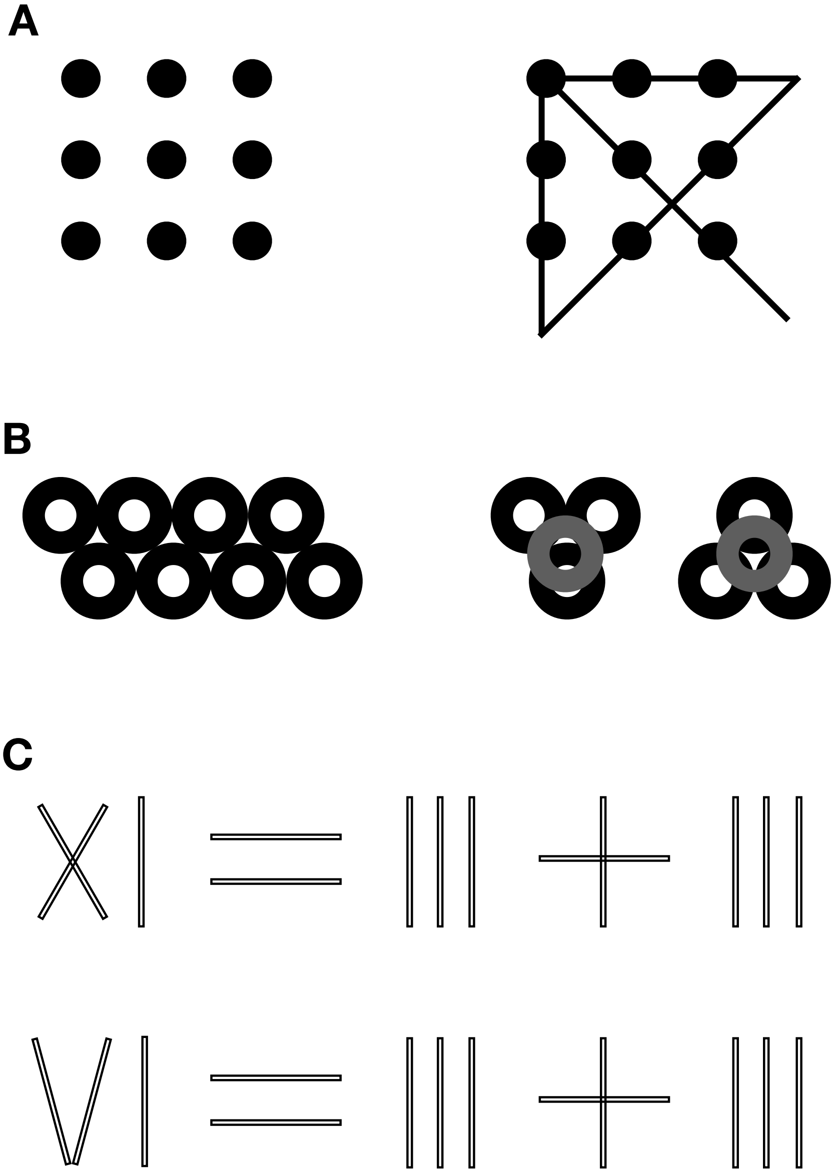


Figure 2: Classic insight tasks.

(A) The 9-dot problem and the solution (right) of how to connect all dots with four connected lines. **(B)** The 8-coin problem and the solution (right) for the problem of altering an array of 8 coins by moving 2 coins only, to create a final array in which each coin touches exactly 3 other. **(C)** The matchstick arithmetic problem showing a false Roman numeral equation made of matchsticks and the solution (bottom) of transforming the equation into a mathematically correct one by moving only one matchstick.

solution: “apple = crabapple, pineapple, applesauce”), whereas the RAT does not pose this constraint (i.e. “falling, actor, dust”, solution: “star = falling star, movie star and stardust”) (Tulver et al., 2023). As mentioned above, Stuyck (2021) noticed that the CRA has not been found to elicit impasse (Cranford & Moss, 2012). This points towards certain task specifics of the RAT/CRA differing from classic insight problems, such as a short trial duration during the solution search as well as the lack of a misleading problem component (Stuyck et al., 2021).

Another common verbal insight problem is solving anagrams (Smith & Kounios, 1996; Kounios et al., 2008; Aziz-Zadeh, Kaplan, & Iacoboni, 2009; Oh, Chesebrough, Erickson, Zhang, & Kounios, 2020). Since it has been found that the subliminal presentation of solution-related words increased insight-like solving of the anagrams (Bowden, 1997), this suggests that insight solutions are preceded by substantial unconscious processing rather than spontaneous generation (Kounios & Beeman, 2014).

Visuo-Spatial Insight A popular non-verbal insight task is the matchstick arithmetic problem (Knoblich et al., 1999) (Fig. 2C). Participants are shown false Roman numeral equations made of matchsticks and are instructed to transform the equation into a mathematically correct one by moving only one matchstick (Knoblich et al., 1999).

Mooney images are thresholded two-tone images of usually real-world objects that are hard to recognise due to the high contrast (Imamoglu, Koch, & Haynes, 2013). Solving Mooney images is a purely visual task that is specifically used to assess perceptual insight (Imamoglu et al., 2013; Kizilirmak, Galvao Gomes da Silva, Imamoglu, & Richardson-Klavehn, 2016; Becker, Yu, & Cabeza, 2023). Insight can be gained through a visual regrouping of the abstracted shapes which can then lead to sudden object recognition (Kizilirmak et al., 2016; Ludmer, Dudai, & Rubin, 2011; Becker et al., 2023).

Hidden Rule Besides the serial reaction time task (SRTT) used as a pattern extraction implicit learning task (Nissen & Bullemer, 1987), the number reduction task

(NRT) (Thurstone & Thurstone, 1941; Woltz, Bell, Kyllonen, & Gardner, 1996) provides a different kind of insight problem: hidden task regularities. In the NRT, participants are given numerical sequences and a certain mathematical rule to solve the sequence. Unmentioned to participants, there is a hidden rule (e.g. the second number will always be the final solution), enabling the participant to solve the task much faster and easier if they have an insight about this hidden task regularity (Wagner et al., 2004; Verleger et al., 2013; Yordanova, Kolev, Wagner, Born, & Verleger, 2012).

1.3 Neural Correlates of Insight

Although insight has been a topic of interest for more than a century, the advance of neuroimaging methods has offered new ways to explore the neural correlates of problem solving only in the last two decades.

The first study investigating verbal RAT insight problems using both functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) in separate experiments found that insight solutions were associated with greater neural activity in the right anterior superior temporal gyrus (aSTG) (Jung-Beeman et al., 2004). Insight solutions were preceded by bursts of high-frequency gamma band neural activity in the same right aSTG area 0.3 seconds prior to the solution, which was interpreted to reflect the transition of the unconscious to conscious insight solution (Jung-Beeman et al., 2004). Additionally, sudden increases in alpha band frequency about 1.5 seconds prior to the insight solution lead the authors to suggest a focus shift, allowing solution information to gain conscious access (Bowden et al., 2005). Reviews of EEG insight studies however find a consistently reported decrease in alpha power in frontal, parietal and temporal areas (Dietrich & Kanso, 2010; Sprugnoli et al., 2017).

Sandkühler and Bhattacharya's (2008) EEG study sought to find neural correlates for different behavioural markers of verbal insight. Impasse was associated with effects in parieto-occipital and occipital areas in the gamma frequency band 3 to 2.5 seconds as well as in the theta band 2.25 to 1.5 seconds prior to reaching impasse, which they interpreted to reflect selective attentional processes (Sandkühler & Bhattacharya, 2008). Restructuring was linked to a right prefrontal effect from 1.5 to 0.25 s before solution in the alpha frequency band (Sandkühler & Bhattacharya, 2008). Individual ratings of suddenness were found to be expressed in the same parieto-occipital area gamma frequency band from 1.5 to 1 and from 0.75 to 0 seconds before the insight solution response, reflecting the gamma band effects of Jung-Beeman et al. (2004), but in more posterior regions, potentially reflecting successful retrieval of a new solution word (Sandkühler & Bhattacharya, 2008).

An fMRI study employing the NRT to assess insight into a hidden task regularity found that subjects gaining insight already showed distinct neural correlates during implicit learning, before conscious access to the insight solution (Darsaud et al., 2011). Hippocampal activity in insight solvers increased proportionally to frontal responses during implicit learning, while changes in learning-related responses overnight were detected in the ventromedial prefrontal cortex (Darsaud et al., 2011).

An activation likelihood estimation based meta-analysis of 36 fMRI studies found evidence for an integrated insight network of left inferior frontal gyrus (IFG), right medial dorsal frontal gyrus, right hippocampal gyrus, as well as the left amygdala (Shen et al., 2018). The review further associates Wallas' (1926) four stages of insight with specific neural correlates, finding the left anterior cingulate cortex to be associated with preparation and the right IFG with verification, while incubation and illumination were associated with more complex network activations (Shen et al., 2018).

The first 7T fMRI study was conducted by Tik et al. (2018) which enabled them to investigate the relation of subcortical structures to insight. The dopaminergic pathway and particularly the nucleus accumbens (NAcc) were found to be modulated by insight with higher insight ratings being linked to higher NAcc activations, potentially reflecting the accompanying sensations of relief and pleasure (Tik et al., 2018). Higher insight was further related to caudate nucleaus, as well as left anterior medial temporal gyrus activation (Tik et al., 2018). Another study similarly found greater NAcc activity when participants solved Mooney images with high compared to low insight, relating the NAcc activity to positive affect associated with internal reward (Becker et al., 2023). These findings of insight-related neural reward signals are substantiated by Oh et al. (2020) who found that the primary neural insight correlate, a burst of gamma band activity about 500 ms prior to solving anagrams with insight, was modulated by dispositional reward sensitivity and expressed by a separate anterior prefrontal gamma band burst 100 ms later. Source modeling of the reward-related insight effect showed a positive correlation of the insight-related left temporal lobe activity with reward-related activity in the right orbitofrontal gyrus (Oh et al., 2020). Taken together these studies suggest that reward is an accompanying feature of insight, rather than a necessary prerequisite.

Overall, these studies and reviews can be concluded as painting a heterogeneous picture of insight where besides certain commonly observed neural signatures of insight (gamma and alpha frequency bands, prefrontal, temporal and reward areas as well as the hippocampus), the underlying neural correlates are highly task dependent. It will be of great interest for future research to not only understand the neural computations underlying the insight *moment*, but also leading up to insight and moreover to understand the causes of insight occurring in only a subset of participants or certain trials. In this thesis I combine behavioural insight data, computational models and sleep EEG to elucidate these questions further.

1.4 Computational Models of Insight

While insight has been extensively studied by psychologists, few researchers have attempted to investigate the insight phenomenon from a computational perspective.

Langley and Jones (1986) were the first to model insight problem solving using a conceptual cognitive model of spreading activation. Semantic networks are here proposed to spread from a source node after it has been activated by the environment (Langley & Jones, 1986). According to their model, insight is then achieved upon sudden retrieval of an analogy from long-term memory (Langley & Jones, 1986).

The most influential computational account of insight, focusing on incubation and impasse, has been presented by Hélie and Sun (2010) (see *Incubation* above). Their EII (explicit-implicit interaction) model uses an architecture which has two separate, but interacting modules for explicit and implicit knowledge. Insight occurs after transferring the solution from implicit to explicit knowledge – this is achieved upon crossing an internal confidence level threshold through an active, iterative problem solving process (Hélie & Sun, 2010).

A recent reinforcement learning model suggested that loss aversion was predictive of insight by linking the parameters of a Q-learning model to individual traits (Harada, 2023). This study was the first instance of linking computational parameters to inter-individual differences in insight problem solving. Albeit conceptual models of insight are valuable for the mechanistic understanding of insight, they teach us little about the underlying neuro-computational processes. A promising avenue for investigating insight on this level are neural network models. By modelling insight problem solving with artificial neural networks, different aspects of the architecture and parameters which might be linked to insight can be identified and moreover manipulated. These parameters could potentially explain not only inter-individual differences, but also trial and task dependencies. Additionally, tweaking neural network parameters allows to modify and assess representations in ways that would be impossible to test in humans. Chapter 3 compares learning dynamics of human behaviour and artificial neural networks on the same hidden rule insight task and identifies crucial parameters for inducing insight. To our knowledge, this is the first account of analysing neural networks to understand the cognitive phenomena underlying insight.

1.5 Outline of this thesis

From the above it becomes clear that, although insight has attracted researchers' attention for over a century, the underlying cognitive and neural mechanisms are still debated. Most insight research focuses on juxtaposing insightful problem solving with analytical problem solving. This is problematic for multiple reasons: (i) insight is defined based on subjects' self-reports, meaning that it will only be classified as such when it has reached the subject's conscious thought, neglecting preceding subconscious processing, (ii) insight is usually assessed at the end of the trial, making it difficult to precisely match neural to cognitive events during the problem solving event, and (iii) most commonly used insight paradigms require subjects to actively search for solutions which ignores scenarios where participants have an unexpected insight about a phenomenon they were not aware existed.

A hallmark of understanding insight will thus be to understand the neural and cognitive processes *that give rise to* insight. Particularly in cases where insight has not been instructed and happens intrinsically, it will be of fundamental interest to understand which processes and mechanisms cause the representational restructuring that eventually leads to insight. A second important quest is to understand the selectivity of insight, i.e. why insight happens only in a subset of participants or certain trial types. While some researchers have tried correlating insight likelihood with specific personality traits (Ovington, Saliba, & Goldring, 2016; Kounios & Beeman, 2014), it is not understood yet what causes these differences. An important milestone along the path to understanding insight will furthermore be to identify factors facilitating insight. In this thesis I thus investigate the mechanisms underlying insight and its selective occurrence on a computational level, as well as probe sleep as a factor promoting insight using sleep EEG. Comparing learning dynamics between humans and neural networks allows me to identify and modify parameters to elucidate insight in ways that would not be possible using solely human behaviour.

Section 2 of the **General Introduction** introduces the Perceptual Spontaneous Strategy Switch Task (PSSST) – a hidden rule insight task that allows trial wise tracking of knowledge of both hidden and learnt task regularities. We developed this insight task, because it allows us to investigate and model cognitive processes leading up to insight with high temporal resolution on a trial basis. Furthermore, it can be employed to investigate insight selectivity on a computational level, since it can be used in artificial agents as well as biological ones.

Chapter 2 employs the PSSST on a sample of 99 human participants and distinguishes three main behavioural characteristics of insight: suddenness, selectivity and delay.

Chapter 3 investigates whether insight as a learning phenomenon could arise in artificial neural networks. The PSST was used on simple neural networks to compare insight-related learning dynamics with the human sample of Chapter 2. We find that neural networks mimic all three human insight characteristics and moreover that insight-like learning depends on regularised gating as well as gradient noise.

Finally, **Chapter 4** scrutinises sleep as a candidate factor of increasing insight. Specifically, this preregistered study tested the effect of different sleep phases on insight and further built on our computational findings of regularisation as a critical component for inducing insight. EEG was recorded while participants took a nap between two sessions of the task. N2 sleep had a beneficial effect on post-nap insight likelihood, with the spectral slope predicting insight beyond just sleep stage, suggesting a role for regularisation during sleep.

1.6 Measuring Insight: The Perceptual Spontaneous Strategy Switch Task (PSSST)

1.6.1 Task Rationale

As described in the previous section, classic insight paradigms require participants to actively search for novel problem solutions. These problems might range from concrete tasks such as solving magic tricks (Danek et al., 2014; Hedne et al., 2016) to more abstract verbal or geometric problems (e.g. Remotes Associates Tasks (Mednick, 1968; Bowden & Jung-beeman, 2003; Knoblich et al., 1999)). Oftentimes, these paradigms distinguish between solutions that subjects reached via analytical problem solving vs solutions gained through insight (Bowden et al., 2005; Kounios & Beeman, 2014).

These classic insight tasks usually measure a binary outcome, namely: did subjects solve a task using insight (y/n)? Besides reaction times, above mentioned paradigms thus tell us little about the cognitive processes leading up to insight during problem solving. We therefore developed the Perceptual Spontaneous Strategy Switch Task (PSSST) which measures insight with high temporal resolution on a trial basis. This allows us to model cognitive processes around insight moments with high precision. Essentially, the PSSST involves a hidden rule that subjects can have an insight about, which offers a more efficient way to solve the task. Since participants are unaware that the task at hand might involve an insight process, the PSSST offers a more naturalistic way of investigating insight.

The PSSST is a random dot motion task that is based on the similar Spontaneous Strategy Switch Task employed by earlier research investigating insight-like strategy switches (Schuck et al., 2015; Gaschler, Schuck, Reverberi, Frensch, & Wenke, 2019; Schuck et al., 2022). In the task that the PSSST was built on, participants were presented a patch made of little squares, coloured in either red or green, that was positioned inside a white reference frame in a way so that it was always marginally closer to one of the four corners of the frame (Schuck et al., 2015; Gaschler et al., 2019; Schuck et al., 2022; Gaschler, Marewski, & Frensch, 2015; Gaschler, Vaterrodt, Frensch, Eichler, & Haider, 2013). Participants' task was to judge which corner the patch was closest to and indicate this with one of two buttons. Unmentioned to participants, a hidden task regularity of colour being predictive of the correct button press appeared after a few blocks of the task. This was particularly relevant on ”ambiguous” trials where the patch was at equal distance from all frame corners and the instructed strategy could thus not be used to press correctly (Schuck et al., 2015; Gaschler et al., 2019; Schuck et al., 2022; Gaschler et al., 2015, 2013). The above mentioned studies using this task paradigm consistently found that i) not all participants discovered the hidden rule ii) if they noticed it, they started using it and thus improved their behaviour in an abrupt manner and ii) the task switches occurred at different times for different participants (Schuck et al., 2015; Gaschler et al., 2019; Schuck et al., 2022; Gaschler et al., 2015, 2013). Interestingly, even children discovered and used the hidden task rule at the same rate as young adults, although their task performance and cognitive control measures were generally significantly worse (Schuck et al., 2022). An fMRI study using this task paradigm found that the colour strategy could be decoded from the medial prefrontal cortex immediately before participants started using it – as though the alternative strategy was unconsciously simulated or processed in parallel while participants were still consciously focused on the instructed corner strategy (Schuck et al., 2015).

While studies using this task show strong behavioural analogies to insight, the switch phenomenon has been framed as spontaneous strategy switches until now. Below, we develop the task further and explain why we deem it an ideal task for measuring insight.

1.6.2 Stimuli

The perceptual decision task requires a binary choice about circular arrays of moving dots (Rajananda, Lau, & Odegaard, 2018). Dots are characterised by two features, (1) a motion direction (four possible orthogonal directions: NW, NE, SW, SE) and (2) a colour (orange or purple, Fig. 3A). The noise level of the motion feature is varied in 5 steps (5%, 10%, 20%, 30% or 45% coherent motion), making motion judgement relatively harder or easier. Colour difficulty is constant, thus consistently allowing easy identification of the stimulus colour. The condition with most noise (5% coherence) occurs slightly more frequently than the other conditions (30 trial per 100, vs 10, 20, 20, 20 for the other conditions respectively).

On every trial, participants are presented a cloud of 200 moving dots with a radius of 7 pixels each. In order to avoid tracking of individual dots, dots have a lifetime of 10 frames before they are replaced. Within the circle shape of 400 pixel width, a single dot moves 6 pixel lengths in a given frame. Each dot is either designated to be coherent or incoherent and remains so throughout all frames in the display, whereby each incoherent dot follows a randomly designated alternative direction of motion.

The trial duration is 2000 ms and a response can be made at any point during that time window. After a response has been made via one of the two button presses, the white fixation cross at the centre of the stimulus turns into a binary feedback symbol (happy or sad smiley) that is displayed until the end of the trial (Fig. 3B). An inter trial interval (ITI) of either 400, 600, 800 or 1000 ms is randomly selected. If no response is made, a ”TOO SLOW” feedback is displayed for 300 ms before being replaced by the fixation cross for the remaining time of the ITI.

1.6.3 Task Design

For the first 400 trials, the correct binary choice is only related to stimulus motion (two directions each on a diagonal are mapped onto one choice), while the colour changes randomly from trial to trial (Fig. 3C). For the binary choice, participants are given two response keys, ”X” and ”M”. The NW and SE motion directions correspond to a left key press (”X”), while NE and SW correspond to a right key press (”M”) (Fig. 3A). Participants receive trial-wise binary feedback (correct or incorrect), and therefore can learn which choice they have to make in response to which motion direction (Fig. 3B).

Participants are not specifically instructed to pay attention to the motion direction. Instead, they are instructed to learn how to classify the moving dot clouds using the two response keys, so that they would maximise their number of correct choices. To ensure

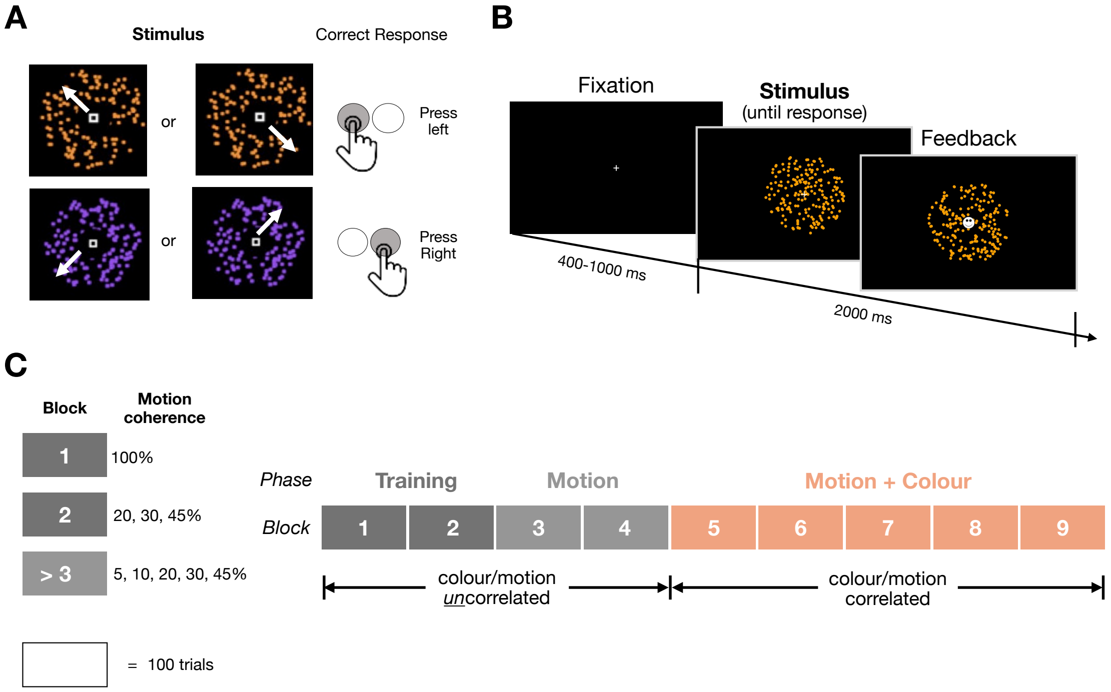


Figure 3: The Perceptual Spontaneous Strategy Switch Task (PSSST).

(A) Stimuli and stimulus-response mapping: dot clouds are either coloured in orange or purple and move to one of the four directions NW, NE, SE, SW with varying coherence. A left response key, ”X”, corresponds to the NW/SE motion direction diagonal, while a right response key ”M” corresponds to NE/SW direction diagonal, so that a diagonal each maps onto one of the two response keys. **(B)** Trial structure: a fixation cue is shown for a duration that is shuffled between 400, 600, 800 and 1000 ms. The random dot cloud stimulus is displayed for 2000 ms. A response can be made during these entire 2000 ms, but a central feedback cue will replace the fixation cue immediately after a response. Feedback is administered in terms of a happy or sad smiley depending on the choice made. **(C)** Task structure: each block consists of 100 trials. A first training block contains only 100% motion coherence trials to familiarise subjects with the stimulus-response mapping. The second training block contains only high coherence (20, 30, 45%) trials. All motion coherence levels (5, 10, 20, 30, 45%) are included starting in block 3. In both the *training* and *motion phase*, colour changes randomly and is not predictive. Colour becomes predictive of correct choices and correlates with motion directions as well as correct response buttons only in the last five blocks (*motion and colour phase*). Participants are instructed to use colour before the very last block 9, which serves as a sanity check.

that participants pick up on the motion relevance and the correct stimulus-response mapping, motion coherence is set to be at 100% in the first block (100 trials), meaning that all dots move towards one coherent direction. In the second task block, the lowest, and therefore easiest, three levels of motion noise (20%, 30% and 45% coherent motion) are introduced, before all five noise levels start in block 3. Since choices during this phase should become solely dependent on motion, they should be affected by the level of motion noise. It is assessed how well participants learn to discriminate the motion direction after the fourth block. Participants that do not reach an accuracy level of at least 85% in the three lowest motion noise levels during this last task block of the *motion phase* are excluded from the *motion and colour phase*.

After the *motion phase*, in the *motion and colour phase*, the colour feature becomes predictive of the correct choice in addition to the motion feature (Fig. 3C). This means that each response key, and thus motion direction diagonal, is consistently paired with one colour, and that colour is fully predictive of the required choice. Orange henceforth corresponds to a correct ”X” key press and a NW/SE motion direction, while purple is predictive of a correct ”M” key press and NE/SW motion direction (Fig. 3A). This change in feature relevance is not announced to participants, and the task continues for another 500 trials as before - the only change being the predictiveness of colour.

Before the last task block participants are asked whether they 1) noticed a rule in the experiment, 2) how long it took until they noticed it, 3) whether they used the colour feature to make their choices and 4) to replicate the mapping between stimulus colour and motion directions. Participants are then instructed about the correct colour mapping and asked to rely on colour for the last task block. This serves as a proof that subjects are in principle able to do the task based on the colour feature and to show that, based on this easier task strategy, accuracy should be near ceiling for all participants in the last instructed block.

1.6.4 Differences to Other Insight Tasks

In the PSSST, participants initially learn a functional, but suboptimal, strategy. This strategy is spontaneously replaced by some participants with a more optimal solution, mirroring an insight process (Schuck et al., 2015, 2022; Gaschler et al., 2019; Allegra et al., 2020). Participants are not made aware of this superior strategy. The PSSST therefore differs from other insight tasks where participants are asked to actively search for a novel problem solution (e.g. Remotes Associates Tasks (Mednick, 1968) or Compounds Remotes Associates Tasks (Bowden & Jung-beeman, 2003)). Nevertheless, the task aligns with the core concept of almost all insight tasks, which require a modification of the initial problem representation (Tulver et al., 2023). Indeed, our task is very similar to the well established Number Reduction Task (NRT) (Thurstone & Thurstone, 1941; Woltz et al., 1996; Wagner et al., 2004). Both the PSSST and the NRT measure 'intrinsic' insight where the hidden rule as a potential strategy improvement is never mentioned to participants, and both tasks can be solved in principle even if the hidden rule is not discovered, by using the initially learned rule.

We thus deem the PSSST most suitable to formal analysis of insight, since participants' knowledge of both the hidden and learnt strategy can be tracked with high temporal resolution on a trial basis (Haider & Rose, 2007).

2 Behavioural Characteristics of Insight

Adapted from:

Löwe, A.T., Touzo, L.T., Muhle-Karbe, P.S., Saxe, A.M., Summerfield, C.* & Schuck, N.W.* (2024). *PLoS Computational Biology*.

Abstract

Humans sometimes have an insight that leads to a sudden and drastic performance improvement on the task they are working on. Sudden strategy adaptations are often linked to insights, considered to be a unique aspect of human cognition tied to complex processes such as creativity or meta-cognitive reasoning. To study the learning dynamics involved in insight, we let 99 participants perform a perceptual decision task that included a hidden regularity to solve the task more efficiently. As opposed to other common insight paradigms, our task allows us to track both the hidden and learnt strategy with high temporal resolution on a trial basis, thus offering a fine grained temporal window into the learning dynamics underlying insight. Our results show that only a subset of participants discovers this regularity, whose behaviour was marked by a sudden and abrupt strategy switch that reflects an Aha! moment. Furthermore, the insight into the hidden rule occurred delayed, at various moments of the task, and always suddenly, within a few trials only. We can thus characterise insight based on three main attributes: suddenness, selectivity and delay. These characteristics allow to measure and compare insight-like behaviour across natural and artificial intelligence, therefore opening new avenues for insight research.

This chapter was reprinted under a CC-BY 4.0 license. doi: . Author contributions: ATL, CS, PMK and NWS conceived the experiments. ATL carried out the experiments and analysed the data. ATL, AS and LT provided simulations. All authors discussed the results and contributed to the final manuscript.

2.1 Introduction

Humans sometimes learn and improve on a task in a seemingly spontaneous and abrupt manner. These striking cases have been related to insights or Aha! moments (Köhler, 1925; Durstewitz, Vittoz, Floresco, & Seamans, 2010), which are thought to reflect a qualitatively different, discrete learning mechanism (Stuyck et al., 2021; Weisberg, 2015). One prominent idea, dating back to Gestalt psychology (Köhler, 1925; Wertheimer, 1925), is that an insight occurs when an agent has found a novel problem solution by restructuring an existing task representation (Kounios & Beeman, 2014; Ohlsson, 1992). It has also been noted that humans often lack the ability to trace back the cognitive process leading up to an insight (Jung-Beeman et al., 2004), suggesting that insights involve unconscious processes becoming conscious. Moreover, so called “Aha! moments” can sometimes even be accompanied by a feeling of relief or pleasure in humans (Kounios & Beeman, 2014; Danek et al., 2014; Kounios & Beeman, 2015). Such putative uniqueness of the insight phenomenon would also be in line with work that has related insights to brain regions distinct from those associated with gradual learning (Shen et al., 2018; Jung-Beeman et al., 2004; Tik et al., 2018). Altogether, these findings have led psychologists and neuroscientists to propose that insights are governed by a distinct learning or reasoning process (Jung-Beeman et al., 2004) that cannot be accounted for by common gradual theories of learning.

Here, we focus on characterising insight based on the following three behavioural observations (Schuck et al., 2015, 2022; Gaschler et al., 2019, 2013, 2015): First, insights trigger abrupt behavioural changes. These sudden behavioural changes are often accompanied by fast neural transitions (Durstewitz et al., 2010; Karlsson, Tervo, & Karpova, 2012; Miller & Katz, 2010; Schuck et al., 2015; Allegra et al., 2020), and by meta-cognitive suddenness (a “sudden and unexpected flash”) (Bowden et al., 2005; Gaschler et al., 2013; Metcalfe & Wiebe, 1987; Weisberg, 2015; Tulver et al., 2023). Second, insights occur selectively in some subjects, while for others improvement in task performance arises only gradually, or never (Schuck et al., 2015). Finally, insights occur “spontaneously”, i.e. without the help of external cues (Friston et al., 2017), and are therefore observed after a seemingly random duration of impasse or delay (Ohlsson, 1992) that differs between participants. In other words, participants seem to be “blind” to the new solution for an extended period of time, before it suddenly occurs to them. Insights are thus characterised by a suddenness, selectivity, and variable delay of occurrence.

We study insight using the PSSST, a high temporal resolution insight task, described in detail in Chapter 1.

2.2 Results

To study insight-like learning dynamics, 99 participants performed the PSSST, which requires a binary choice about a stimulus characterised by two features. A circular array of coloured and moving dots (Rajananda et al., 2018) served as the stimulus (Fig. 3A,B). The motion coherence was varied in five levels such as to produce an accuracy gradient, while the dot colour was easy to recognise throughout (either orange or purple, see below). Participants had to learn the correct choice in response to each stimulus from trial-wise binary feedback (Fig. 3B), and were not instructed which features of the stimulus to pay attention to.

Participants first underwent an initial *training phase* with high motion coherence (2 blocks, 100 trials each), followed by a *motion phase* that marked the onset of motion coherence variability (2 blocks). During both phases, only the motion direction predicted the correct choice, while stimulus colour was random (see Fig. 3C). Without any announcement, stimulus colour became predictive of the correct response in the *motion and colour phase*, such that from then on both features could be used to determine choice (5 blocks, Fig. 3C). The experiment concluded with a questionnaire, in which participants were asked whether (1) they had noticed a rule, (2) how long it took them to notice it, (3) whether they had paid attention to colour during choice. The questionnaire was followed by an instruction block that served as a sanity check (see Methods).

Phases were not cued and or announced to participants. Previous work has shown that participants discover the hidden opportunity to use the stimulus colour that arises in the *motion and colour phase* through insight, as evidenced by sudden, delayed and selective behavioural changes that go hand in hand with gaining consciousness about the new regularity (Gaschler et al., 2019). This setup differs from traditional problem-solving tasks where participants actively seek solutions (Danek et al., 2014; Jung-Beeman et al., 2004; Kounios et al., 2006), but it closely aligns with the often used Number Reduction Task (NRT) (Wagner et al., 2004), where abrupt task improvements emerge through insights about uninstructed task elements, even though the task can in principle be performed using the initially learned/instructed strategy.

Baseline Task Performance (Training and Motion Phases)

Data from the *training phase*, during which motion directions were highly coherent but uncorrelated to colours (Block 1-2, dark grey tiles in Fig. 3C), showed that participants learned the response mapping for the four motion directions well (78% correct, t-test against chance: $t(98) = 30.8, p < .001$). In the following *motion phase*, noise was added to the motion, while the colour remained uncorrelated (blocks 3-4, grey tiles in Fig. 3C). This resulted in an accuracy gradient that depended on noise level (linear mixed effects model of accuracy: χ^2(1) = 726.36, $p < .001$; RTs: χ^2(1) = 365.07, $p < .001$; N = 99, Fig. 5A). Crucially, while performance in the condition with least noise was very high (91%), it was heavily diminished in the conditions with the largest amounts of motion noise, i.e. the two lowest coherence conditions, where participants only performed at only 60% and 63%, and did not change over time (paired t-test block 3 vs 4: $t(195.9) = -1.13$, $p = 0.3$, $d = 0.16$). Hence, substantial performance (improvements) in the high noise condition can only be attributed to colour use, rather than heightened motion sensitivity.

Task Performance After Correlation Onset (Motion and Colour Phase)

The noise level continued to influence performance in the *motion and colour phase*, as evidenced by a difference between performance in high vs. low coherence trials (20, 30 & 45% vs 5 & 10 % coherent motion, respectively; $M = 93 \pm 6\%$ vs $M = 77 \pm 12\%$; $t(140.9) = 12.5$, $p < .001$, $d = 1.78$, see Fig. 5A-B). Notably, however, the onset of the colour correlation triggered performance improvements across all coherence levels ($t(187.2) = -12.4$, $p < .001$, $d = 1.8$; end of *motion phase*: $M = 78 \pm 7\%$ vs. end of *motion and colour phase*: $M = 91 \pm 8\%$), contrasting the stable performance found during the motion phase and suggesting that at least some participants leveraged colour information once available.

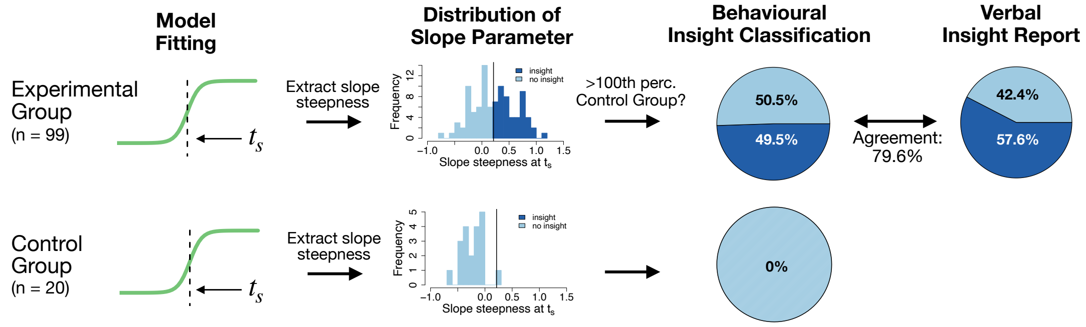


Figure 4: Insight classification procedure.

We fitted a sigmoid model to data from the lowest motion coherence condition data of both the Experimental Group and a Control Group (where colour never becomes predictive), and derived the distributions of the slope steepness at the estimated switch point (inflection point t_s of sigmoid function). We then asked which participants from the Experimental group had fitted slopes that were steeper than the 100th percentile of the Control Group. The resulting purely behavioural classification of insights agreed to 79.6% with verbal insight reports from a post-task questionnaire and predicted a number of behavioural features (see text). Importantly, using this method allowed us to apply the same procedure to neural networks.

2.2.1 Insight Classification

We asked whether these improvements are related to gaining conscious insight by analysing the post-experimental questionnaire. Results show that conscious knowledge about the colour regularity arose in some, but not all, participants: 57.6% (57/99) reported in the questionnaire to have used colour, while 42.4% indicated to not have noticed or used the colour. Hence insights were selective. As expected, the verbal report also related to performance differences in the lowest coherence condition in Block 8, where participants performed at 91.4% vs 68.7% if they had reported an insight vs not.

We next developed a behavioural classification of insight-like strategy switches. Previous work (Schuck et al., 2015; Gaschler et al., 2019; Schuck et al., 2022) has used a simple performance threshold of 75% in low coherence trials that also correlates highly with verbal insight reports in our sample (see above, $r(97) = .67$, $p < .001$). But this metric is not specifically related to suddenness, and therefore might identify participants who learned gradually about the colour. We therefore used a cross-validated approach in which we first identified the maximum suddenness than can occur by chance in a control experiment, and used this threshold for our main sample (Fig. 4). Details about the control experiment, in which participants (N=20) performed an identical task, except that colour never started to correlate with motion, and hence no insight was possible, can be found in the Methods.

To calculate suddenness, we first fitted a simple sigmoid function to participant accuracy with free parameters, slope m, inflection point t_s and function maximum y_{max} (details see Methods). We then calculated the rate of performance change at the inflection point (see Methods) of the fitted sigmoid function, and asked how many subjects from the experimental group had steeper behavioural transitions than the maximum value observed in the control group. This showed that about half of participants (49/99, 49.5%) had values larger than the 100% percentile of the control distribution (Fig. 4). Hence, in some participants behaviour changed so suddenly as to suggest insights (Fig. 5F). While this behavioural classification resulted in a more conservative estimate of the number of insight participants compared to the questionnaire, it highly overlapped with these verbal reports. Of the 49 participants classified as insight subjects 39 (79.6%) also self-reported to have used colour to make correct choices, see Fig. 4, 9A-B. This again correlates highly with the previously used performance threshold ($r(97) = .44$, $p < .001$). Hence, our behavioural marker of unexpectedly sudden performance changes can serve as a valid indicator for sudden insight.

Note that while in the above model fitting we used bins of 15 trials to identify suddenness, behavioural transitions often occurred within just a few trials (Fig. 9), i.e. on the order of 15-30 seconds, as is common for insights. The averaging was necessary in order to stabilise model fits. We also note that our choice of fitting function more generally was validated by group-wise model comparisons against a linear ramp (free parameters intercept y_0 and slope m) or a step function (free parameters inflection point t_s and function max y_{max}), BICs -6.7, -6.4 and -6.5, protected exceedance probabilities: 1, 0, 0, for sigmoid, linear and ramp models, respectively (see Fig. 5D-E, Fig. 6).

2.2.2 Behavioural Differences Between Participants With and Without Insights

We validated our behavioural metric of selectivity through additional analyses. Splitting participants into separate insight and no-insight groups based on the above procedure showed that, as expected based on the dependency of accuracy and our behavioural metric, insight subjects started to perform significantly better in the lowest coherence trials once the *motion and colour phase* (Fig. 5C) started, (mean proportion correct in *motion and colour phase*: $M = 83 \pm 10\%$), compared to participants without insight ($M = 66 \pm 8\%$) ($t(92) = 9.5$, $p < .001$, $d = 1.9$). Unsurprisingly, a difference in behavioural accuracy between insight participants and no-insight participants also held when the average across all coherence levels was considered ($M = 91 \pm 5\%$ vs. $M = 83 \pm 7\%$, respectively, t-test: $t(95.4) = 6.9$, $p < .001$, $d = 1.4$). Accuracy in the *motion phase*, which was not used in steepness fitting, did not differ between groups (low coherence trials: $M = 59\%$, vs. $M = 62\%$; $t(94.4) = -1.9$, $p = 0.07$, $d = 0.38$; all noise levels: $M = 76\%$ vs $M = 76\%$,$t(96) = 0.45$, $p = 0.7$, $d = 0.09$). Reaction times, which are independent from the choices used in model fitting and thus served as a sanity check for our behavioural metric split, reflected the same improvements upon switching to the colour strategy. Subjects who showed insight about the colour rule ($M = 748.47 \pm 171.1$ ms) were significantly faster ($t(96.9) = -4.9$, $p < .001$, $d = 0.97$) than subjects that did not ($M = 924.2 \pm 188.9$ ms) on low coherence trials, as well as over all noise levels ($t(97) = -3.8$, $p < .001$, $d = 0.87$) ($M = 675.7 \pm 133$ ms and

$M = 798.7 \pm 150.3$ ms, respectively).

2.2.3 Delay of Insights

Having established that behavioural changes were sudden and occurred for only some participants, we next asked whether insights occurred with random delays. To quantify this key characteristic, insight moments were defined as the time points of inflection of the fitted sigmoid function, i.e. when performance exhibited abrupt increases (see Methods). We verified the precision of our switch point identification by time-locking the data to the individually fitted switch points. This showed that accuracy steeply increased between the trial bins (50 trials) immediately before vs. after the switch, as expected ($M = 62\%$ vs $M = 83\%$ $t(89) = -11.2$, $p < .001$, $d = 2.34$, Fig. 5C, Fig. 16A). Additionally, reaction times dropped steeply from pre- to post-switch ($M = 971.63$ ms vs. $M = 818.77$ ms, $t(87) = 3.34$, $p < .001$, $d = 0.7$). The average delay of insight onset was 130 trials (± 95 trials), corresponding to 2.6 trial bins (Fig. 5G). The distribution of delays among insight participants ranged from 0 to 6 trial bins after the start of the *motion and colour phase*, and statistically did not differ from a uniform distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(49) = 0.25$, $p = 0.11$).

Hence, we find all characteristics of insight: sudden improvements in performance that occurred only in a subgroup and with variable delays.

Effects of Feedback and Low Coherence

Two possible objections to the idea that the above described insight-like strategy switches are internally generated and truly spontaneous could be made: first, the non-random nature of motion and feedback provided in the lowest coherence condition could mean that performance on these trials reflects heightened motion sensitivity coupled to reward guided choices, rather than the use of a colour strategy. Second, the presence of low coherence trials in and of itself might prompt participants to search for an alternative

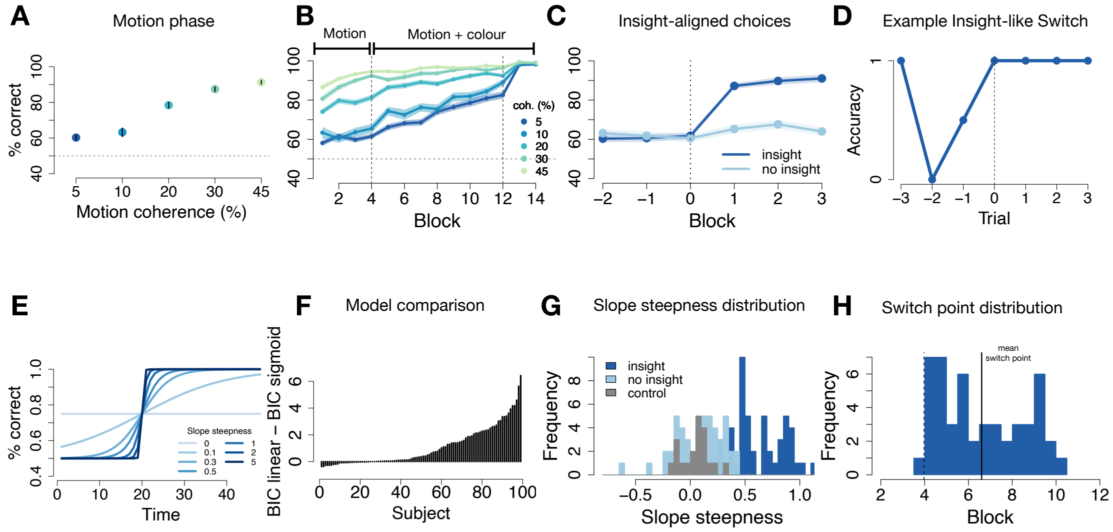


Figure 5: Task performance and insight-like strategy switches.

(A) Accuracy (% correct) during the *motion phase* increases with increasing motion coherence. N = 99, error bars signify standard error of the mean (SEM). **(B)** Accuracy (% correct) over the course of the experiment for all motion coherence levels. First dashed vertical line marks the onset of the colour predictiveness (*motion and colour phase*), second dashed vertical line the "instruction" about colour predictiveness. Blocks shown are halved task blocks (50 trials each). N = 99, error shadows signify SEM. **(C)** Switch point-aligned accuracy on lowest motion coherence level for insight (49/99) and no-insight (50/99) subjects. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **(D)** Trial-wise switch-aligned smoothed binary responses on lowest motion coherence level for an example insight subject. **(E)** Illustration of the sigmoid function for different slope steepness parameters. **(F)** Difference between BICs of the linear and sigmoid function for each human subject. N = 99. **(G)** Distributions of fitted slope steepness at inflection point parameter for control experiment and classified insight and no-insight groups. **(H)** Distribution of switch points. Dashed vertical line marks onset of colour predictiveness. Blocks shown are halved task blocks (50 trials each).

strategy. We dismiss these objections through two control experiments. The first (N=61) replicated the original task, but incorporated truly random stimuli in the lowest coherence condition (0% coherence) and replaced feedback with instructions. The second (N=29) introduced the two lowest motion coherence levels (5% and 10%) only in the sixth task block. As detailed in the SI, both experiments elicited a substantial number of insight-like switches, thus supporting the generality of the here reported phenomenon across a number of task variations (in line with Gaschler et al. Gaschler et al. (2019)).

2.3 Discussion

We investigated learning dynamics underlying insight using a binary decision making-task with a hidden regularity that entailed an alternative way to solve the task more efficiently. Jump-like learning that signified the sudden discovery of the hidden regularity describes insight moments – often deemed "mysterious" – boiled down to their simplest expression. Our results point towards three key characteristics of insight behaviour: suddenness, selectivity and delay.

It is worth emphasising that, like the commonly used Number Reduction Task (NRT) (Wagner et al., 2004), the PSSST does not mention the possibility of a hidden strategy in the instructions, thus differing from cognitive control paradigms involving deliberate attention switching tasks. One difference in task structure between the PSSST and the NRT is the onset of the hidden rule. In the NRT, the alternative strategy is present from the start, while in the PSSST, the *motion phase* persists for the first 45% of the trials. The rationale behind this is that we do not use explicit instructions in our paradigm, requiring participants to learn the feature relevance through trial-and-error learning. If colour was predictive from the start, subjects would start using it right away, since it is significantly easier than the noisy motion feature. In the NRT, participants are explicitly instructed with a certain task rule. Both tasks thus involve a phase of learning a certain rule at first that can be overcome by insight about different contingencies in the environment.

We note that, besides the NRT, our task differs from other common insight paradigms in one important aspect. Usually, subjects are actively trying to find a solution to a certain problem which then suddenly rises to consciousness after experiencing a period of being stuck (impasse) (Bowden et al., 2005). Here, in the PSSST, subjects learn a certain strategy and are unaware that task contingencies will change during the task, which then offers a hidden, more efficient strategy to solve the task that can be discovered through insight. However, the behavioural signature of insight in our task paradigm shares the fundamental characteristic of happening suddenly and after a variable delay. We regard the latter as analogous to impasse since subjects are blind to the alternative solution during that period. Based on these behavioural findings, we conclude that albeit differing from classic insight paradigms (Danek et al., 2014; Jung-Beeman et al., 2004; Kounios et al., 2006), the PSSST captures a special case of the general insight phenomenon. Further, insight has recently been conceived as a core cognitive mechanism of a unifying framework accounting for changes in mental representations, that reach beyond problem solving and include sub-fields as various as psychotherapy, psychedelic research and meditation (Tulver et al., 2023).

These results make an important contribution to the general understanding of learning dynamics and representation formation in environments with non-stationary feature relevance. Further, by defining three main insight characteristics that can be measured purely on a behavioural level, we provide metrics to investigate insight in humans as well as animals and even artificial agents, therefore fostering an integrative approach to understanding insight.

2.4 Methods

2.4.1 Task

Stimuli

To measure insight with high temporal resolution, we employed the PSSST. As described in Chapter 1, the task required a binary choice about circular arrays of moving dots (Rajananda et al., 2018). Dots were characterised by two features, (1) a motion direction (four possible orthogonal directions: NW, NE, SW, SE) and (2) a colour (orange or purple, Fig. 3A). The noise level of the motion feature was varied in 5 steps (5%, 10%, 20%, 30% or 45% coherent motion), making motion judgement relatively harder or easier. Colour difficulty was constant, thus consistently allowing easy identification of the stimulus colour. The condition with most noise (5% coherence) occurred slightly more frequently than the other conditions (30 trial per 100, vs 10, 20, 20, 20 for the other conditions).

The task was coded in JavaScript and made use of the jsPsych 6.1.0 plugins. Participants were restricted to use desktops (no tablets or mobile phones) of at least 13 inch width diagonally. Subjects were further restricted to use either a Firefox or Google Chrome browser to run the experiment.

On every trial, participants were presented a cloud of 200 moving dots with a radius of 7 pixels each. In order to avoid tracking of individual dots, dots had a lifetime of 10 frames before they were replaced. Within the circle shape of 400 pixel width, a single dot moved 6 pixel lengths in a given frame. Each dot was either designated to be coherent or incoherent and remained so throughout all frames in the display, whereby each incoherent dot followed a randomly designated alternative direction of motion.

The trial duration was 2000 ms and a response could be made at any point during that time window. After a response had been made via one of the two button presses, the white fixation cross at the centre of the stimulus would turn into a binary feedback symbol (happy or sad smiley) that would be displayed until the end of the trial (Fig. 3B). An inter trial interval (ITI) of either 400, 600, 800 or 1000 ms was randomly selected. If no response was made, a ”TOO SLOW” feedback was displayed for 300 ms before being replaced by the fixation cross for the remaining time of the ITI.

Task Design

For the first 400 trials, the *motion phase*, the correct binary choice was only related to stimulus motion (two directions each on a diagonal were mapped onto one choice), while the colour changed randomly from trial to trial (Fig. 3C). For the binary choice, participants were given two response keys, ”X” and ”M”. The NW and SE motion directions corresponded to a left key press (”X”), while NE and SW corresponded to a right key press (”M”) (Fig. 3A). Participants received trial-wise binary feedback (correct or incorrect), and therefore could learn which choice they had to make in response to which motion direction (Fig. 3B).

We did not specifically instruct participants to pay attention to the motion direction.

Instead, we instructed them to learn how to classify the moving dot clouds using the two response keys, so that they would maximise their number of correct choices. To ensure that participants would pick up on the motion relevance and the correct stimulus-response mapping, motion coherence was set to be at 100% in the first block (100 trials), meaning that all dots moved towards one coherent direction. Participants learned this mapping well and performed close to ceiling (87% correct, t-test against chance: $t(98) = 37.4$, $p < .001$). In the second task block, we introduced the lowest, and therefore easiest, three levels of motion noise (20%, 30% and 45% coherent motion), before starting to use all five noise levels in block 3. Since choices during this phase should become solely dependent on motion, they should be affected by the level of motion noise. We assessed how well participants had learned to discriminate the motion direction after the fourth block. Participants that did not reach an accuracy level of at least 85% in the three lowest motion noise levels during this last task block of the pretraining were excluded from the *motion and colour phase*. The 85% accuracy threshold on low motion noise trials was based on previous studies, where subjects' performance was at a comparable level (Schuck et al., 2015; Gaschler et al., 2013; Schuck et al., 2022). All subjects were notified before starting the experiment, that they could only advance to the second task phase (*motion and colour phase*, although this was not communicated to participants) if they performed well enough in the first phase and that they would be paid accordingly for either one or two completed task phases. Based on the early stopping of the experiment for participants who performed below threshold in the *motion phase*, a post-hoc change of the performance criterion is not possible. Results from other studies using this or a similar task imply that the reported insight phenomenon is not an artefact of the early stopping threshold (A. Löwe, Petzka, Tzegka, & Schuck, 2024; Schuck et al., 2015; Gaschler et al., 2019; Schuck et al., 2022).

After the *motion phase*, in the *motion and colour phase*, the colour feature became predictive of the correct choice in addition to the motion feature (Fig. 3C). This meant that each response key, and thus motion direction diagonal, was consistently paired with one colour, and that colour was fully predictive of the required choice. Orange henceforth corresponded to a correct ”X” key press and a NW/SE motion direction, while purple was predictive of a correct ”M” key press and NE/SW motion direction (Fig. 3A). This change in feature relevance was not announced to participants, and the task continued for another 400 trials as before - the only change being the predictiveness of colour.

Before the last task block we asked participants whether they 1) noticed a rule in the experiment, 2) how long it took until they noticed it, 3) whether they used the colour feature to make their choices and 4) to replicate the mapping between stimulus colour and motion directions. We then instructed them about the correct colour mapping and asked them to rely on colour for the last task block. This served as a proof that subjects were in principle able to do the task based on the colour feature and to show that, based on this easier task strategy, accuracy should be near ceiling for all participants in the last instructed block.

Participants

Participants between eighteen and 30 years of age were recruited online through Prolific.

Participation in the study was contingent on showing learning of the stimulus classification. Hence, to assess whether participants had learned to correctly identify motion directions of the moving dots, we probed their accuracy on the three easiest, least noisiest coherence levels in the last block of the uncorrelated task phase. If subjects reached an accuracy level of at least 85%, they were selected for participation in the experiment.

Ninety-six participants were excluded due to insufficient accuracy levels after the *motion phase* as described above. 99 participants learned to classify the dots’ motion direction, passed the accuracy criterion and completed both task phases. These subjects make up the final sample included in all analyses. Note that in previous studies where we did not apply such a criterion, similar spontaneous strategy switch behaviour was observed (Schuck et al., 2015; Gaschler et al., 2015; Allegra et al., 2020). 34 participants were excluded due to various technical problems or premature quitting of the experiment. All participants gave informed consent prior to beginning the experiment. The study protocol was approved by the local ethics committee of the Max Planck Institute for Human Development. Participants received 3£ for completing only the first task phase and 7£ for completing both task phases.

2.4.2 Modelling of Insight-like Switches

Models of Colour Use

In order to probe whether strategy switches in low coherence trials occurred abruptly, we compared three different models with different assumptions about the form of the data. First, we fitted a linear model with two free parameters:

$$y = m_t + y_0$$

where m is the slope, y_0 the y-intercept and t is time (here, task blocks)(Fig. 6). This model should fit no-insight participants' data well where colour use either increases linearly over the course of the experiment or stays at a constant level.

Contrasting the assumptions of the linear model, we next tested whether colour-based responses increased abruptly by fitting a step model with three free parameters, a switch point t_s, the step size s and a maximum value y_{max} (Fig. 6), so that

$$y = \begin{cases} y_{max} - s & \text{if } t < t_s \\ y_{max} & \text{if } t \geq t_s \end{cases}$$

We also included a sigmoid function with three free parameters as a smoother approximation of the step model:

$$y = \frac{y_{max} - y_{min}}{1 + e^{-m(t-t_s)}} + y_{min}$$

where y_{max} is the fitted maximum value of the function, m is the slope and t_s is the inflection point (Fig. 6). y_{min} was given by each individual's averaged accuracy on 5% motion coherence trials in block 3-4.

Comparing the model fits across all subjects using the Bayesian Information Criterion (BIC) and protected exceedance probabilities yielded a preference for the sigmoid function over both a step and linear model (Fig. 5E). On the one hand, this supports our hypothesis that insight-like strategy switches do not occur in an incremental linear fashion, but abruptly, with variance in the steepness of the switch. Secondly, this implies that at least a subset of subjects shows evidence for an insight-like strategy switch.

To investigate these insight-like strategy adaptations, we modelled participants' data using the individually fitted sigmoid functions (Fig. 7). The criterion we defined in order to assess whether a subject had switched to the colour strategy, was the slope at the inflection point, expressing how steep the performance jump was after having an insight about colour (Fig. 4). We obtained this value by taking the sigmoid function's partial derivative of time

$$\frac{\partial y}{\partial t} = (y_{max} - y_{min})\frac{me^{-m(t-t_s)}}{(1+e^{-m(t-t_s)})^2}$$

and then evaluating the above equation for the fitted switch point, $t = t_s$, which yields:

$$y'(t_s) = \frac{1}{4}m(y_{max} - y_{min})$$

Switch misclassifications can happen that are caused by irregularities and small jumps in the data - irrespective of a colour strategy switch. We therefore corrected for a general fit of the data to the model by subtracting the individually assessed general model fit from the slope steepness at the inflection point. Insight subjects were then classified as those participants whose corrected slope steepness at inflection point parameters were outside of the 100% percentile of a control group's (no change in predictiveness of colour) distribution of that same parameter (Fig. 4). By definition, insights about a colour rule cannot occur in this control condition, hence our derived out-of-sample distribution evidences abrupt strategy improvements hinting at insight (Fig. 22F).

Before the last task block we asked participants whether they used the colour feature to make their choices. 57.6% of participants indicated that they used colour to press correctly. The 49.5% insight participants we identified using our classification method overlapped to 79.6% with participants' self reports (Fig. 4).

2.5 Supplementary Information

Behavioural Effect Without Feedback

To exclude the possibility that insight-like behavioural switches were driven by the feedback signal acting as an external reward cue after the contingency change, we ran another task version without trial-wise feedback. In this experiment (N = 61), participants were presented with the same random dot motion stimuli, but motion coherence ranged from 0% to 50% in increments of 10. 0% coherence trials were thus truly random and could only be solved if participants had an insight about the predictiveness of the colour. Instead of learning feature relevance from trial-wise feedback, participants were here instructed about the stimulus-response mapping between the different dot motion directions and respective response keys and performed three training blocks (out of which the first two administered trial-wise feedback) before they started with the main task. During the main experiment, the only feedback participants received was their average accuracy at the end of each block. The onset of the *motion and colour phase* occurred after two task blocks of the main experiment.

We then used the same procedure of classifying insight-like behaviour by assessing in how many participants the steepness of the performance increase exceeded the chance level defined by the baseline distribution of steepness. About a fifth of participants (6/29, 20.1%) had steepness values larger than the 100% percentile of the control distribution. As expected, insight subjects also started to perform significantly better in 0% coherence trials once the *motion and colour phase* started, (mean proportion correct in *motion and colour phase*: $M = 75 \pm 6\%$), compared to participants without insight ($M = 56 \pm 5\%$) ($t(18.4) = 4.8$, $p < .001$, $d = 1.4$).

Late Onset of Difficult Trials

To further control whether the most difficult trials with the lowest motion coherence were driving insight-like behaviour, we ran another task version (N=29) of the same structure as described above, but delayed the onset of the two lowest motion coherence levels, 5% and 10%, until the sixth task block. In this task version subjects were also instructed about the motion relevance and received two training blocks, but were also presented with trial-wise feedback during the main experiment. There was no performance difference on 5% coherence trials in the last two task blocks between this and another task version with low coherence trials from the first block on ($M = 89 \pm 3\%$ vs $M = 89 \pm 3\%$, $t(54.1) = 0.04$, $p = 0.97$, $d = 0.01$), demonstrating that the majority of subjects had switched to using colour irrespective of difficult trials. These results are in line with earlier work by Gaschler et al. Gaschler et al. (2019) using a similar version of our task.

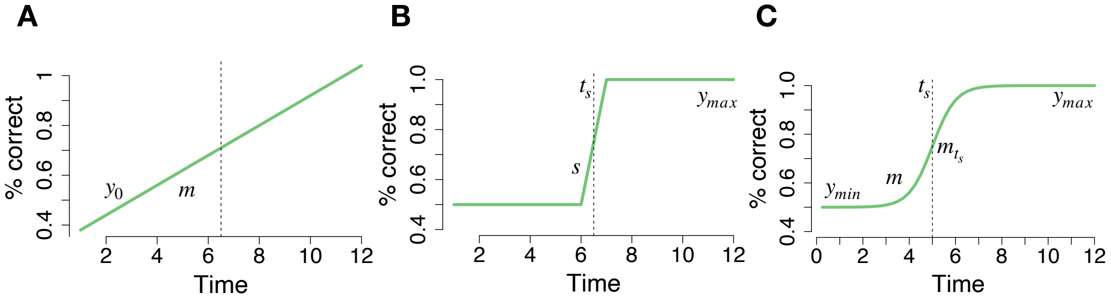
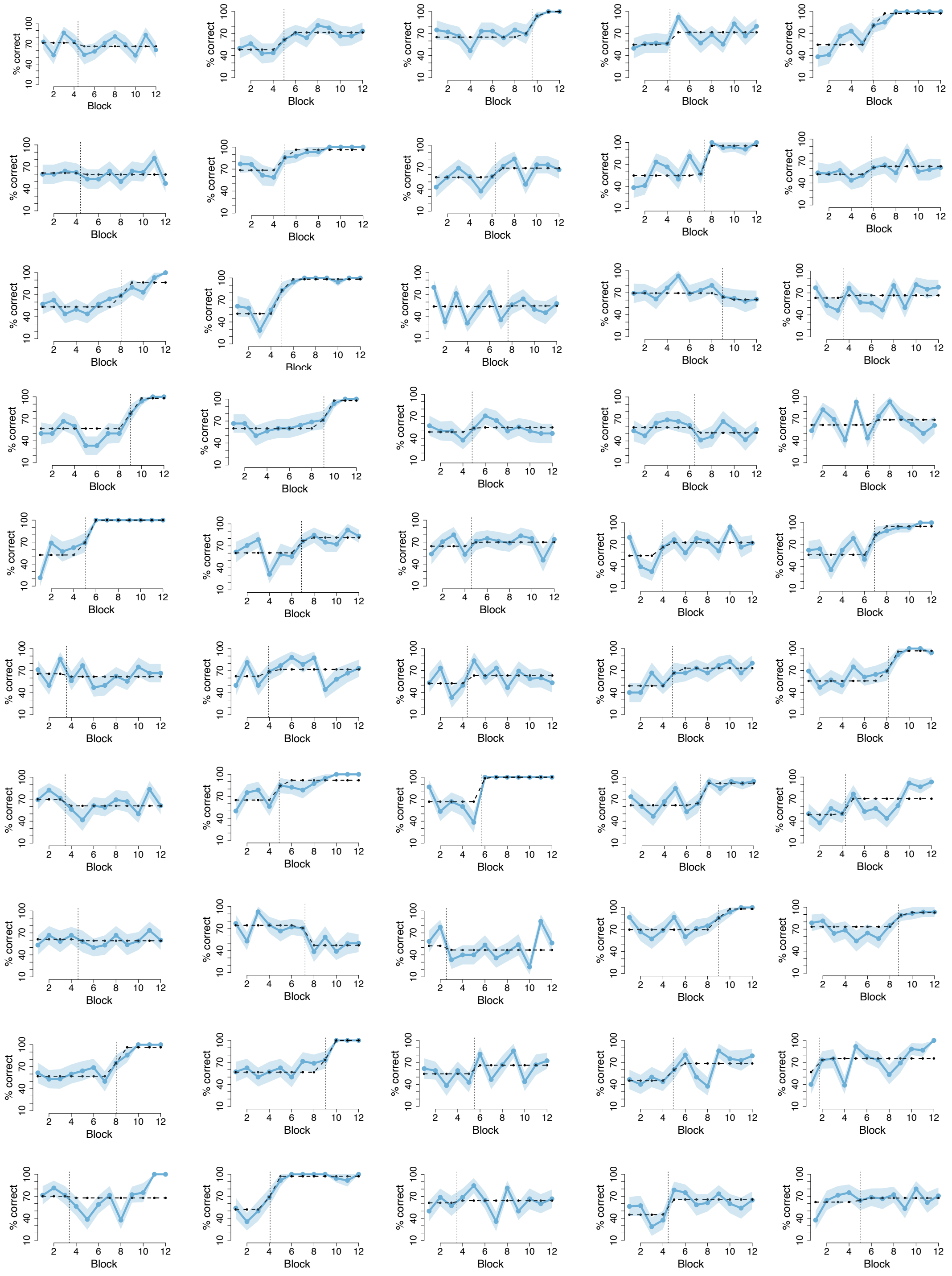


Figure 6: Illustrations of models and respective parameters.

(A) Linear function with free parameters intercept y_0 and slope m. **(B)** Step function with free parameters inflection point t_s and function maximum y_{max}. **(C)** Generalised logistic regression function with free parameters slope m, inflection point t_s and function maximum y_{max}.


% correct
10
40
70
100
Block
2
4
6
8
10
12

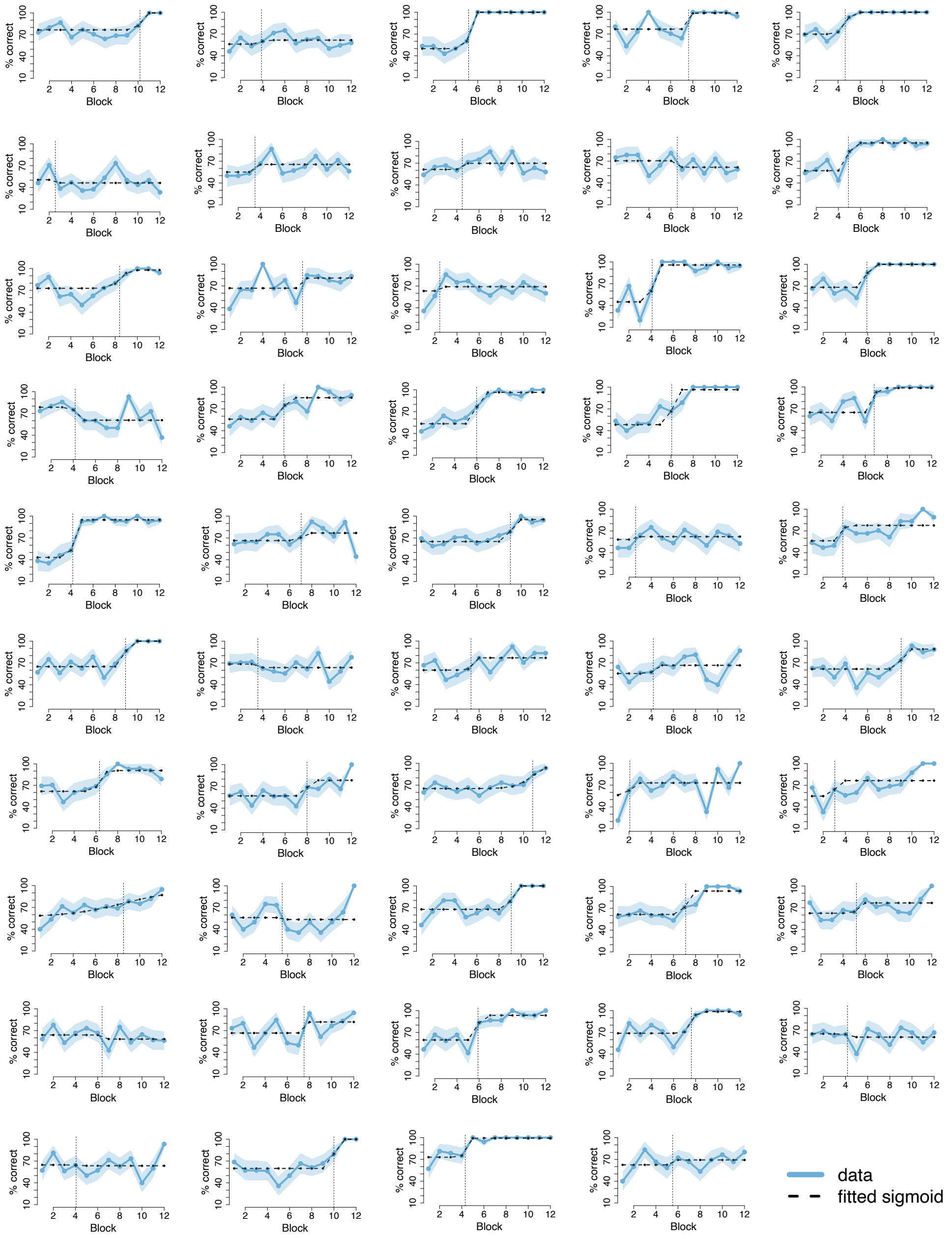


Figure 7: Performance and model predictions.

Performance on highest motion noise trials (blue) and model predictions (black) for every human participant. Blocks shown are halved task blocks (50 trials each). Error shadows signify SEM.

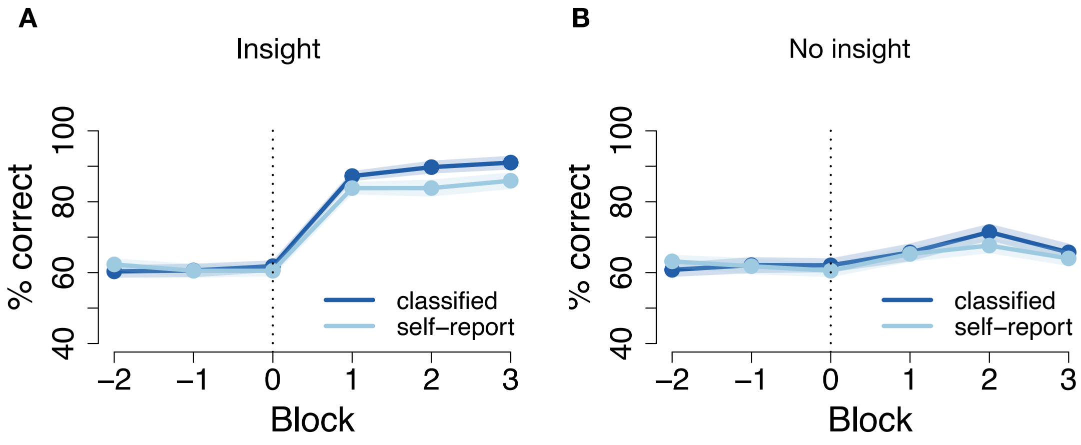


Figure 8: Switch-aligned performance and overlap between classification and self-reported colour use.

(A) Switch-aligned performance and overlap (39) between classified insight subjects (49/99) and self-reported colour use (57/99). **(B)** Switch-aligned performance and overlap (32) between classified no-insight subjects (50/99) and self-reported no colour use (42/99).

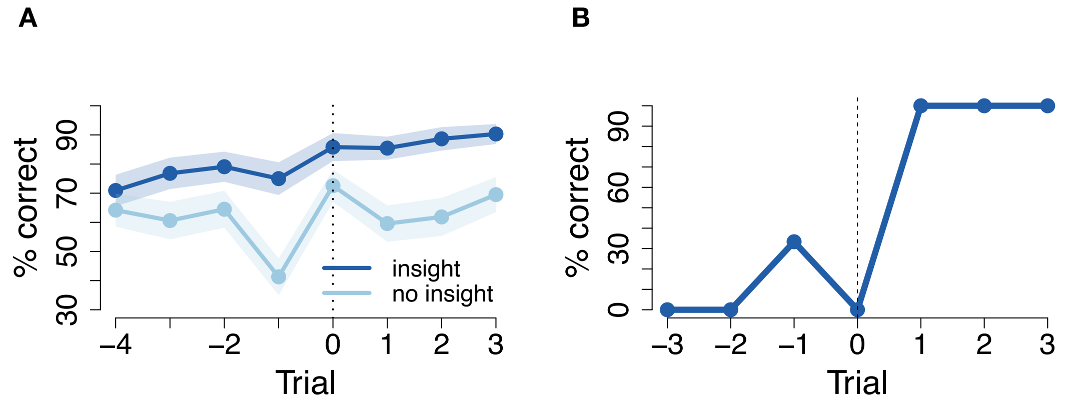


Figure 9: Trial-wise insight-like strategy improvements for 5% motion coherence trials

(A) Trial-wise switch-aligned performance between classified insight subjects (48/99) and no insight subjects (51/99). **(B)** Trial-wise switch-aligned performance for an example subject.

3 Insight in Neural Networks

Adapted from:

Löwe, A.T., Touzo, L.T., Muhle-Karbe, P.S., Saxe, A.M., Summerfield, C.* & Schuck, N.W.* (2024). *PLoS Computational Biology*.

Abstract

The previous chapter showed that humans sometimes have an insight that leads to sudden strategy adaptations and characterised these insight moments as happening selectively, abruptly and delayed. Here, we take a learning perspective and ask whether insight-like behaviour can occur in simple artificial neural networks, even when the models only learn to form input-output associations through gradual gradient descent. We compared learning dynamics in humans and regularised neural networks in a perceptual decision task that included a hidden regularity to solve the task more efficiently. We showed that only some humans discover this regularity, whose behaviour was marked by a sudden and abrupt strategy switch that reflects an Aha! moment. Notably, we find that simple neural networks with a gradual learning rule and a constant learning rate closely mimicked behavioural characteristics of human insight-like switches, exhibiting delay of insight, suddenness and selective occurrence in only some networks. Analyses of network architectures and learning dynamics revealed that insight-like behaviour crucially depended on a regularised gating mechanism and noise added to gradient updates, which allowed the networks to accumulate "silent knowledge" that is initially suppressed by regularised gating. This suggests that insight-like behaviour can arise from gradual learning in simple neural networks, where it reflects the combined influences of noise, gating and regularisation. These results have potential implications for more complex systems, such as the brain, and guide the way for future insight research.

This chapter was reprinted under a CC-BY 4.0 license. doi: . Author contributions: ATL, CS, PMK and NWS conceived the experiments. ATL carried out the experiments and analysed the data. ATL, AS and LT provided simulations. All authors discussed the results and contributed to the final manuscript.

3.1 Introduction

Neural networks trained with stochastic gradient descent (SGD) are a current theory of human learning that can account for a wide range of learning phenomena. At face value, SGD trained network models seem to imply that all learning is gradual. Yet, as discussed in Chapter 1, humans sometimes learn in an abrupt manner and improve on task in a seemingly spontaneous and abrupt manner.

The uniqueness of insight has led psychologists and neuroscientists to propose that insights are governed by a distinct learning or reasoning process (Jung-Beeman et al., 2004) that cannot be accounted for by common gradual theories of learning.

Here, we show that sudden and abrupt changes in behaviour in a task that elicits insights in humans can occur in a simple gradual learning system devoid of any dedicated insight mechanism. Our argument does not concern the subjective experiences related to insights, but focuses on showing how insight-like *behaviour* can emerge from gradual learning algorithms. Specifically, we aim to explain the following three behavioural characteristics discussed in Chapter 1: Insights are defined by a suddenness, selectivity, and variable delay of occurrence.

Current computational accounts of insight have proposed a number of specific mechanisms that operate in parallel to gradual learning and cause abrupt strategy changes linked to Aha! moments. One interesting model that can dynamically switch between strategies was reported by Collins and colleagues (Collins & Koechlin, 2012; Donoso, Collins, & Koechlin, 2014). The model is able to maintain multiple concurrent behavioural strategies, switch between them based on current task requirements, and devise new strategies. Other models have focused on representational restructuring (Kralik, Mao, Cheng, & Ray, 2016), integration of explicit and implicit knowledge (Hélie & Sun, 2010) or metacognitive monitoring (Dubey, Ho, Mehta, & Griffiths, 2021). Our work introduces a much simpler but unified model where both gradual and insight-like learning stem from a singular, delta-rule-based algorithm. Our model distinguishes itself from previous approaches by utilising this single updating rule before, during and after strategy switches, eliminating the need for dedicated strategy monitoring, maintenance, or multiple memory systems. As we will show below, our model also does not require a heightened occurrence of errors or low rewards to switch strategy. To emphasise our theoretical point, we focus on the most concise model that can exhibit insight-like behaviour. This is achieved through a simple neural network comprising merely two input nodes and one output node, whereby each input node is regulated by a single multiplicative gate that modulates its respective weight on the output. During learning, gates are L1-regularised and noise is added to the gradients. Stripped down to such minimal assumptions, our model demonstrates how insight-like behaviour can theoretically emerge from a system devoid of complex mechanisms such as restructuring. Furthermore, we show how our model qualitatively and quantitatively aligns with human behaviour across the three insight dimensions suddenness, selectivity and delay.

The idea that insight-like behaviour can arise from gradual learning is supported by previous work on human behaviour (Durso, Rea, & Dayton, 1994) and neural networks trained with gradient descent (Power, Burda, Edwards, Babuschkin, & Misra, 2022). Saxe and colleagues (Saxe, McClelland, & Ganguli, 2014), for instance, have shown that non-linear learning dynamics, i.e. suddenness in the form of saddle points and stage-like transitions, can result from gradient descent even in linear neural networks, which could explain sudden behavioural improvements. Other work has shown a delayed or stage-like mode of learning in neural networks that is reminiscent of the period of impasse observed in humans, reflecting for instance the structure of the input data (Saxe, McClelland, & Ganguli, 2019; Schapiro & McClelland, 2009; McClelland & Rogers, 2003), or information compression of features that at some point seemed task-irrelevant (Flesch, Juechems, Dumbalska, Saxe, & Summerfield, 2022; Saxe, Bansal, et al., 2019). Finally, previous work has also found substantial individual differences between neural network instances that are induced by random differences in weight initialisation, noise, or the order of training examples (Bengio, Louradour, Collobert, & Weston, 2009; Flesch, Balaguer, Dekker, Nili, & Summerfield, 2018), which can become larger with training (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020). Notably, while different behavioural aspects of sudden and abrupt strategy switches have been shown, so far no study has made a detailed comparison to behaviour in humans and specifically asked whether delay, suddenness and selectivity can occur jointly in a single network model.

Two factors that influence discontinuities in learning in neural networks are regularisation and gating. A simple mechanism to attain regularisation involves adding a penalty term to the error function that prevents coefficients from reaching large values, and which thereby leads to suppression of input features (Bishop, 2006). While these forms of explicit regularisation share similarities with other (implicit) regularisation techniques, these two forms are not identical and we focus only on the former. From a cognitive neuroscience perspective, regularisation may correspond to mechanisms that limit the number of factors that are taken into account during decision making, reminiscent of the effects of priors or attentional mechanisms on human cognition (Parpart, Jones, & Love, 2018). Crucially, while regularisation is useful in that it avoids overfitting or getting stuck in a local minimum (Liu, Papailiopoulos, & Achlioptas, 2020), the lingering suppression of some inputs might also cause above-mentioned “blindness” to a solution. The second factor – gating – describes a multiplicative interaction between learnable parameters and is ubiquitously found in real neurons whose activity is often gain modulated (S. J. Mitchell & Silver, 2003). It is known that such multiplicative interactions cause exponential transitions in learning, which are for instance widely used in multiplicative dynamical systems like the logistic growth model or recurrent neural networks. Hence, regularisation and gating are both commonly used in artificial neural networks (Bishop, 2006; Krishnamurthy, Can, & Schwab, 2022; Jozefowicz, Zaremba, & Sutskever, 2015), and are inspired by biological brains (Groschner, Malis, Zuidinga, & Borst, 2022; Poggio, Torre, & Koch, 1985; Costa, Assael, Shillingford, De Freitas, & Vogels, 2017). This makes regularisation and gating natural candidate aspects of network structure and training that could be related to insight-like behaviour, as evidenced by a temporary impasse followed by a sudden performance change.

A simple neural network architecture with multiplicative gates and L1-regularisation served as our candidate model. We focused specifically on L1-regularisation – which penalises the loss by the absolute rather than quadratic gate values – as it forces gates of irrelevant inputs most strongly towards 0, causing a sustained suppression period before the fast transition, similar to the impasse observed in humans. Our model is meant to be conceptual and not biologically realistic, as we aim to demonstrate how insight-like behaviour and fast representational restructuring can occur in simple architectures. We also show that our findings generalise to more complex neural networks with multiple input nodes and hidden layers which show qualitatively similar behaviour to what is described below.

We study insight-like strategy switches using an adapted, numerical version of the PSSST.

3.2 Results

To study insight-like learning dynamics, 99 neural networks performed the PSSST (Fig. 3). While for human participants a circular array of coloured and moving dots (Rajananda et al., 2018) served as the stimulus in humans, two scalar inputs represented the stimulus features symbolically for networks (Fig. 10A). In humans, the motion coherence was varied in five levels such as to produce an accuracy gradient, while the dot colour was easy to recognise throughout. The network inputs for motion were set such that for each human one network performed with the same behavioural accuracy in the different coherence levels (see below for details). Like human participants, networks had to learn the correct choice in response to each stimulus from trial-wise binary feedback.

For networks, the initial *training phase* with high motion coherence was extended to 6 blocks (2 blocks, 100 trials each in humans), followed by a *motion phase* that marked the onset of motion coherence variability (2 blocks in humans and networks). During both phases, only the motion direction predicted the correct choice, while stimulus colour was random (see Fig. 10B). Without any announcement, stimulus colour became predictive of the correct response in the *motion and colour phase*, such that from then on both features could be used to determine choice (5 blocks for humans and networks, Fig. 10B).

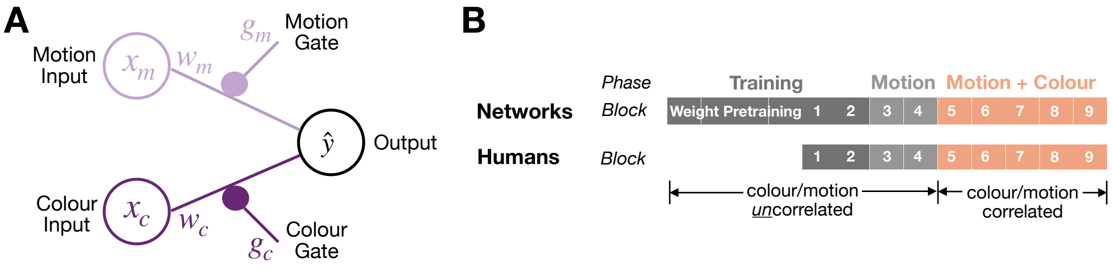


Figure 10: Network architecture and task design.

(A) Schematic of the neural network with regularised gate modulation used to model insights. **(B)** Task structure of the two-alternative forced choice task for neural networks compared to humans: each block consisted of 100 trials. Neural networks have an extended Weight Pretraining phase of 6 blocks in total. Otherwise, the task structure is the same as for human participants.).

To probe whether insight-like behaviour can arise in simple neural networks trained with gradient descent, we simulated 99 network models performing the same decision making task as 99 human subjects.

Architecture

The networks had two input nodes (x_c, x_m, for colour and motion, respectively), two input-specific gates (g_m, g_c) and weights (w_m, w_c), and one output node ($\hat{y}$, Fig. 10A). Network weights and their respective gates were initialised at 0.01. Gates only differ from the weights by the applied regularisation. The network multiplied each input node by two parameters, a corresponding weight, and a gate, and returned a decision based on the output node's sign $\hat{y}$:

$$\hat{y} = \operatorname{sign}(g_m w_m x_m + g_c w_c x_c + \eta) \tag{1}$$

where $\eta \sim \mathcal{N}(0, \sigma = 0.05)$ is Gaussian noise, and weights and gates are the parameters learned online through gradient descent.

We note that our architecture is functionally equivalent to a 2-layer diagonal linear network with L1-regularisation on the second layer. Hence, although we focus on the gating interpretation, our results can equally be interpreted as pertaining to a particular simple form of linear 2-layer neural networks.

Learning Algorithm

To train L1-networks we used a simple squared loss function with L1-regularisation of gate weights:

$$\mathcal{L} = \frac{1}{2}(g_m w_m x_m + g_c w_c x_c + \eta - y)^2 + \lambda(|g_m| + |g_c|) \tag{2}$$

with a fixed level of regularisation $\lambda = 0.07$ and a fixed learning rate of $\alpha = 0.6$. $\mathcal{L}$ was minimised by updating weights and gates after trial, while adding Gaussian noise $\xi \sim \mathcal{N}(\mu_\xi = 0, \sigma_\xi = 0.05)$ to each gradient update to mimic learning noise and induce variability between individual networks (same gradient noise level for all networks). This yielded the following trialwise update equations for noisy SGD of the network's weights given the correct decision y:

$$\Delta w_m = -\, \alpha x_m g_m (x_m g_m w_m + x_c g_c w_c + \eta - y) + \xi_{w_m}, \tag{3}$$

and gates,

$$\Delta g_m = -\, \alpha x_m w_m (x_m g_m w_m + x_c g_c w_c + \eta - y) - \alpha\lambda\, \mathrm{sign}(g_m) + \xi_{g_m} \tag{4}$$

where we have not notated the dependence of all quantities on trial index t for clarity; and analogous equations hold for colour weights and gates with all noise factors ξ_{g_m}, ξ_{w_m} etc, following the same distribution.

We used this setup, L1-regularisation of multiplicative gate weights, for two reasons: First, by adding a penalty term to the error function, regularisation leads to a suppression of (irrelevant) input features, which we reasoned would introduce competitive dynamics between the input channels. These competitive dynamics mimic a simple 'attentional' gating, and can lead to non-linear learning dynamics. We used L1- rather than L2 regularisation because it adds an absolute rather than squared penalty that forces gates of irrelevant inputs most strongly towards 0, compared to L2-regularisation, which is less aggressive in particular once gates are already very small (which we also show empirically in the Supplementary Material). Second, we implemented multiplicative weights and gates because they induce non-linear quadratic and cubic gradient dynamics. Applying L1-regularisation to the gates then will lead to a sustained suppression period before the fast transition (see Methods for details).

Network Training

The stimulus features motion and colour were reduced to one input node each, which encoded colour/motion direction of each trial by taking on either a positive or a negative value. More precisely, given the correct decision $y = \pm 1$, the activities of the input nodes were sampled from i.i.d. normal distributions with means $\pm M_m$ and $\pm M_c$ and standard deviations $\sigma_m = 0.01$ and $\sigma_c = 0.01$ for colour and motion respectively. Hence M_m and M_c determine the signal to noise ratio in each input. While σ models perceptual noise, signal to noise ratio, and therefore the difficulty, is determined by the ratio of the means M_m and M_c to sigma. Hence, different difficulty levels in the motion inputs could have equivalently been modelled as changing variances in the presence of a constant mean. In order to match human performance, we fixed the colour mean shift M_c to 0.22, while the mean shifts of the motion node differed by noise level and were fitted individually such that each human participant had one matched network with comparable pre-insight task accuracy in each motion noise condition (see below).

Networks received an extended pre-task training phase of 6 blocks, but then underwent a training curriculum precisely matched to the human task (2 blocks of 100 trials in the *motion phase* and 5 blocks in the *motion and colour phase*, see Fig. 10B). We adjusted direction specificity of motion inputs (i.e. difference in distribution means from which x_m was drawn for left vs right trials) separately for each participant and coherence condition, such that performance in the motion phase was equated between each pair of human and network (Fig. 11A, see Methods). Moreover, the colour and motion input sequences used for network training were sampled from the same ten input sequences that humans were exposed to. The learning rate of $\alpha = 0.6$ (same for all participants) was selected to match average learning speed.

Deep Neural Networks

We also trained a simple deep neural network on the same task. Briefly, this network also had 2 inputs, x_m and x_c, 48 units in a hidden layer, and two outputs $\hat{y}$. The activation function was ReLu and each weight connecting the inputs with a hidden unit had one associated multiplicative gate g, where we again applied L1-regularisation on the gate weights g. The network was trained on the Cross Entropy loss using stochastic gradient descent with λ = 0.002 and α = 0.1. As for the one-layer network, we trained this network on a curriculum precisely matched to the human task, and adjusted hyperparameters (noise levels) as described above, such that baseline network performance and learning speed were carefully equated between humans and simple deep neural networks as well.

3.2.1 Behaviour of L1-regularised Neural Networks

Overall Network Performance

Networks learned the motion direction-response mapping well in the training phase, during which colour inputs changed randomly and output should therefore depend only on motion inputs (75% correct, t-test against chance: $t(98) = 33.1$, $p < .001$, the accuracy of humans in this phase was $M = 76 \pm 6\%$). As in humans, adding noise to the motion inputs (*motion phase*) resulted in an accuracy gradient that depended on noise level (linear mixed effects model of accuracy: χ^2(1) = 165.61, $p < .001$; N = 99, Fig. 11A), as expected given that input distributions were set such that network performance would equate to human accuracy (Fig. 11A-B). Networks also exhibited low and relatively stable performance levels in the two lowest coherence conditions (58% and 60%, paired t-test to assess stability in the *motion phase*: $t(98) = -0.7$, $p = 0.49$,

$d = 0.02$), and had a large performance difference between high vs low coherence trials ($M = 88\%{\pm}6\%$ vs. $M = 74{\pm}13\%$, $t(137.3) = 9.6$, $p < .001$, $d = 1.36$ for high, i.e. $\geq$ 20% coherence, vs. low trials). Finally, networks also performed comparably well to humans at the end of learning (last block of the *colour and motion phase*: $M(nets) = 79\% \pm 17\%$ vs. $M(humans) = 82 \pm 17\%$, $t(195.8) = 1.1$, $p = 0.27$, $d = 0.16$, Fig. 19), suggesting that at least some networks did start to use colour inputs. Hence, networks' baseline performance and learning were successfully matched to humans.

Insight-like Behavioural Characteristics of Network Behaviour

To look for characteristics of insight in network performance, we employed the same approach used for modelling human behaviour (Fig. 4), and investigated suddenness, selectivity, and delay. To identify sudden performance improvements, we fitted each network's time course of accuracy on low coherence trials by (1) a linear model and (2) a non-linear sigmoid function, which would indicate gradual performance increases or insight-like behaviour, respectively. As in humans, network performance on low coherence trials was best fit by a non-linear sigmoid function, indicating at least a subsection of putative "insight networks" (BIC sigmoid function: $M = -10$, $SD = 1.9$, protected exceedance probability: 1, BIC linear function: $M = -9$, $SD = 2.4$, protected exceedance probability: 0)(Fig. 11D).

We then tested whether insight-like behaviour occurred only in a subset of networks (selectivity) by assessing in how many networks the steepness of the performance increase exceeded a chance level defined by a baseline distribution of the steepness. As in humans, we ran simulations of 99 control networks with the same architecture, which were trained on the same task except that during the *motion and colour phase*, the two inputs remained uncorrelated. About half of networks (46/99, 46.5%) had steepness values larger than the 100% percentile of the control distribution, closely matching the value we observed in the human sample. The L1-networks that showed sudden performance improvements were not matched to insight humans more often than chance (χ^2(47) = 27.9, $p = 0.99$), suggesting that network variability did not originate from baseline performance levels or trial orders. Hence, a random subset of networks showed sudden performance improvements comparable to those observed during insight moments in humans (Fig. 11E).

For simplicity reasons in comparing network behaviour to humans, we will refer to the two groups as “insight and no-insight networks”. Analysing behaviour separately for the insight and no-insight networks showed that switches to the colour strategy improved the networks’ performance on the lowest coherence trials once the *motion and colour phase* started, as compared to networks that did not show a strategy shift ($M = 83 \pm 11\%$, vs. $M = 64 \pm 9\%$, respectively, $t(89.8) = 9.2$, $p < .001$, $d = 1.9$, see Fig. 11C). The same performance difference between insight and no-insight networks applied when all coherence levels of the *motion and colour phase* were included ($M = 88 \pm 7\%$ vs. $M = 77 \pm 6\%$, $t(93.4) = 7.8$, $p < .001$, $d = 1.57$). Unexpectedly, insight networks performed slightly worse on low coherence trials in the motion phase, i.e. before the change in predictiveness of the features, ($t(97) = -3.1$, $p = 0.003$, $d = 0.62$) (insight networks: $M = 58 \pm 8\%$; no-insight networks: $M = 64 \pm 9\%$), and in contrast to the lack of pre-insight differences we found in humans.

Finally we asked whether insight-like behaviour occurred with random delays in neural networks, again scrutinising the time points of inflection of the fitted sigmoid function, i.e. when performance exhibited abrupt increases (see Methods). Time-locking the data to these individually fitted switch points verified that, as in humans, the insight-like performance increase was particularly evident around the switch points: accuracy was significantly increased between the halved task blocks preceding and following the insight-like behavioural switch, for colour switching networks ($M = 66 \pm 8\%$ vs. $M = 86 \pm 7\%$, $t(91.6) = -12.7$, $p < .001$, $d = 2.6$, see Fig. 11C, Fig. 16B).

Among insight networks, the delay distribution ranged from 2 to 8 trial bins after the start of the *motion and colour phase*, and did not differ from a uniform distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(46) = 0.13$, $p = 0.85$). The average delay of insight-like switches was 3.5 trial bins (± 1.05), corresponding to 175 trials (Fig. 11F). The insight networks’ delay was thus slightly longer than for humans ($M = 130 \pm 95$ trials vs. $M = 175 \pm 105$ trials, $t(92.7) = -2.1$, $p = 0.04$, $d = 0.42$). The variance of insight-like strategy switch onsets as well as the relative variance in the abruptness of the switch onsets thus qualitatively matched our behavioural results observed in human participants. The behaviour of L1-regularised neural networks therefore showed all characteristics of human insight: sudden improvements in performance that occurred selectively only in a subgroup with variable random delays.

We investigated whether this effect was specific to the form of regularisation and found that neither L2-regularisation nor non-regularised networks showed the insight key behavioural characteristics of selectivity and delay (see Supplementary Material).

Spontaneous Strategy Switches in Deep Neural Networks

Analysing the results from the models with a more complex network architecture revealed qualitatively similar results. When we applied L1-regularisation with a regularisation parameter of $\lambda = 0.002$ on the gate weights of hidden layer networks, 18.2% of the networks exhibited *abrupt* and *delayed* learning dynamics, resembling insight-like behaviour in humans (Fig. 14). Insight-like switches to the colour strategy thereby again improved the networks' performance significantly. We also observed a wide distribution of delays, for L1-regularised networks with a hidden layer (Fig. 14C-D). Taken together, these results from simple deep neural networks mirror our observations from simulations with a simplified setup. We can thus confirm that our results of L1-regularised neural networks' behaviour exhibiting all key characteristics of human insight-like spontaneous switches (suddenness, selectivity and delay) are not an artefact of the one-layer linearity.

Origins of Insight-like Behaviour in Neural Networks

Having established the behavioural similarity between L1-networks and humans in an insight task, we asked what gave rise to insight-like switches in some networks, but not

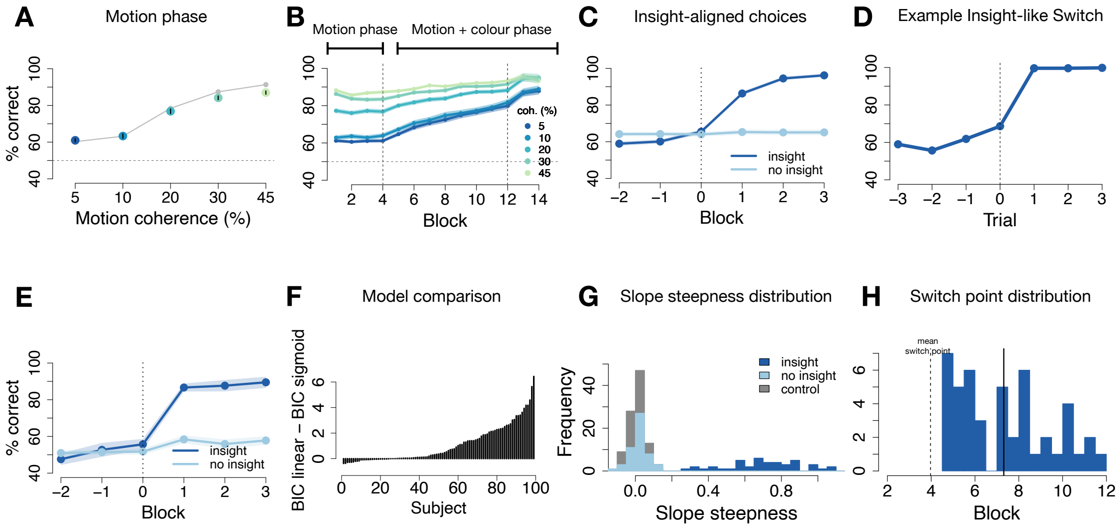


Figure 11: L1-regularised neural networks: task performance and insight-like strategy switches.

(A) Accuracy (% correct) during the *motion phase* increases with increasing motion coherence. N = 99, error bars signify SEM. Grey line is human data for comparison. **(B)** Accuracy (% correct) over the course of the experiment for all motion coherence levels. First dashed vertical line marks the onset of the colour predictiveness (*motion and colour phase*), second dashed vertical line the "instruction" about colour predictiveness. Blocks shown are halved task blocks (50 trials each). N = 99, error shadows signify SEM. **(C)** Switch point-aligned accuracy on lowest motion coherence level for insight (46/99) and no-insight (53/99) networks. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **(D)** Trial-wise switch-aligned continuous outputs on lowest motion coherence level for an example insight network. **(E)** Switch point-aligned accuracy on lowest motion coherence level for insight (18/99) and no-insight (81/99) hidden layer networks. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **(F)** Difference between BICs of the linear model and sigmoid function for each network. **(G)** Distributions of fitted slope steepness at inflection point parameter for control networks and classified insight and no-insight groups. **(H)** Distribution of switch points. Dashed vertical line marks onset of colour predictiveness. Blocks shown are halved task blocks (50 trials each).

others. We therefore investigated the dynamics of gate weights and the effects of noise in insight vs. no-insight networks, and the role of regularisation strength parameter λ.

Insight Networks Immediately Learn More About Colour Once It Becomes Predictive

Our first question was how learning about stimulus colour differed between insight and no-insight L1-networks, as expressed by the dynamics of network gradients. We time-locked the time courses of gradients to each network's individual switch point. Right when the switch occurred (at t of the estimated switch), colour gate weight gradients were significantly larger in insight compared to no-insight L1-networks ($M = 0.06 \pm 0.06$ vs. $M = 0.02 \pm 0.03$, $t(73.2) = 5.1$, $p < .001$, $d = 1.05$), while this was not true for motion gate weight gradients ($M = 0.18 \pm 0.16$ vs. $M = 0.16 \pm 0.16$, $t(97) = 0.7$, $p = 0.5$, $d = 0.13$).

Notably, insight networks had larger colour gate weight gradients even before any behavioural changes were apparent, right at the beginning of the *motion and colour phase* (first 5 trials of *motion and colour phase*: $M = 0.05 \pm 0.07$ vs. $M = 0.01 \pm 0.01$; $t(320) = 8.7$, $p < .001$), whereas motion gradients did not differ ($t(576.5) = -0.1$, $p = 0.95$). This increase in colour gate weight gradients for insight networks happened within a few trials after correlation onset (colour gradient last trial of *motion phase*: $M = 0 \pm 0$ vs. 5th trial of *motion and colour phase*: $M = 0.06 \pm 0.08$; $t(47) = -5.6$, $p < .001$, $d = 1.13$), and suggests that insight networks start early to silently learn more about colour inputs compared to their no-insight counterparts. A change point analysis considering the mean and variance of the gradients confirmed the onset of the *motion and colour phase* to be the change point of the colour gradient mean, with a difference of 0.04 between the consecutive pre-change and change time points for insight networks vs 0.005 for no-insight networks (with a change point detected two trials later), indicating considerable learning about colour for insight networks.

3.2.2 "Silent" Colour Knowledge Precedes Insight-like Behaviour

A core feature of our network architecture is that inputs were multiplied by two factors, a gate g, and a weight w, but only gates were regularised. This meant that some networks might have developed larger colour weights, but still showed no signs of colour use, because the gates were very small. This could explain the early differences in gradients reported above. To test this idea, we investigated the absolute size of colour gates and weights of insight vs no-insight L1-networks before and after insight-like switches had occurred.

Comparing gates at the start of learning (first trial of the *motion and colour phase*), there were no differences between insight and no-insight networks for either motion or colour gates (colour gates: $M = 0 \pm 0.01$ vs. $M = 0 \pm 0.01$; $t(95.3) = 0.8$, $p = 0.44$, motion gates: $M = 0.5 \pm 0.3$ vs. $M = 0.6 \pm 0.3$; $t(93.1) = -1.7$, $p = 0.09$, see Fig. 12A, Fig. 12F,H). Around the individually fitted switch points, however, the gates of insight and no-insight networks differed only for colour gates (colour gates: 0.2 ± 0.2 vs 0.01 ± 0.02 for insight vs no-insight networks, $t(48) = 6.7$, $p < 0.001$, $d = 1.4$, motion gates: 0.5 ± 0.3 vs 0.5 ± 0.3 for insight vs no-insight networks, $t(95.6) = 0.2$, $p = 0.9$, $d = 0.04$). Insight networks' increased use of colour inputs was particularly evident at the end of learning (last trial of the *motion and colour phase*) and reflected in larger colour gates (0.7 ± 0.3 vs 0.07 ± 0.2 for insight vs no-insight networks, $t(73.7) = 13.4$, $p < 0.001$, $d = 2.7$) while the reverse was true for motion gates ($M = 0.2 \pm 0.2$ vs $M = 0.5 \pm 0.3$, respectively, $t(81) = -7.5$, $p < 0.001$, $d = 1.5$, see Fig. 12B, Fig. 12F,H). Hence, differences in gating between network subgroups were only present after, but not before learning, and did not explain the above reported gradient differences or which network would show insight-like behaviour.

A different pattern emerged when investigating the weights of the networks. Among insight networks colour weights were significantly larger already at the start of learning (first trial of the *motion and colour phase*), as compared to no-insight networks (insight: $M = 1.2 \pm 0.6$; no-insight: $M = 0.4 \pm 0.3$, $t(66.2) = 8.1$, $p < .001$, $d = 1.7$, see Fig. 12C, Fig. 12E,G). This was not true for motion weights (insight: $M = 3.4 \pm 0.7$; no-insight: $M = 3.5 \pm 0.5$, $t(89.5) = -1.1$, $p = 0.3$, $d = 0.2$, see Fig. 12C, Fig. 12E,G). Thus, colour information appeared to be encoded in the weights of insight networks already before any insight-like switches occurred. Because the colour gates were suppressed through the L1-regularisation mechanism before learning, the networks did not differ in any observable colour sensitivity. An increase of colour gates reported above could then unlock the “silent knowlegde” of colour relevance.

To experimentally test the effect of pre-learning colour weights, we ran a new sample of L1-networks ($N = 99$), and adjusted the colour and motion weight of each respective network to the mean absolute colour and motion weight size we observed in insight networks at start of learning (first trial of *motion and colour phase*). Gates were left untouched. This increased the number of insight networks from 46.5% to 70.7%, confirming that encoding of colour information at an early stage was an important factor for later switches, but also not sufficient to cause insight-like behaviour in all networks. Note that before weights adjustments were made, the performance of the new networks did not differ from the original L1-networks ($M = 0.8 \pm 0.07$ vs $M = 0.8 \pm 0.07$, $t(195) = 0.2$, $p = 0.9$, $d = 0.03$). In our new sample, networks that would later show insight-like behaviour or not also did not differ from each other (insight: $M = 0.7 \pm 0.07$ vs $M = 0.7 \pm 0.07$, $t(100.9) = 1.4$, $p = 0.2$, $d = 0.3$, no-insight: $M = 0.8 \pm 0.05$ vs $M = 0.8 \pm 0.07$, $t(71) = 0.9$, $p = 0.4$, $d = 0.2$). Weight and gate differences between L1- and L2-networks are reported in the Supplementary Material (see Fig. 18).

3.2.3 Networks Need Noise For Insight-like Behaviour

One possible factor that could explain the early differences between the weights of network subgroups is noise. The networks were exposed to noise at two levels: on each trial noise was added at the output stage ($\eta \sim \mathcal{N}(0,\, \sigma_\eta^2)$), and to the gate and weight gradients during updating ($\xi \sim \mathcal{N}(0,\, \sigma_\xi^2)$).

We probed whether varying the level of noise added during gradient updating, i.e. σ_ξ, would affect the proportion of networks exhibiting insight-like behaviour. Parametrically varying the variance of noise added to colour and motion gates and weights led to increases in insight-like behaviour, from no single insight network when no noise was added to 100% insight networks when σ_{ξ_g} reached values of larger than approx. 0.05 (Fig. 13A). Since gate and weight updates were coupled (see Eq. 4-7), noise during

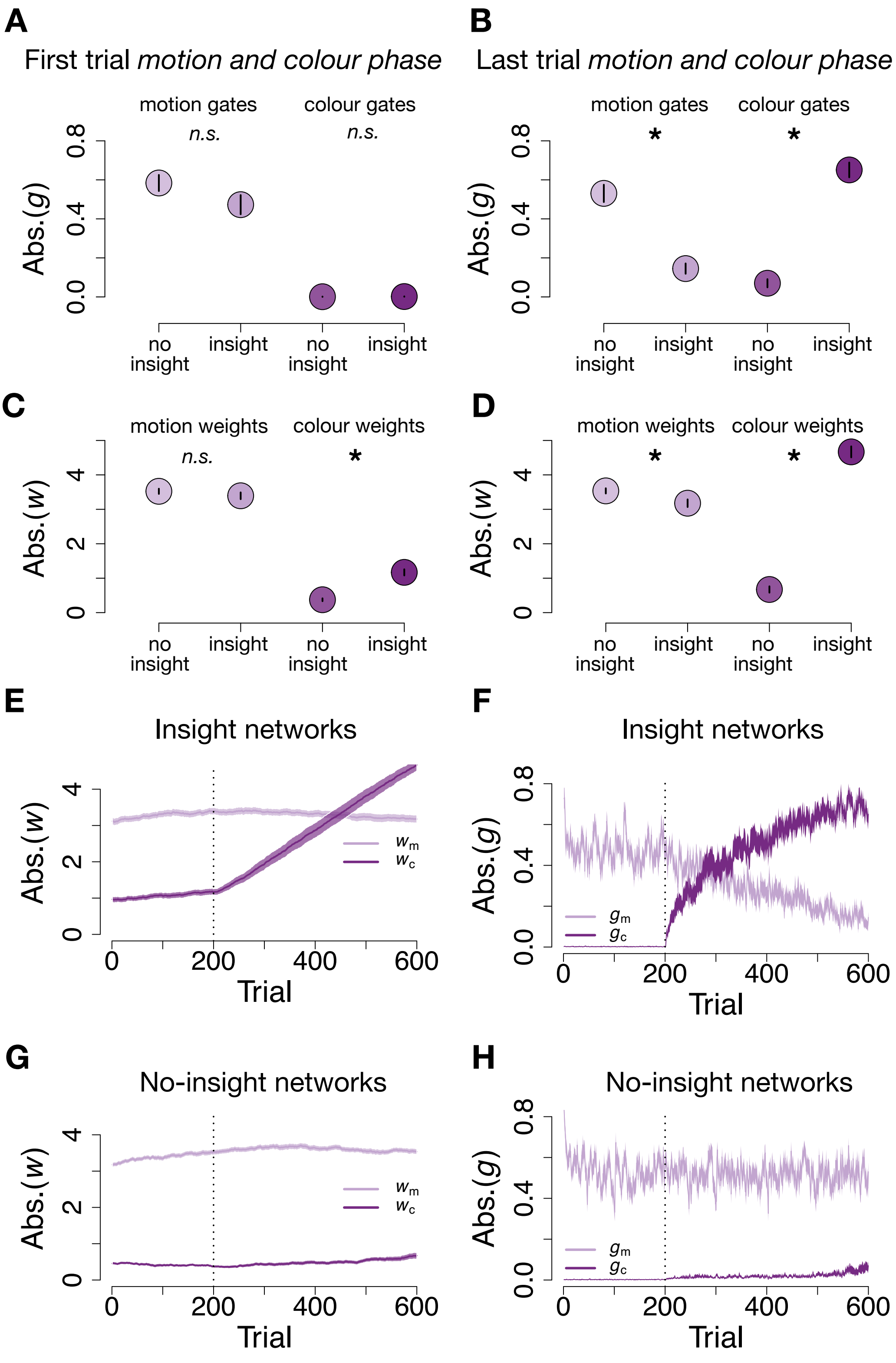


Figure 12: Differences in parameter dynamics between insight and no-insight networks.

Caption on the next page.

Figure 12: Absolute average magnitude of motion (light purple) and colour (dark purple) gates at the first trial **(A)** and the last trial **(B)** of the *motion and colour phase*, shown separately for insight and no-insight networks. Differences between the gates of insight vs no-insight networks emerged only at the end of *motion and colour phase*, i.e. only after insight-like behaviour had occurred. Panels **(C)** and **(D)** show the same for absolute weight magnitudes. Unlike gates, colour weights differed between network types already before insight-like switches were detectable, i.e. at the start of the *motion and colour phase*. Both absolute colour weight **(E)** and gate **(F)** magnitudes increase after start of the *motion and colour phase* (trial 200, dashed vertical line) for insight networks, while absolute colour gates decrease in magnitude after colour correlation onset. For no-insight networks absolute weight **(G)** and gate **(H)** magnitudes do not show such dynamics and remain stable after the onset of the *motion and colour phase*. Error bars/shadows signify SEM.

one gradient update could in principle affect other updates as well. We therefore separately manipulated the noise added to updates of colour gates and weights, motion gates and weights, all weights and all gates. This showed that adding noise to only weights during the updates was sufficient to induce insight-like behaviour (Fig. 13B). In principle, adding noise to only gates was sufficient for insight-like switches as well, although noise applied to the gates had to be relatively larger to achieve the same effect as applying noise to weight gradients (Fig. 13B), presumably due the effect of regularisation. Adding noise only to the gradients of motion gates or weights was not sufficient to induce insight-like switches (Fig. 13B). On the other hand, noise added only to the colour parameter updates quickly led to substantial amounts of insight-like behavioural switches (Fig. 13B).

An analysis of *cumulative* noise showed that the effects reported above are mostly about momentary noise fluctuations: cumulative noise added to the output did not differ between insight and no-insight networks at either the start (first trial of the *motion and colour phase*) or end of learning (last trial of the *motion and colour phase*) (start: $M = -0.3 \pm 4.7$ vs. $M = -0.6 \pm 3.9$; $t(91.2) = 0.4$, $p = 0.7$, end: $M = 0.6 \pm 7.1$ vs. $M = 0.5 \pm 7.1$; $t(96.7) = 0.07$, $p = 1$), and the same was true for cumulative noise added during the gradient updates to weights and gates (see Supplementary Material for details).

We therefore conclude that Gaussian noise added to updates of particularly colour gate weights, in combination with "silent knowledge" about colour information stored in suppressed weights, is a crucial factor for insight-like behavioural changes.

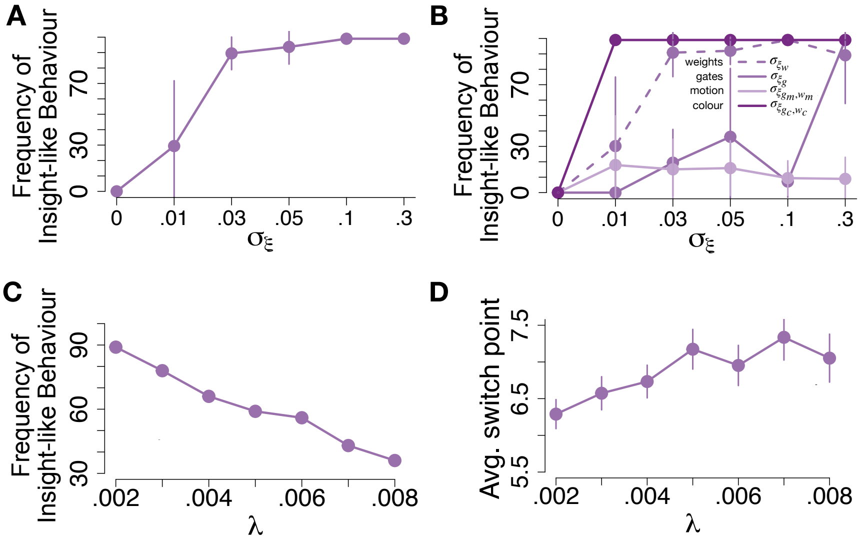


Figure 13: Influence of gradient noise σ_ξ and regularisation λ on insight-like switches. **(A)** Influence of gradient noise (standard deviation σ_ξ) on the frequency of absolute numbers of insight-like networks. The frequency of insight-like switches increases gradually with σ_ξ until it reaches a ceiling around σ_ξ = .05. Error bars are SD. Results from 10 simulations of 99 networks each. **(B)** Effects of gradient noise added only to either all weights (σ_{ξ_w}, light purple dashed line), all gates (σ_{ξ_g}, solid purple line), all motion parameters (i.e. motion weight and motion gates, $\sigma_{\xi_{g_m,w_m}}$, light purple solid line) and all colour parameters ($\sigma_{\xi_{g_c,w_c}}$, dark purple solid line) on the frequency of insight-like switches. While only small amounts of noise in the colour parameters $\sigma_{\xi_{g_c,w_c}}$ suffice to induce 100% insight-like strategy switches among networks (dark purple line), adding noise to the motion parameters $\sigma_{\xi_{g_c,w_c}}$ (light purple solid line) only had a a very minor effect. Furthermore, we find that adding noise specifically to the weights (σ_{ξ_w}, dashed purple line), has a much larger effect than adding noise to the gates (σ_{ξ_g}, solid purple line). Error bars are SD. Results from 10 simulations of 99 networks each. Colour scheme as in Fig. 1B **(C)** The frequency of insight-like switches declines with increasing λ. **(D)** The average switch point occurs later in the task with increasing λ. Error bars signify SEM.

3.2.4 Regularisation Parameter λ Affects Insight Behaviour Delay and Frequency

In our previous results, the regularisation parameter λ was arbitrarily set to 0.07. We next tested the effect of of λ on insight-like behaviour. The number of L1-regularised insight networks linearly decreased with increasing λ (Fig. 13C). Lambda further had an effect on the delay of the insight-like switches, with smaller λ values leading to decreased average delays of switching to a colour strategy after predictiveness of the inputs had changed (Fig. 13D). The regularisation parameter λ thus affects two of the key characteristics of human insight – selectivity and delay.

3.3 Discussion

We investigated insight-like learning behaviour in neural networks. In a binary decision making-task with a hidden regularity that entailed an alternative way to solve the task more efficiently, a subset of L1-regularised neural networks with multiplicative gates of their input channels displayed spontaneous, jump-like learning that signified the sudden discovery of the hidden regularity – insight moments often deemed "mysterious" boiled down to the simplest expression. Networks exhibited all key characteristics of human insight-like behaviour in the same task (suddenness, selectivity, delay). Crucially, neural networks were trained with standard online stochastic gradient descent that is often associated with gradual learning. Our results therefore suggest that the behavioural characteristics of Aha! moments can arise from gradual learning mechanisms, and hence suffice to mimic human insight. To our knowledge, this is the first time that insight-like behaviour has been shown in a gradual delta rule learning system, devoid of complex cognitive processes.

Network analyses identified the factors which caused insight-like behaviour in L1-networks: noise added during the gradient computations accumulated to non-zero weights in some networks. As long as colour information was not useful yet, i.e. prior to the onset of the hidden regularity, close-to-0 colour gates rendered these weights "silent", such that no effects on behaviour can be observed. Once the hidden colour regularity became available, the non-zero colour weights helped to trigger non-linear learning dynamics that arise during gradient updating, and depend on the starting point. Hence, our results hint at important roles of gating as an "attentional" mechanism, noise, and L1-regularisation as the computational origins of sudden, insight-like behavioural changes. We report several findings that are in line with this interpretation: addition of gradient noise ξ in particular to the colour weights and gates, pre-learning adjustment of colour weights and a reduction of the regularisation parameter λ all increased insight-like behaviour. We note that our networks did not have a hidden layer, witnessing the fact that no hidden layer is needed to produce non-linear learning dynamics. This is however not a necessity, as we can reproduce the same results in more complex networks with a hidden layer (see Supplementary Material).

Our choice to include explicit regularisation of only the gates was motivated by the idea that capacity constraints enforce regularisation only during the output stage, while the initial sensory processing is unaffected. Note, however, that regularising only the weights, but not the gates, would have functionally equivalent effects, and regularisation of both suppresses any learning.

Furthermore, since our architecture is functionally equivalent to a 2-layer diagonal linear network with L1-regularisation on the second layer, our results could equally pertain to a particular simple form of a linear network with 2 layers.

Our findings have implications for the conception of insight phenomena in humans. While present-day machines clearly do not have the capacity to have Aha! moments due to their lack of meta-cognitive awareness, our results show that the remarkable behavioural signatures of insight-like strategy switches by themselves do not necessitate a dedicated process. This raises the possibility that sudden behavioural changes which occur even during gradual learning could in turn lead to the subjective effects that accompany insights (Frensch et al., 2003; Esser, Lustig, & Haider, 2022).

Our neural network model differs from other computational accounts of insight in the same way that our task neither includes a classic insight problem to actively solve, nor an off-task incubation period. The EII (explicit-implicit interaction) model (Hélie

& Sun, 2010) for example uses the CLARION architecture which has two separate, but interacting modules for explicit and implicit knowledge. The insight goal in this account is reached when an internal confidence level threshold is crossed through an active, iterative problem solving process. Our model architecture differs from this as information is represented and processed in one single way. Insight-like behaviour is furthermore a phenomenon occurring ”for free” at the same time that gradual learning happens in our task which stands in contrast to the EII model which defines insight, not performance, as the ultimate goal.

Our results highlight noise and regularisation as aspects of brain function that are involved in the generation of insights. Cellular and synaptic noise is omnipresent in brain activity (Faisal, Selen, & Wolpert, 2008; Waschke, Kloosterman, Obleser, & Garrett, 2021), and has a number of known benefits, such as stochastic resonance and robustness that comes with probabilistic firing of neurons based on statistical fluctuations due to Poissonian neural spike timing (Rolls, Tromans, & Stringer, 2008). It has also been noted that noise plays an important role in jumps between brain states, when noise provokes transitioning between attractor states (Rolls & Deco, 2012). Previous studies have therefore noted that stochastic brain dynamics can be advantageous, allowing e.g. for creative problem solving (as in our case), exploratory behaviour, and accurate decision making (Rolls & Deco, 2012; Faisal et al., 2008; Garrett et al., 2013; Waschke et al., 2021). Albeit our conceptual model was not meant to be biologically realistic, this work adds a computationally precise explanation of how noise can lead to insight-like behaviour to this literature. Questions about whether inter-individual differences in neural variability predict insights (Garrett et al., 2013), or about whether noise that occurs during synaptic updating is crucial remain an interesting topic for future research.

While our simulations focused on the specific explicit regularisation mechanism of adding a penalty term to the error function, many other forms of implicit regularisation, such as weight sharing, might exist and be possibly implemented in the brain. To what extent our results generalise to such other regularisation techniques is unknown and speculative at this point. Regularisation has for instance been implied in synaptic scaling, which helps to adjust synaptic weights in order to maintain a global firing homeostasis (Lee & Kirkwood, 2019), thereby aiding energy requirements and reducing memory interference (Tononi & Cirelli, 2014; De Vivo et al., 2017). It has also been proposed that regularisation modulates the threshold for induction of long-term potentation (Lee & Kirkwood, 2019). These mechanisms therefore present possible synaptic factors that contribute to insight-like behaviour in humans and animals. We note that synaptic scaling has often been linked to sleep (Tononi & Cirelli, 2014), and regularisation during sleep has also been suggested to help avoid overfitting to experiences made during the day, and therefore generalisation (Hoel, 2021). Chapter 4 therefore investigates the effect of a daytime nap intervention on insight, finding that the steepness of the spectral slope during sleep, which has been indirectly linked to regularisation (Lendner et al., 2023), predicted insight above and beyond sleep stages (see Chapter 4) (A. Löwe et al., 2024). The relationship between insight-like behaviour and regularisation in the broader sense - and different regularisation mechanisms specifically - thus remains an interesting area for future research.

On a more cognitive level, regularisation has been implied in the context of heuristics. In this notion, regularisation has been proposed to function as an infinitely strong prior in a Bayesian inference framework (Parpart et al., 2018). This infinitely strong prior would work as a sort of attention mechanism and regularise input and information in a way that is congruent with the specific prior, whereas a finite prior would under this assumption enable learning from experience (Parpart et al., 2018). Another account regards cognitive control as regularised optimisation (Ritz, Leng, & Shenhav, 2022). According to this theory, better transfer learning is supported by effort costs regularising towards more task-general policies. It therefore seems possible that the factors that impact regularisation during learning can also lead to a neural switch between states that might be more or less likely to govern insights.

The occurrence of insight-like behaviour with the same characteristics as found in humans was specific to L1-regularised networks, while no comparable similarity occurred in L2- or non-regularised networks. While no hidden layer is necessary to produce this result, the same L1-specific effect can be replicated in a model with a hidden layer (see Supplementary Material). Although L2-regularised neural networks learned to suppress initially irrelevant colour feature inputs and showed abrupt performance increases reminiscent of insights, only L1 networks exhibited a wide distribution of time points when the insight-like switches occur (delay) as well as a selectivity of the phenomenon to a subgroup of networks, as found in humans. We note that L2- and non-regularised networks technically performed better on the task, because they collectively improve their behavioural efficiency sooner. One important question therefore remains under which circumstances L1 would be the most beneficial form of regularisation. One possibility could be that the task is too simple for L1-regularisation to be beneficial. It is conceivable that L1-regularisation only starts being advantageous in more complex task settings when generalisation across task sets is required and a segregation of task dimensions to learn about at a given time would prove useful.

Taken together, gradual training of neural networks with gate modulation leads to insight-like behaviour as observed in humans, and points to roles of regularisation, noise and “silent knowledge” in this process.

While general scalability remains an issue for future research, we believe that the mechanistic insights from our model have implications for more complex systems, such as the brain. Our results imply a link between behavioural markers of insight and noise measurements as well as regularisation in the brain. Towards this goal of generalisation we replicated our results using a multi-neuron, deep, non-linear model with a 48-unit hidden layer and investigated effects of different types of explicit regularisation (L1 vs L2 vs no regularisation), showing that only L1-regularised networks exhibit all three key characteristics of human insight. We further explained that our network is identical to a 2-layer diagonal network with L1-regularisation on the second layer and that L1-regularisation of weights has equivalent effects of gate regularisation, while L1-regularising both entirely changes the model behaviour. We speculate that even more complex, deeper networks, which would allow to disentangle several factors of variation (e.g. motion, shape, colour, lighting) and apply gating variables to these, might exhibit similar shifts between relying on subsets of these features.

These results make an important contribution to the general understanding of learning dynamics and representation formation in environments with non-stationary feature relevance in both biological and artificial agents.

3.4 Methods

3.4.1 L1-regularised Neural Networks

We utilise a simple neural network model to reproduce the observations of the human behavioural data in a simplified supervised learning regression setting. We trained a simple neural network with two input nodes, two input gates and one output node on the same decision making task (Fig. 20D).

The network received two inputs, x_m and x_c, corresponding to the stimulus motion and colour, respectively, and had one output, $\hat{y}$. Importantly, each input had one associated multiplicative gate (g_m, g_c) such that output activation was defined as $\hat{y} = \operatorname{sign}(g_m w_m x_m + g_c w_c x_c + \eta)$ where $\eta \sim \mathcal{N}(0, \sigma)$ is Gaussian noise (Fig. 20D).

To introduce competitive dynamics between the input channels, we added L1-regularisation on the gate weights g, resulting in the following loss function:

$$\mathcal{L} = \frac{1}{2}(g_m w_m x_m + g_c w_c x_c + \eta - y)^2 + \lambda(|g_m| + |g_c|) \tag{5}$$

The network was trained in a gradual fashion through online gradient descent with Gaussian white noise ξ added to the gradient update and a fixed learning rate α. Given the loss function, this yields the following update equations for noisy stochastic gradient descent:

$$\Delta w_m = -\alpha x_m g_m (x_m g_m w_m + x_c g_c w_c + \eta - y) + \xi_{w_m} \tag{6}$$

$$\Delta g_m = -\alpha x_m w_m (x_m g_m w_m + x_c g_c w_c + \eta - y) - \alpha\lambda \operatorname{sign}(g_m) + \xi_{g_m} \tag{7}$$

$$\Delta w_c = -\alpha x_c g_c (x_c g_c w_c + x_m g_m w_m + \eta - y) + \xi_{w_c} \tag{8}$$

$$\Delta g_c = -\alpha x_c w_c (x_c g_c w_c + x_m g_m w_m + \eta - y) - \alpha\lambda \operatorname{sign}(g_c) + \xi_{g_c} \tag{9}$$

with λ = 0.07, α = 0.6 and $\sigma_\xi = 0.05)$.

This implies that the evolution of the colour weights and gates will exhibit non-linear quadratic and cubic dynamics, driven by the interaction of w_c and g_c. Multiplying the weights w with the regularised gate weights g leads to smaller weights and therefore initially slower increases of the colour weights w_c and respective gate weights g_c after colour has become predictive of correct choices.

To understand this effect of non-linearity analytically, we used a simplified setup of the same model without gate weights:

$$\mathcal{L} = [w_m x_m + w_c x_c + \eta - y]^2 \tag{10}$$

Using this model, we observe exponential increases of the colour weights w_c after the onset of the *motion and colour phase*. This confirms that the interaction of w_c and g_c, as well as the regularisation applied to g_c are necessary for the insight-like non-linear dynamics including a distribution of insight-like strategy switch onsets as well as variety in slope steepness of insight-like switches.

The accuracy is given by:

$$\begin{aligned}&\mathbb{P}[\hat{y} = y | w_m, g_m, w_c, g_c]\\ &= \frac{1}{2}[1 + \mathrm{erf}(\frac{g_m w_m x_m + g_c w_c x_c}{\sqrt{2((g_m w_m \sigma_m)^2 + (g_c w_c \sigma_c)^2 + \sigma^2)}})]\end{aligned} \tag{11}$$

We trained the network on a curriculum precisely matched to the human task, and adjusted hyperparameters (noise levels), such that baseline network performance and learning speed were carefully equated between humans and networks.

Specifically, we simulated the same number of networks than humans ($N = 99$). We matched the motion noise based performance variance of a given simulation to a respective human subject using a non-linear COBYLA optimiser. While the mean of the colour input distribution (0.22) as well as the standard deviations of both input distributions were fixed (0.01 for colour and 0.1 for motion), the respective motion input distribution mean values were individually fitted for each single simulation as described above.

The input sequences the networks received were sampled from the same ten input sequences that humans were exposed to in task phase two. This means that for the task part where colour was predictive of the correct binary choice, *motion and colour phase* (500 trials in total), networks and humans received the same input sequences.

The networks were given a slightly longer *training phase* of six blocks (600 trials) in comparison to the two blocks *training phase* that human subjects were exposed to (Fig. 10B). Furthermore, human participants first completed a block with 100% motion coherence before doing one block with low motion noise. The networks received six *training phase* blocks containing the three highest motion coherence levels. Both human subjects and networks completed two blocks including all noise levels in the *motion phase* before colour became predictive in the *motion and colour phase*.

3.4.2 L2-regularised Neural Networks

To probe the effect of the aggressiveness of the regulariser on insight-like switch behaviour in networks, we compared our L1-regularised networks with models of the same architecture, but added L2-regularisation on the gate weights g. This yielded the following loss function:

$$\mathcal{L} = \frac{1}{2}(g_m w_m x_m + g_c w_c x_c + \eta - y)^2 + \frac{\lambda}{2}(|g_m| + |g_c|)^2 \tag{12}$$

From the loss function we can again derive the following update equations for noisy stochastic gradient descent:

$$\Delta w_m = -\alpha x_m g_m (x_m g_m w_m + x_c g_c w_c + \eta - y) + \xi_{w_m} \tag{13}$$

$$\Delta g_m = -\alpha x_m w_m (x_m g_m w_m + x_c g_c w_c + \eta - y) - \alpha\lambda\, g_m + \xi_{g_m} \tag{14}$$

$$\Delta w_c = -\alpha x_c g_c (x_c g_c w_c + x_m g_m w_m + \eta - y) + \xi_{w_c} \tag{15}$$

$$\Delta g_c = -\alpha x_c w_c (x_c g_c w_c + x_m g_m w_m + \eta - y) - \alpha\lambda\, g_c + \xi_{g_c} \tag{16}$$

with λ = 0.07, α = 0.6 and $\sigma_\xi = 0.05)$.

The training is otherwise the same as for the L1-regularised networks.

3.4.3 Modelling of Insight-like Switches

We used the same classification procedure for neural networks as described in Chapter 1. All individual sigmoid function fits for L1-regularised networks can be found in the Supplementary Material (15).

3.5 Supplementary Information

Effects of Regularisation Type on Network Behaviour

Following our observation that L1-regularised networks exhibited human-like insight behaviour, we investigated whether this was specific to the form of regularisation. We therefore first trained otherwise identical networks with a L2-regularisation term on the gate weights. We hypothesised that L2-regularisation would also lead to competitiveness between input nodes, but to a lower extent than L1-regularisation. We reasoned that in particular the fact that during the *motion phase* the networks motion weights would not shrink as close to 0 would lead to more frequent and earlier insight-like behavioural switches.

While L2-regularised gate weights indeed led to switches that were similar to those previously observed in their abruptness (Fig. 17), as predicted such insight-like behaviours were much more frequent and clustered: 96% of networks switched to a colour strategy, with a switch point distribution that was much more centred around the onset of the colour predictiveness (Fig. 17F, average delay of 2 task blocks ($SD = 1.1$) corresponding to 100 trials after onset of the colour correlation (*motion and colour phase*). This was significantly shorter than for L1-regularised networks ($M = 1.05 \pm 1.1$ vs. $M = 1.75 \pm 1.05$, $t(59.6) = 4$, $p < 0.001$, $d = 0.9$) and also differed from a uniform distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(95) = 0.26$, $p = 0.005$). Additionally, performance on the lowest coherence level in the last block of the *colour and motion phase* before colour instruction was centred just below ceiling and thus did not indicate a range of colour use like humans and L1-regularised networks ($M(L2 - networks) = 97\% \pm 2\%$ vs. $M(humans) = 82 \pm 17\%$,

$t(101.6) = -8.8$, $p < .001$, $d = 1.25$, Fig. 18). While L2-regularised networks thus showed abrupt behavioural transitions, they failed to show the other two key characteristics of insight: selectivity and delay.

Next, we investigated network behaviour when no regularisation was applied. In non-regularised networks, the effects observed in L2-regularised networks are enhanced. 99% of the networks started using colour inputs (Fig. 19A), but colour use occurred in a more linear, less abrupt way than for L1- or L2-regularised networks. Additionally, there was very little delay of only 1.4 task blocks (70 trials, (± 0.25)) between onset of the *motion and colour phase* and the start of the networks making use of the colour input predictiveness (Fig. 19B). As for L2-networks, this delay was significantly shorter than for L1-regularised networks ($M = 0.7 \pm 0.55$ vs. $M = 1.75 \pm 1.05$, $t(49.3) = 6.6$, $p < 0.001$, $d = 1.6$) and also differed from a uniform distribution taking into account the hazard rate (Exact two-sided Kolmogorov-Smirnov test: $D(98) = 0.35$, $p < .001$). Similarly, performance on the lowest coherence level in the last block indicated that all networks used colour inputs ($M = 100\% \pm 0.3\%$ vs. $M = 82 \pm 17\%$, $t(98) = -10.4$, $p < .001$, $d = 1.5$, Fig. 19). Thus non-regularised networks also did not show the insight key behavioural characteristics of selectivity and delay.

Hidden Layer Model

In order to verify that our results were not merely an artefact of the oversimplified models we used, we tested the task on a simple deep neural network that had one additional hidden layer of fully connected linear units.

The linear neural network received two inputs, x_m and x_c, corresponding to the stimulus motion and colour, respectively, and had two output nodes, $\hat{y}$, as well as one hidden layer of 48 units. Importantly, each weight connecting the inputs with a hidden unit had one associated multiplicative gate g. To introduce competitive dynamics between the input channels, we again applied L1-regularisation on the gate weights g.

The network was trained on the Cross Entropy loss using stochastic gradient descent with λ = 0.002 and α = 0.1.

As for the one-layer network, we trained this network on a curriculum precisely matched to the human task, and adjusted hyperparameters (noise levels), such that baseline network performance and learning speed were carefully equated between humans and networks (see Methods).

We employed the same analysis approach to detect insight-like behaviour (see Methods for details) by running simulations of a ”control” network of the same architecture, but without correlated features and therefore without colour predictiveness in the *motion and colour phase*. We found that when we applied L1-regularisation with a regularisation parameter of $\lambda = 0.002$ on the gate weights, 18.2% of the networks exhibited *abrupt* and *delayed* learning dynamics, resembling insight-like behaviour in humans (Fig. 14A) and thereby replicating the key insight characteristics suddenness and selectivity. Insight-like switches to the colour strategy thereby again improved the networks' performance significantly. Using the same parameters, experimental setup and analyses, but applying L2-regularisation on the gate weights g, yielded an insight-like switch rate of 51.5% (Fig. 14B).

We again also observed a wider distribution of delays, the time point when the switches in the *motion and colour phase* occurred in insight networks, for L1-regularised networks with a hidden layer (Fig. 14C-D).

Taken together, these results mirror our observations from network simulations with a simplified setup. We can thereby confirm that our results of L1-regularised neural networks' behaviour exhibiting all key characteristics of human insight behaviour (suddenness, selectivity and delay) are not an artefact of the one-layer linearity.

Weight and Gate Differences between L1- and L2-regularised Networks

At correlation onset (first trial of *motion and colour phase*), neither motion nor colour weights differed (motion: $M = 3.5{\pm}0.6$ vs $M = 3.4{\pm}0.5$, $t(192.7) = 1.2$, $p = 0.2$, $d = 0.2$, colour: $M = 0.8 \pm 0.6$ vs $M = 0.8 \pm 0.5$, $t(189.2) = 0.4$, $p = 0.7$, $d = 0.1$). After learning, however, i.e. at the last trial of the *motion and colour phase*, the average absolute size of the colour weights was higher in L2- compared to L1-networks ($M = 2.6 \pm 2.2$ vs

$M = 4.7 \pm 0.7$, $t(115.1) = -9$, $p < .001$, $d = 1.3$), while the reverse was true for motion weights ($M = 3.4 \pm 0.7$ vs $M = 2.8 \pm 0.6$, $t(194.9) = 5.6$, $p < .001$, $d = 0.8$). For gate weights, differences between L1- and L2-networks are already apparent at correlation onset (first trial of *motion and colour phase*), where the mean of the motion gate was 0.53 for L1-networks and 0.58 for L2-networks, and hence lower in L1 networks, albeit not significantly ($t(195.1) = -1$, $p = 0.3$, $d = 0.1$). In addition, the average absolute size of the colour gate weights was higher in L2- compared to L1-networks ($M = 0.04 \pm 0.05$ vs $M = 0.002 \pm 0.006$, respectively, $t(100.6) = -7.2$, $p < 0.001$, $d = 1$). The respective distributions also reflected these effects. L1-networks had a much more narrow distribution for colour gates and just slightly narrower distribution for motion gates (L1: colour gates: 0 to 0.04, motion gates: 0 to 1.3, L2: colour gates: 0 to 0.2, motion gates: 0 to 1.4) After learning, i.e. at the last trial of the *motion and colour phase*, the mean colour gate size still was lower in L1- compared to L2-regularised networks ($M = 0.4 \pm 0.4$ vs $M = 0.8 \pm 0.2$, $t(169.1) = -9.3$, $p < 0.001$, $d = 1.3$), while the reverse was true for motion gates ($M = 0.3 \pm 0.3$ vs $M = 0.2 \pm 0.2$, $t(152.4) = 3.9$, $p < 0.001$, $d = 0.6$, see Fig. 18). This was again also reflected in the respective distributions with L1-networks having much wider distributions for motion and slightly shorter width for colour gates (L1: colour gates: 0 to 1.2, motion gates: 0 to 1.3, L2: colour gates: 0 to 1.3, motion gates: 0 to 0.7).

Gaussian Noise Differences at Weights and Gates between Insight and No-Insight Networks

Comparing Gaussian noise $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ at the weights and gates around the individually fitted switch points revealed no differences between insight and no-insight networks for either motion or colour weights (colour weights: $M = -0.08 \pm 1$ vs. $M = 0.04 \pm 0.8$; $t(89.5) = -0.6$, $p = 0.5$, motion weights: $M = 0.5 \pm 0.3$ vs. $M = 0.6 \pm 0.3$; $t(93.1) = -1.7$, $p = 0.09$) or gates (colour gates: $M = -0.1 \pm 0.9$ vs. $M = 0.1 \pm 0.9$; $t(95.3) = 0.8$, $p = 0.44$, motion gates: $M = 0.2 \pm 0.6$ vs. $M = -0.3 \pm 0.8$; $t(94.4) = 2$, $p = 0.05$). There also were no σ_ξ differences at either the start of learning (first trial of the *mo-*

tion and colour phase) (colour weights: $M = -0.06 \pm 0.8$ vs. $M = -0.03 \pm 0.5$; $t(78.1) = -0.2$, $p = 0.8$, motion weights: $M = 0.08 \pm 0.7$ vs. $M = 0.07 \pm 0.7$; $t(96.7) = 1$, $p = 0.3$, colour gates: $M = 0 \pm 0.6$ vs. $M = -0.2 \pm 0.7$; $t(97) = 1.6$, $p = 0.1$, motion gates: $M = -0.04 \pm 0.6$ vs. $M = -0.07 \pm 0.7$; $t(97) = 0.2$, $p = 0.8$) or end of learning (last trial of the *motion and colour phase*)(colour weights: $M = 0.05 \pm 1.3$ vs. $M = 0.08 \pm 1.1$; $t(92.7) = -0.1$, $p = 0.9$, motion weights: $M = 0 \pm 1.2$ vs. $M = -0.02 \pm 1.1$; $t(95.6) = 0.04$, $p = 1$, colour gates: $M = 0.2 \pm 1.1$ vs. $M = -0.2 \pm 1.2$; $t(97) = 1.7$, $p = 0.09$, motion gates: $M = -0.1 \pm 1.3$ vs. $M = 0.05 \pm 1.3$; $t(96) = -0.7$, $p = 0.5$).

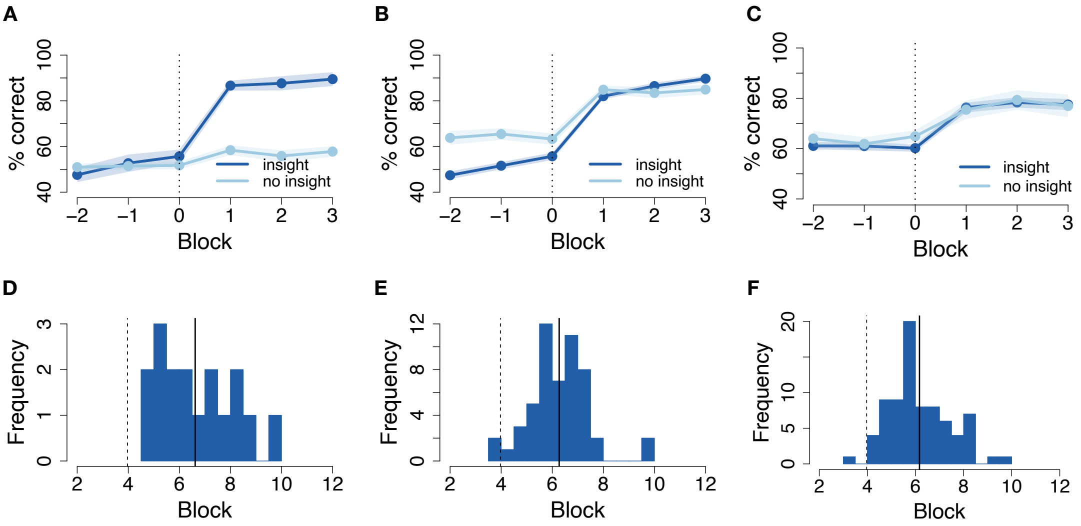
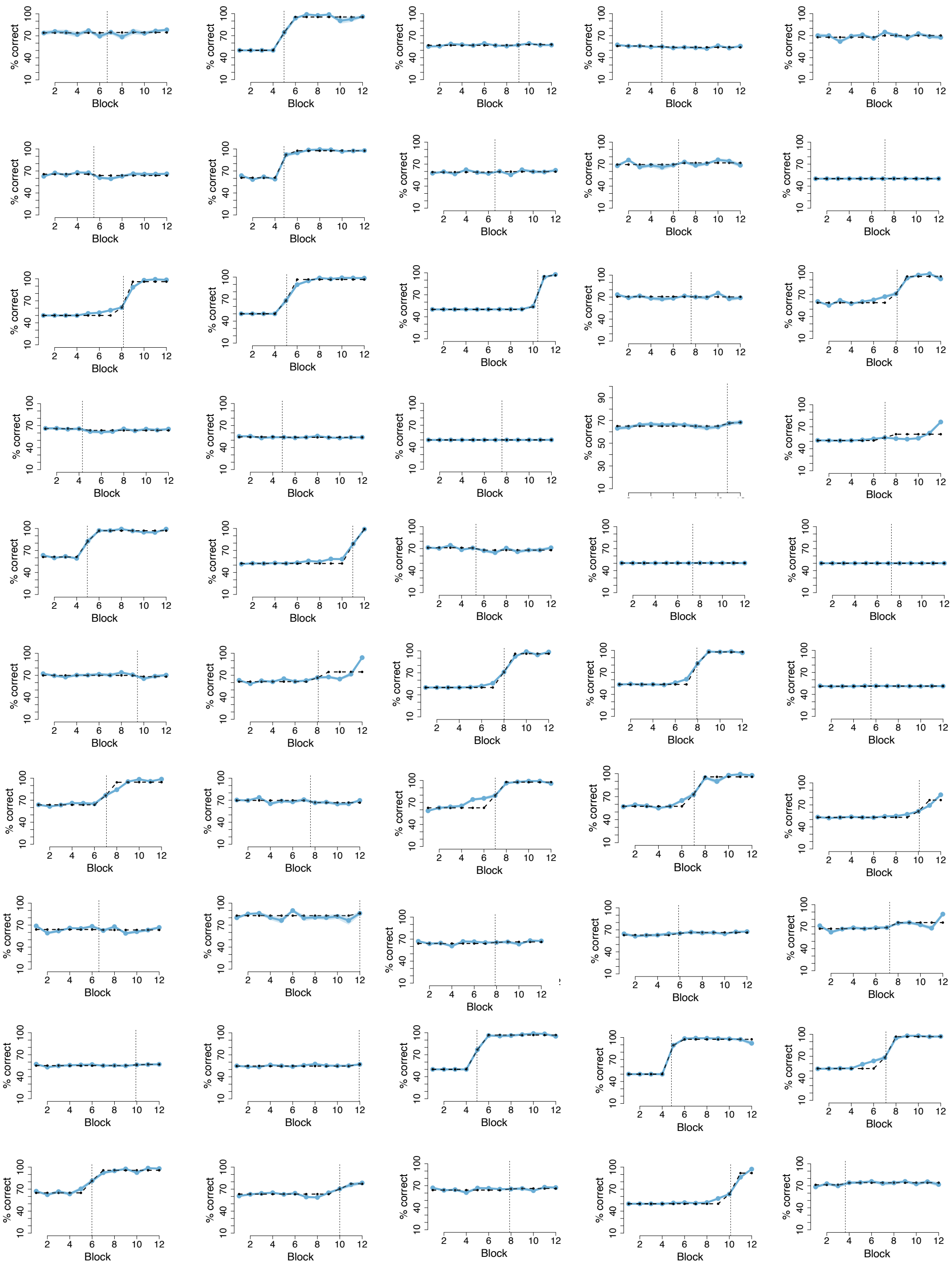


Figure 14: Switch-aligned performance and switch point distributions for L1- and L2-regularised neural networks with a 48 unit hidden layer each.

Blocks shown are halved task blocks (50 trials each). Error shadows signify SEM. **(A)** Switch-aligned performance for insight (18/99) and no-insight groups (81/99) respectively for L1-regularised networks with a hidden layer. **(B)** Switch-aligned performance for insight (51/99) and no-insight (48/99) L2-regularised neural networks with a hidden layer. **(C)** Switch-aligned performance for insight (78/99) and no-insight (21/99) non-regularised neural networks with a hidden layer. **(D)** Switch point distributions for L1-regularised insight networks with a hidden layer. **(E)** Switch point distributions for L2-regularised insight neural networks. Dashed vertical line marks onset of colour predictiveness. **(F)** Switch point distributions for non-regularised insight neural networks. Dashed vertical line marks onset of colour predictiveness.


% correct
10
40
70
100
Block
2
4
6
8
10
12

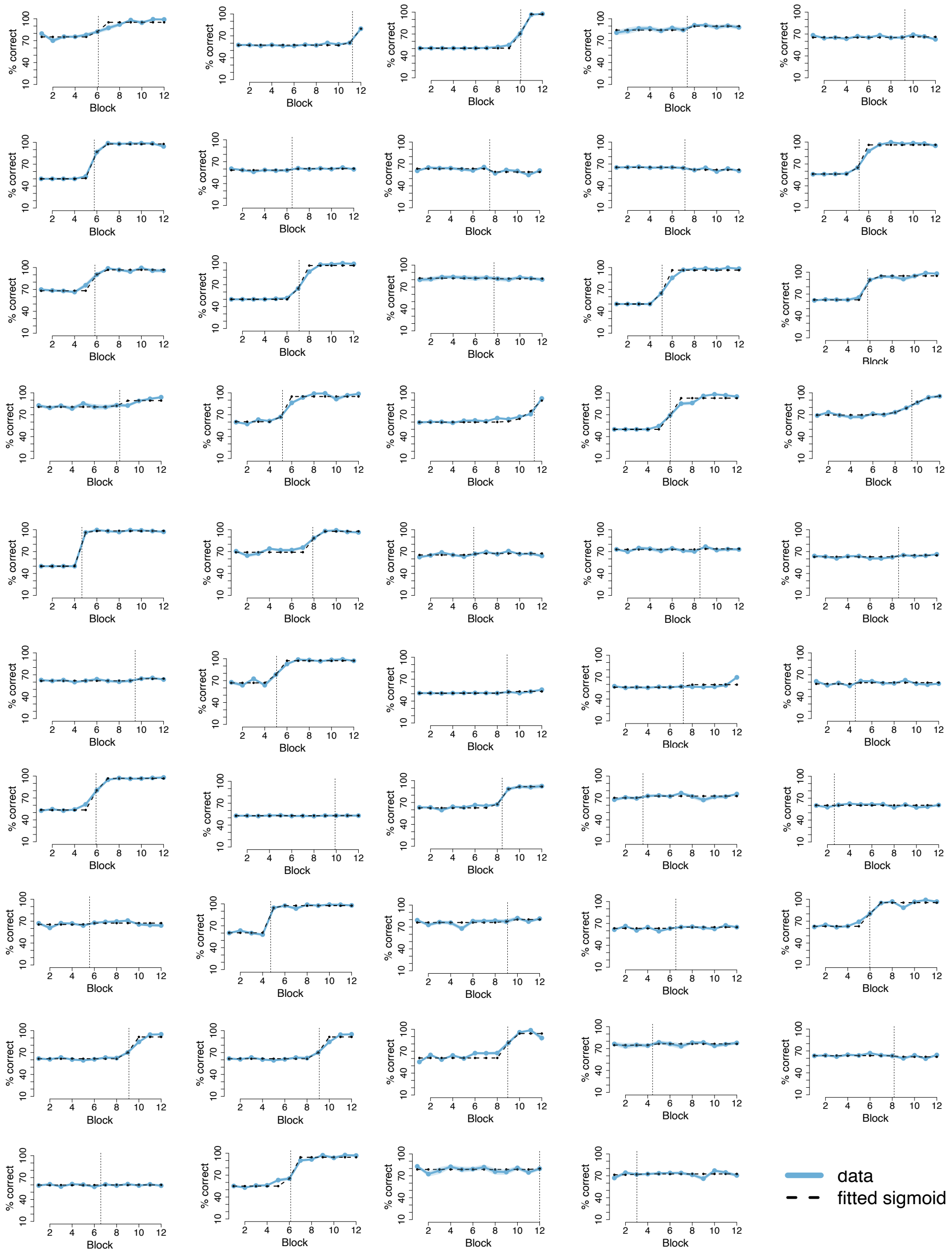


Figure 15: Network performance and model predictions.

Performance on highest motion noise trials (blue) and model predictions (black) for every L1-regularised neural network. Blocks shown are halved task blocks (50 trials each). Error shadows signify SEM.

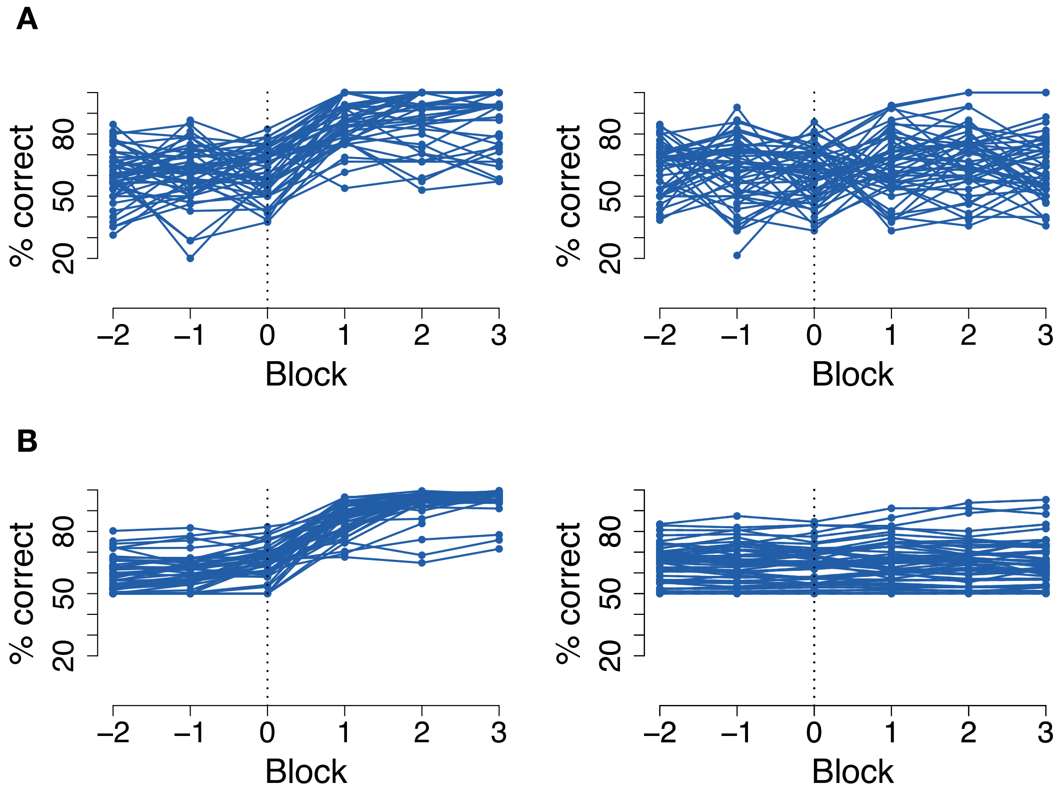


Figure 16: Switch-aligned performance for insight group (left) and no-insight group (right).

(A) Human insight group (49/99). **(B)** L1-regularised neural network insight group (48/99).

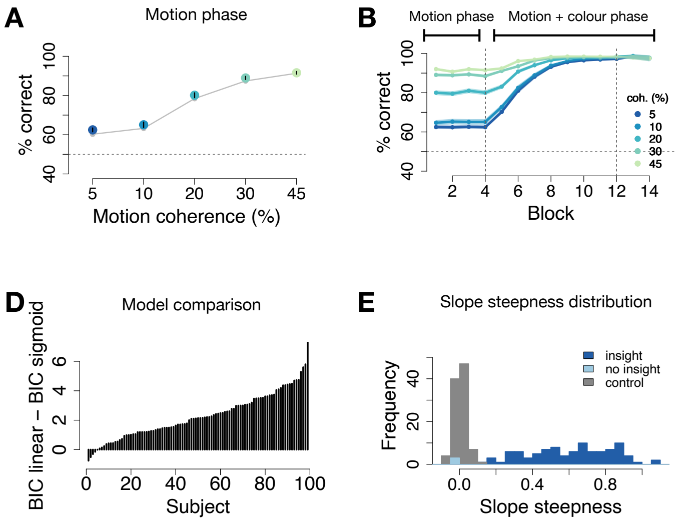


Figure 17: L2 networks: Task performance and insight-like strategy switches.

(A) Accuracy (% correct) during the *motion phase* increases with increasing motion coherence. Blocks shown are halved task blocks (50 trials each). N = 99, error bars signify SEM. Grey line is human data for comparison. **(B)** Accuracy (% correct) over the course of the experiment for all motion coherence levels. First dashed vertical line marks the onset of the colour predictiveness (*motion and colour phase*), second dashed vertical line the ”instruction” about colour predictiveness. N = 99, error shadows signify SEM. **(C)** Switch point-aligned accuracy on lowest motion coherence level for insight (95/99) and no-insight (4/99) networks. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **(D)** Difference between BICs of the linear and sigmoid function for each network. **(E)** Distributions of fitted slope steepness at inflection point parameter for control networks and classified insight and no-insight groups. **(F)** Distribution of switch points for insight networks. Dashed vertical line marks onset of colour predictiveness. Blocks shown are halved task blocks (50 trials each).

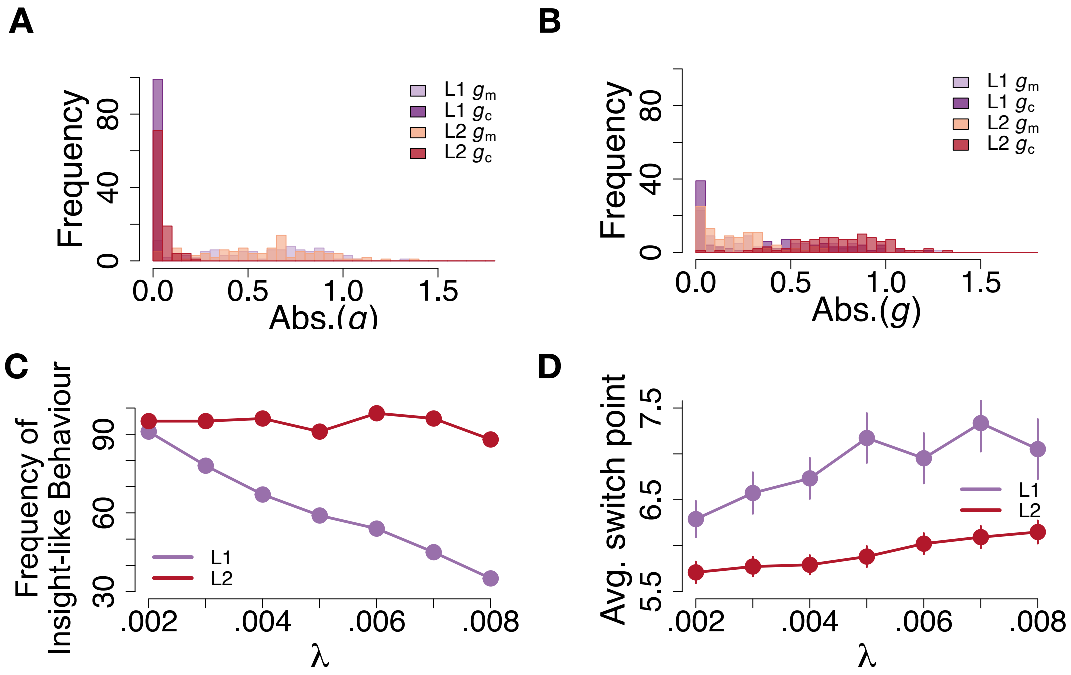


Figure 18: Comparison of gate weight magnitude and influence of λ on insight-like behaviour across L1- and L2-regularised networks.

Gate weight magnitude for colour and motion gate weights at the first trial **(A)** and the last trial **(B)** of the *motion and colour phase* for L1- and L2-regularised networks. **(C)** Influence of λ on the frequency of insight like behaviour in L1- vs L2-networks. The frequency of insight-like switches declines with increasing λ for L1-regularised networks, but is largely unaffected for L2-regularised networks. **(D)** Influence of λ on the average switch points. The average switch point occurs later in the task with increasing λ for both L1 and L2-regularised networks. Error bars signify SEM.

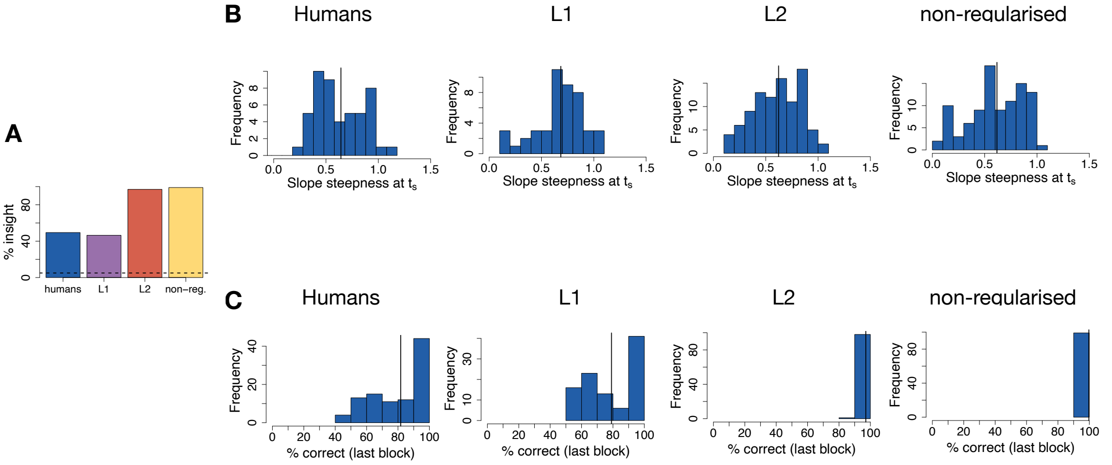


Figure 19: Comparison of insight percentage and performance in the last task block across groups.

(A) % insight-like switches in humans, L1, L2 and non-regularised networks, respectively. Dashed line marks chance percentage of "insight". **(B)** Distributions of slope steepness for humans, L1, L2 and non-regularised networks. **(C)** Distributions of performance (% correct) for humans, L1, L2 and non-regularised networks for the last block of the *colour and motion phase* before the colour instruction.

4 Sleep and Insight

Adapted from:

Löwe, A.T.*, Petzka, M., Tzegka, M. & Schuck, N.W.* (2024). *bioRxiv*.

Abstract

Humans sometimes have an insight that leads to a sudden and drastic performance improvement on the task they are working on. The precise origins of such insights are unknown. Some evidence has shown that sleep facilitates insights, while other work has not found such a relationship. One recent suggestion that could explain this mixed evidence is that different sleep stages have differential effects on insight. In addition, computational work has suggested that neural variability and regularisation play a role in increasing the likelihood of insight. To investigate the link between insight and different sleep stages as well as regularisation, we conducted a preregistered study in which N=90 participants performed a perceptual insight task before and after a 20 minute daytime nap. Sleep EEG data showed that N2 sleep, but not N1 sleep, increases the likelihood of insight after a nap, suggesting a specific role of deeper sleep. Exploratory analyses of EEG power spectra showed that spectral slopes could predict insight beyond sleep stages, which is broadly in line with theoretical suggestions of a link between insight and regularisation. In combination, our findings point towards a role of N2 sleep and aperiodic, but not oscillatory, neural activity for insight.

This chapter was reprinted under a CC-BY 4.0 license. doi: . Author contributions: ATL, MP and NWS conceived the experiment. MT and ATL carried out the experiments and ATL and MP analysed the data. All authors discussed the results and contributed to the final manuscript.

4.1 Introduction

Having an insight, or Aha! moment, is a unique learning phenomenon that has attracted researchers' interest for a century (Köhler, 1925). The cognitive and neural mechanisms that underlie insight are still debated (Stuyck et al., 2021; Weisberg, 2015), and have for instance been described as a restructuring of existing task representations (Wertheimer, 1925; Kounios & Beeman, 2014; Ohlsson, 1992). On a behavioural level, insight is often characterised by three features: an abrupt, non-linear increase in task performance (Haider & Rose, 2007; Durstewitz et al., 2010); a variable delay before the insight occurs 'spontaneously' (Ohlsson, 1992); and selective occurrence in only some, but not all participants (Schuck et al., 2015; A. T. Löwe et al., 2024).

An important milestone along the path to understanding insight will be to define the factors that facilitate its occurrence. One such potential factor is sleep, which is linked to memory consolidation (Rasch & Born, 2013) and restructuring of memories (Cowan et al., 2020), suggesting that it could be a facilitating factor for the incubation of insight. The evidence that sleep supports insight, however, is inconclusive. Work by Wagner et al. (2004) suggests a beneficial effect of a full night's sleep on insight, finding that more than twice as many subjects gained insight into a hidden task rule after sleep, compared to wakefulness. Another study reported similar findings after a daytime nap Lacaux et al. (2021). Other investigations, in contrast, did not find any benefits of sleep for insight, or reported no difference between sleep and awake rest (Cordi & Rasch, 2021; Schönauer et al., 2018; Brodt et al., 2018).

One possibility to explain divergent findings is that particular sleep stages affect insight in different ways. Lacaux et al. (2021) investigated this question by letting participants have a daytime nap in between sessions of a mathematical insight task, where discovering a hidden rule allowed to solve the task much more efficiently. In this case, a beneficial effect of sleep on insight was associated exclusively with sleep stage 1 (N1) (Lacaux et al., 2021), which led to a 83% probability to discover the hidden rule, compared to 30% in participants who stayed awake and 14% in those how reached deeper N2 sleep.

Given the diverging findings on the impact of sleep on insight, we conducted a preregistered daytime nap intervention study based on procedures by Lacaux et al. (2021), but used a different task (pregregistration link: https://osf.io/z5rxg/resources). We first aimed to replicate the above mentioned finding that N1 sleep compared to wakefulness after task exposure would lead to a higher number of insight moments about a hidden strategy during the post-nap behavioural measurement, while N2 sleep would lead to a reduced number of insight moments. A second major interest was to understand which features of the sleep-EEG signal best predict insight. Past work has focused on power in individual frequency bands (Lacaux et al., 2021). However, our own computational work (A. T. Löwe et al., 2024) has suggested that a combination of regularisation and noise had beneficial effects for insight. While a direct mapping between noise or regularisation in neural networks and electrophysiological signals is unknown, the concepts of noise (Voytek et al., 2015) and regularisation (as in synaptic downscaling, (Lendner et al., 2023)) have been indirectly linked to aperiodic activity. Additionally, aperiodic activtiy has been shown to decrease with an increase in sleep depth (Lendner et al., 2020, 2023; Ameen, Jacobs, Schabus, Hoedlmoser, & Donoghue, 2024). Hence, we also asked whether aperiodic activity of the EEG signal might have additional effects on insight, over and above the hypothesised relations to sleep stages.

Instead of the Number Reduction Task (NRT) employed by Lacaux et al. (2021), we employed the Perceptual Spontaneous Strategy Switch Task (PSSST) that also features a hidden task regularity, and which our previous works has shown to invoke insight-based spontaneous strategy switches (A. T. Löwe et al., 2024; Schuck et al., 2015; Gaschler et al., 2019). Similarly to the NRT, participants initially learned a functional, but suboptimal, strategy, which was replaced by some participants with a more optimal solution through an insight (Schuck et al., 2015, 2022; Gaschler et al., 2019; Allegra et al., 2020).

We note that while our task has the benefit to allow for tracking insight on a trial basis, it also differs from other tests in which participants are asked to actively search for a novel problem solution (e.g. Remotes Associates Tasks (Mednick, 1968) or Compounds Remotes Associates Tasks (Bowden & Jung-beeman, 2003)).

4.2 Results

To study the effect of different sleep stages on insight, 90 participants performed the PSSST, before and after a 20-minute nap break. Subjects were presented with a stimulus consisting of dots that were (1) either orange or purple (colour feature) and (2) moved in one of four possible orthogonal directions (motion feature, see Fig. 20A). Dot motion had a varying degree of noise across trials (5%, 23%, 41%, 59% or 76% coherent motion), making motion judgement relatively harder or easier on different trials. Participants were instructed to learn the correct button for each stimulus from trial-wise binary feedback (see Fig. 20A, B). The main task consisted of 9 blocks of 100 trials each in which participants had to press one of two buttons in response to the shown stimulus, and observe the feedback afterwards.

In the first three task blocks, only stimulus motion correlated with the correct response, such that the correct button was deterministically mapped onto the directions of the dots (two directions for each response). However, starting in the middle of block 4, stimulus colour began predicting the correct button as well (i.e. the colour was paired with the two directions that predicted the same response button, see Fig. 5A). After block 4, participants were given an opportunity to nap for 20 minutes in a reclining arm chair. We monitored brain activity and sleep during this phase using a 64-channel electroencephalography (EEG). Participants then completed 5 more blocks of the task, during which colour continued to predict the correct response in addition to motion (Fig. 5A). Additional details about the task can be found in the Methods section.

The subtle, unannounced change in task structure after 3.5 blocks provided a hidden opportunity to improve the decision strategy that could be discovered through insight. Insight was spontaneous in the sense that participants were not instructed about the hidden rule and did not need to switch their strategy to perform the task correctly. Only after a participant incidentally discovered the hidden rule did it become clear that using the colour could make the task easier.

We tracked insight on a trial-by-trial basis by monitoring rapid performance increases on high-noise (i.e. low motion coherence) trials, on which accuracy prior to the onset of colour predictiveness was at only 56% (vs. 92% in low noise trials; how accuracy depended on the noise level is shown in Fig. 20D). Performance in high noise trials was stable before the change in task structure (paired t-test first half of block 3 vs. first half of block 4: 55% vs. 58%, $t(157.8) = -1.51$, $p = 0.13$, $d = 0.23$, Fig. 20C), indicating that improvements do not arise simply due to training. A sudden change towards high accuracy on high noise trials can therefore be interpreted as indicative of insight about the colour-based strategy (Schuck et al., 2015; Gaschler et al., 2019; A. T. Löwe et al., 2024).

4.2.1 20 Minutes of Rest Increase Insight

Fifteen subjects had an insight before the nap and were therefore excluded from analysis. In another 7 cases EEG data quality prevented sleep classification, resulting in a total of 68 subjects for post nap data analysis. 70.6% (48/68) of participants showed abrupt, non-linear performance improvements after the nap and were thus classified as “insight participants” (Fig. 20E). Notably, this percentage is substantially higher than a baseline of 49.5% (49/99) insight that we observed in our previous study with closely related experimental procedures, but without a nap period ($p = .007$, Fisher’s exact test, see Fig. 5B below; N = 99, data from A. T. Löwe et al., 2024). By the first half of block 8, insight participants had significantly higher average accuracy across all trial types ($M = 98.2 \pm 0.3\%$ vs $M = 86.4 \pm 0.9\%$, $t(22.86) = 12.28$, $p < .001$, $d = 4.26$), and lower reaction times ($M = 526.6 \pm 14$ vs $M = 767.4 \pm 30.4$, $t(27.4) = -7.19$, $p < .001$, $d = 2.2$), as expected. Hence, the 20 minute nap period significantly improved insight. Insight showed all three characteristics we observed in previous work: First, insight was selective, i.e. occurred only in some, but not all, participants (see above). Second, the timing of individual strategy switch points differed substantially across participants, indicating the highly variable delay known as impasse in the insight literature (block in which switch occurred: $M = 5.1 \pm 2.6$, range 3.6–6.2, Fig. 20F; analyses based on

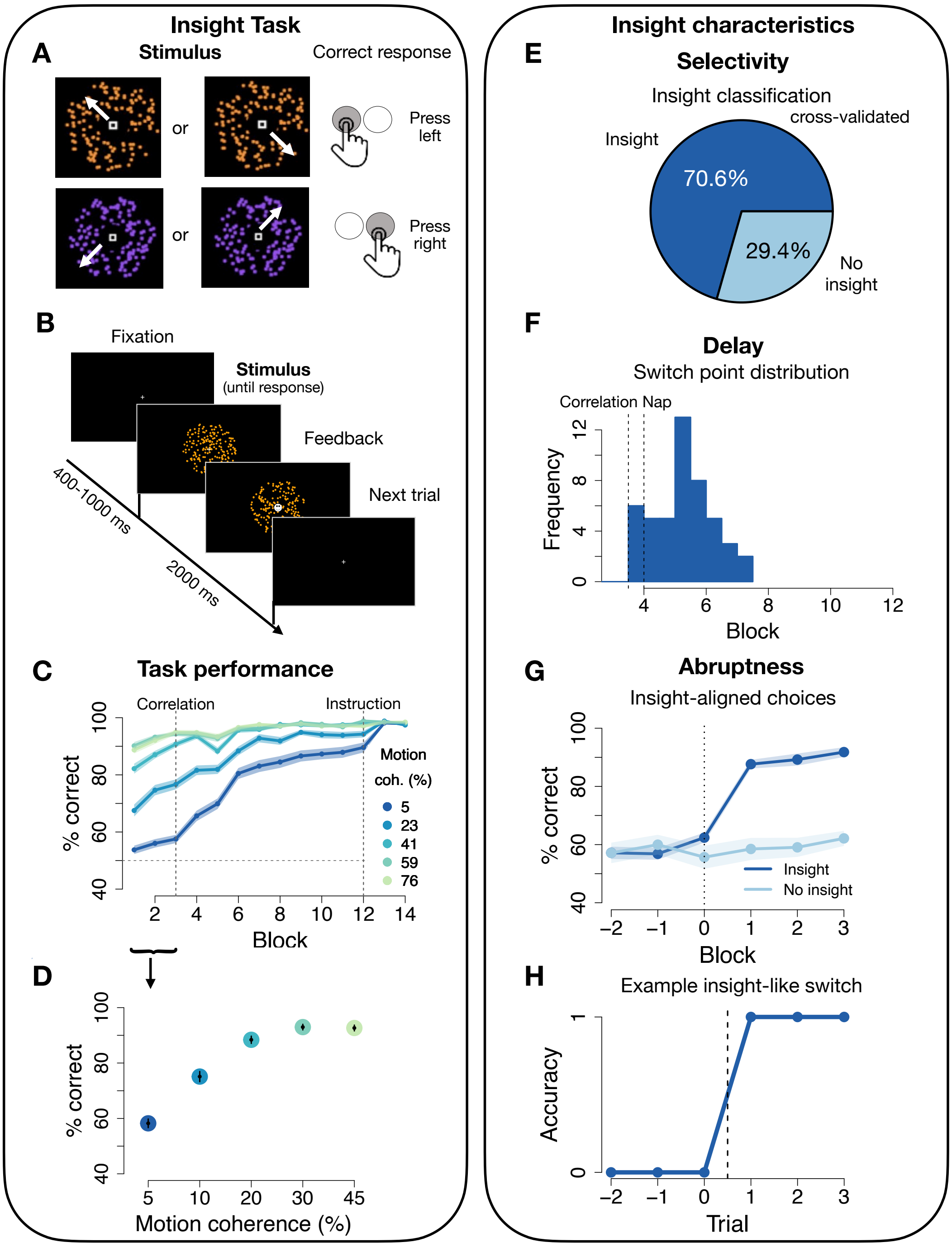


Figure 20: Task stimuli and behavioural results.

Caption on the next page.

Figure 20: **A:** Stimuli and stimulus-response mapping of the PSSST. Dot clouds were either coloured in orange or purple and moved to one of the four directions (NW, NE, SE, SW) with varying coherence. A left response key, ”X”, corresponded to the NW/SE motion directions, while a right response key ”M” corresponded to NE/SW directions. **B:** Trial structure: a fixation cue is shown for a duration that is shuffled between 400, 600, 800 and 1000 ms. The random dot cloud stimulus is displayed for 2000 ms. A response can be made during these entire 2000 ms, but a central feedback cue will replace the fixation cue immediately after a response. **C:** Accuracy (% correct) over the course of the experiment for all motion coherence levels. The first dashed vertical line marks the onset of the colour correlation, the second dashed vertical line the instruction about colour predictiveness. Blocks shown are halved task blocks (50 trials each). N = 90, error shadows signify standard error of the mean (SEM). **D:** Accuracy (% correct) during the motion phase increases with increasing motion coherence. N = 90, error bars signify SEM. **E:** 70.6% of subjects (48/68) were classified as insight subjects based on non-linear increases in performance on the lowest motion coherence level (5%). **F:** Distribution of switch points. The first dashed vertical line marks onset of the colour correlation, the second dashed vertical line the nap period. Blocks shown are halved task blocks (50 trials each). **G:** Switch point-aligned accuracy on the lowest motion coherence level for insight (48/68) and no-insight (20/68) subjects. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **H:** Trial-wise switch-aligned binary responses on lowest motion coherence level for an example insight subject.

logistic function fits, see Methods). Third, if participants had an insight, their accuracy increased very abruptly within a short time window, i.e. time-locking performance to their individual switch point indicated an average 25% performance jump within merely 15 trials ($M = 62.4 \pm 16.9\%$ vs $M = 87.6 \pm 15.1\%$, $t(92.8) = -11.16$, $p < .001$, Fig. 20G), which often reflected performance changes within a single trial only (Fig. 20H).

4.2.2 No Evidence For N1 but for N2 Sleep Promoting Insight

We followed the procedure of Lacaux et al. (2021) and divided participants into three groups based on their vigilance state during rest. Sleep was manually scored according to the guidelines from the American Academy of Sleep Medicine (Berry et al., 2016) based on 30 sec EEG (O2, O1, Pz, Cz, C3, C4, F3 and F4), EOG and EMG epochs. Using these criteria, participants were categorised as having had either no sleep, N1 sleep, or N2 sleep. This analysis showed that during the 20-minute nap period 28 participants reached N2 sleep, 22 reached only N1 sleep, and 18 subjects remained awake. Within the N2 group, 85.7% (24/28) gained insight into the hidden strategy, while only 63.6% (14/22) of participants in the N1 group and 55.5% (10/18) of the Wake group gained an insight in our task (Fig. 5B). We validated the manual sleep stage scoring with a convolutional neural network trained on external polysomnography data (U-Sleep, Perslev et al. (2021)). This categorisation correlated highly with manual scoring, $r(66) = 0.82$, $p < 0.001$), and results reported here can be replicated qualitatively using this alternative approach (see Supplemental Information (SI)). Similarly, splitting participants based on subjective sleep reports also results in the same pattern of results (see Fig 5, SI), although subjective reports did not match objective sleep staging closely (see SI).

Based on the paper by Lacaux et al. (2021), our main preregistered hypothesis proposed that N1 sleep would lead to an increased number of insight compared to the Wake and N2 sleep groups, respectively. We further hypothesised that N2 sleep would lead to decreased insight compared to N1. We find no support for either the first or second hypothesis (Fisher's exact test N1 vs. Wake: $p = 0.75$; N1 vs. N2: $p = 0.1$). To explain the above reported heightened incidence of insight after the nap generally, we explored whether N2 sleep was the main driver of insight. Interestingly, we observed a significantly higher number of insight after N2 sleep compared to Wake (Fisher's exact test, $p = 0.038$, Fig. 5B). In line with these analyses, a generalised linear model (GLM) with sleep stage as a predictor of insight fits the data better than a model with just an intercept (AIC 82.5 vs. 84.4). As expected, post-hoc tests also showed a significant N2 sleep coefficient in this model ($p = 0.03$), while N1 sleep and Wake remained non-significant (Wake: $p = 0.64$, N1: $p = 0.6$). Investigating Bayes Factors (BF) supports this finding and shows strong evidence for an effect of N2 $>$ N1 (BF = 24.71) as well as N2 $>$ Wake (BF = 8.19), while there is no substantial evidence for our preregistered hypotheses of N1 $>$ W (BF = 1.19) and N1 $>$ N2 (BF = 0.04). We thus find no evidence that N1 sleep promotes insight as reported by Lacaux et al. (2021). Instead, in our data N2 sleep showed a significant association with insight frequency.

The increased occurrence of insight in the N2 group had no major associations with overall performance after the nap. Accuracy on the lowest motion coherence trials only trended to be better in N2 compared to Wake participants (t-test block 5-12, N2 vs. Wake: $M = 85 \pm 3\%$ vs $M = 76 \pm 2.9\%$, $t(14) = 2.06$, $p = 0.06$, $d = 1.03$, N2 vs. N1: $M = 85 \pm 3\%$ vs $M = 81 \pm 2\%$, $t(12.1) = 1.06$, $p = 0.31$, $d = 0.53$, 23A). No effects on the corresponding reaction times could be found (N2 vs. Wake: $M = 757.6 \pm 48$ms vs $M = 809 \pm 35$ms, $t(12.8) = -0.86$, $p = 0.4$, $d = 0.43$, N2 vs. N1: $M = 757.6 \pm 48$ms vs $M = 787.8 \pm 45$ms, $t(13.9) = -0.46$, $p = 0.66$, $d = 0.23$, 23B). Thus, sleep seemed to increase insight frequency, but not alter overall performance characteristics.

To explore more directly whether the characteristics of insight differed between sleep groups, we next focused on the individually determined time points of insight, and participants' performance thereafter. We investigated differences in delay using the individually defined switch points in high noise trials (Fig. 20G,F; details see Methods), and found no significant differences across groups ($M_{\text{N2}} = 4.96 \pm 0.1\%; M_{\text{N1}} = 5.22 \pm 0.16\%; M_{\text{Wake}} = 5.21 \pm 0.15\%$, see Fig. 21C; all ts < 1.39, ps $> .18$). The switch point distributions also did not differ between groups (Kolmogorov-Smirnov test: N1–Wake: $D = 0.33$, $p = 0.47$, N1–N2: $D = 0.29$, $p = 0.36$, N2–Wake: $D = 0.33$, $p = 0.36$). Accuracy of insight subjects after their switch did not differ between sleep groups either ($M_{\text{N2}} = 90.9 \pm 0.3\%; M_{\text{N1}} = 94.5 \pm 0.3\%; M_{\text{Wake}} = 90.2 \pm 0.3\%$, see Fig. 21D; all ts < 1.06, ps $> .3$). Finally, we also found no group differences between reaction times after the insight ($M_{\text{N2}} = 688.4 \pm 42; M_{\text{N1}} = 607.1 \pm 54.7; M_{\text{Wake}} = 711 \pm 73$, see Fig. 21E; all ts < -0.27, ps $> .25$). Thus, while N2 sleep increased the prevalence of insight, it does not seem to affect its characteristics, i.e. abruptness, selectivity and delay.

4.2.3 Aperiodic Neural Activity Predicts Insight

Above, we performed pre-registered analyses investigating sleep stages and their impact on insight. They revealed that N2 sleep in particular is associated with insight. In a next step, we follow up on these findings with exploratory analyses investigating a potential association between insight and aperiodic activity. Our previous work on

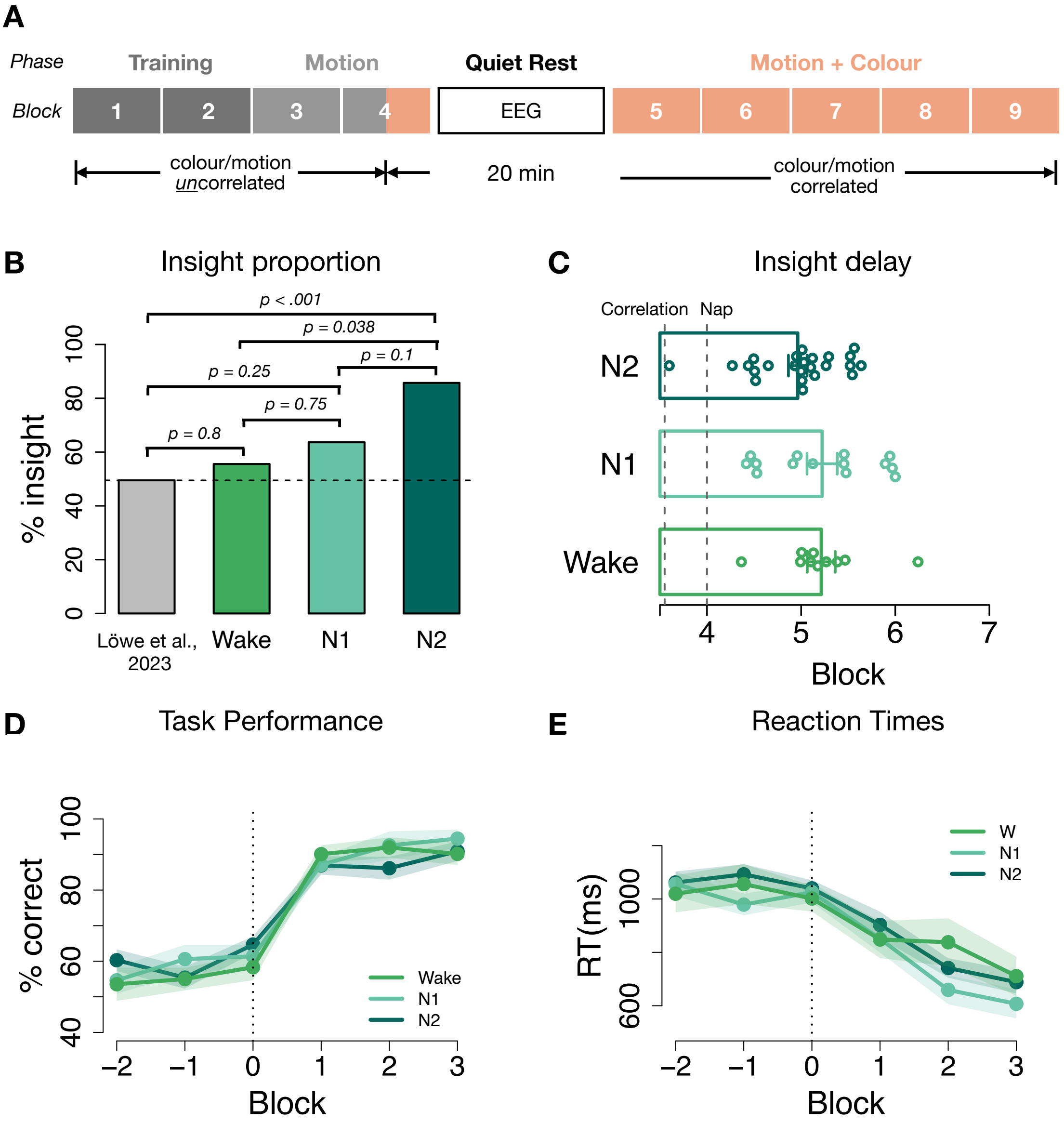


Figure 21: Task structure and behavioural results across sleep groups.

Caption on the next page.

Figure 21: **A:** Task structure of the PSSST: each block consisted of 100 trials. A first training block contained only 100% motion coherence trials to familiarise subjects with the S-R mapping. The remaining training block contained only high coherence (41%, 59%,76%) trials. In the motion phase, colour changed randomly and was not predictive and all motion coherence levels were included. Colour started to be predictive of correct choices and correlate with motion directions as well as correct response buttons in the second half of the 4th block to expose subjects to the hidden rule before the nap. Participants were then given 20 minutes to nap while EEG was recorded. Before the very last block 9, which served as sanity check, participants were instructed to use colour. **B:** Insight proportion among the different sleep groups. The insight ratio was significantly higher for the N2 sleep group (85.7%) than for the Wake group (55.5%). The N1 sleep group ratio (63.6%) did not differ significantly from the other two groups. The insight baseline ratio of 49.5% was derived from our previous work using the same task without a nap period. **C:** Distribution of switch points for the different sleep groups. One beeswarm point is one insight participant. Barplots show the mean, error bars signify SEM. **D:** Switch point-aligned accuracy and **E:** reaction times on the lowest motion coherence level for insight subjects of the respective sleep groups. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM.

neural networks (A. T. Löwe et al., 2024) suggests that noise as well as regularisation facilitate sudden and abrupt performance changes characterising insight. Although the precise mapping of these parameters in neural networks onto electrophysiological markers is unclear, noise (Voytek et al., 2015) and regularisation (as in synaptic downscaling, (Lendner et al., 2023)) have both been associated with aperiodic activity. Additionally, aperiodic activity has been shown to decrease along the sleep cycle, translating into a steeper spectral slope with deeper sleep (Lendner et al., 2020, 2023; Ameen et al., 2024). This led us to ask whether aperiodic activity during the nap period relates to insight, over and above the effects of sleep stages. We quantified aperiodic neural activity by the spectral slope of the power spectrum in log-log space (FOOOF algorithm by Donoghue et al., 2020, range 1-45Hz, 0.2Hz frequency resolution, 4sec epochs with 50% overlap). We verified that spectral slopes differ between the Wake, N1 and N2 groups, as expected (Lendner et al., 2020, 2023; Ameen et al., 2024). This showed a global association (across all channels) between the spectral slope and sleep stages ($p_{\text{cluster}} = 0.003$) such that the spectral slope was the steepest in the N2 group and the flattest in the Wake group (post-hoc t-tests, channel C4: Wake vs. N1:

$M_{\text{Wake}} = -1.30 \pm 0.08$ vs. $M_{\text{N1}} = -1.51 \pm 0.05$, $t(26.6) = 2.06$, $p = 0.05$, $d = 0.68$, N1 vs. N2: $M_{\text{N1}} = -1.51 \pm 0.05$ vs. $M_{\text{N2}} - 1.78 \pm 0.06$, $t(47.7) = 3.48$, $p = 0.001$, $d = 0.95$, Fig. 22A).

Our main question was whether the spectral slope relates to insight beyond the association between sleep stages and insight reported above. Given the substantial association between sleep stages and spectral slope, we used a nested model comparison approach and tested a baseline model containing only sleep stage as a predictor for insight against a model containing sleep stage and spectral slope. This showed that spectral slope over fronto-central areas improved insight prediction compared to the baseline model (e.g., channel C4: AIC: 82.5 vs. 81.2, Fig. 22B), with a steeper spectral slope relating to a higher insight likelihood (e.g., channel C4: $\beta = 1.86$). Interestingly, comparing this full model (with both sleep stage and spectral slope as predictors) with the more parsimonious model containing only the spectral slope showed that the spectral slope alone is the best predictor for insight, yielding the best of all considered models (e.g., channel C4: AIC: 81.2 vs. 78.8, Fig. 22B). As anticipated based on these results, contrasting participants with versus without insight also indicated clear differences in spectral slope ($p_{\text{cluster}} = 0.01$, Fig. 22C; for channel C4: Insight vs. No Insight: $M_{\text{Insight}} = -1.51 \pm 0.05$ vs. $M_{\text{NoInsight}} = -1.78 \pm 0.06$, $t(47.7) = 3.48, p = 0.001, d = 0.75$, Fig. 22D).

Investigation of oscillatory activity, in contrast, did not reveal any correlation with insight. Although oscillatory activity changed across sleep stages, and Lacaux et al. (2021) reported links between alpha and delta power and insight, we did not find such associations in our data (see SI for an overview of the analyses).

In conclusion, variations in aperiodic activity during a nap period predict whether participants will gain insight, with steeper spectral slopes, particularly over fronto-central areas, linked to higher insight likelihood. This association exists across sleep stages, and is stronger than previously described links between sleep stages or oscillatory power and insight.

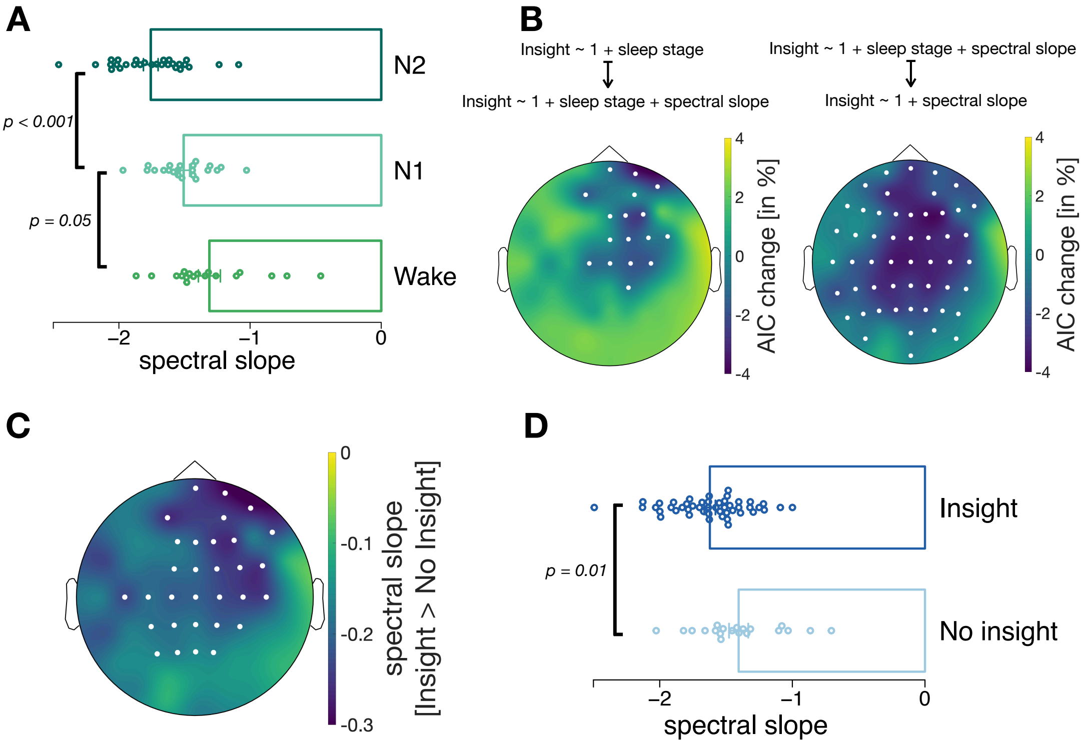


Figure 22: The spectral slope predicts insight.

A: The spectral slope significantly decreased from Wake to N1 to N2, as expected. For the corresponding topoplot see Supplemental Information, Fig. 25. **B:** Topographies of model comparison results testing a model of interest that included sleep stage and spectral slope (left) or only spectral slope (right) against a baseline model (left: insight $\sim$ 1 + sleep stage, right: insight $\sim$ 1 + sleep stage + slope). Shown are channel-wise model fit improvements obtained by including the spectral slope (left) or removing sleep stage (right; AIC in percentage, negative numbers indicate better fit of the main models). Channels with AIC differences < 0 were located over fronto-central areas (left) or central areas (right) and are highlighted in white. **C:** The spectral slope was significantly steeper (i.e., more negative) for participants with insight vs. participants without insight, over fronto-central areas. All channels that are part of the significant cluster are highlighted in white. **D:** The comparison of the spectral slope between participants with vs. without an insight for channel C4 (part of the significant cluster in C).

4.3 Discussion

We investigated the effect of sleep on insight. Our preregistered study set out to conceptually replicate findings of Lacaux et al. (2021), who reported that effects of sleep on insight were driven entirely by N1 sleep. While we did find a general increase in insight following the nap, the insight ratio of N1 subjects did not differ from subjects of the Wake group, thus providing no support for the hypothesis that N1 sleep fosters insight, contrary to (Lacaux et al., 2021). Instead, we found a beneficial effect of N2 sleep on post-nap insight likelihood, suggesting a need for deeper sleep for insight. Naturally, reaching N2 sleep implies longer total sleep duration. We thus cannot exclude the possibility that sleep duration instead of sleep stage is predictive of insight. However, an exploratory analysis showed that the 1/f slope of the power spectrum did explain additional variance in insight probability above and beyond sleep stages. In contrast, neither power in the alpha nor in the spindle frequency range could predict insight. Hence, aperiodic but not oscillatory neural activity emerged as an additional factor that promotes insight.

The 1/f slope has been linked to consciousness and sleep depth, where a steeper slope signifies less consciousness under anaesthesia, or deeper sleep (Miskovic, MacDonald, Rhodes, & Cote, 2019; Colombo et al., 2019; Lendner et al., 2020; Horváth et al., 2022; Schneider et al., 2022). Compared to ordinal sleep staging, the 1/f slope is a continuous measurement that offers a more fine grained measure of sleep depth. Hence, the fact that the spectral slope predicts insight beyond sleep stages alone supports the idea that deeper sleep is needed for insight.

This begs the question what the insight promoting processes during deeper sleep are. Our previous computational work (A. T. Löwe et al., 2024) pointed towards a role of regularisation and noise for the formation of insight. Proponents of the synaptic homeostasis hypothesis (Tononi & Cirelli, 2003, 2006, 2014) have related regularisation to synaptic downscaling (Hoel, 2021), a process that regulates synaptic strength depending on the synapses' firing rates during wake. By pruning synaptic connections with low activity, overall excitability is renormalised during sleep (Turrigiano & Nelson, 2004; Olcese, Esser, & Tononi, 2010; Hashmi, Nere, & Tononi, 2013). Computational work correlated this excitation-inhibition (E:I) balance with the spectral slope of aperiodic EEG activity (Gao, Peterson, & Voytek, 2017). Beyond just being a fine grained measure of sleep depth, the 1/f slope might thus reflect regularisation, which potentially plays an important role in generating insight.

It should be noted, however, that to date it is unclear if synaptic downscaling occurs during NREM sleep. Some evidence has linked E:I balance adjustments to REM sleep (Lendner et al., 2023), and evidence for synaptic downscaling during NREM sleep has remained indirect (Suppermpool, Lyons, Broom, & Rihel, 2024; Norimoto et al., 2018). Future work should thus investigate the role of sleep beyond NREM and include a full night of sleep.

What amount of regularisation is beneficial for insight is also uncertain. While our previous work (A. T. Löwe et al., 2024) has suggested that a certain amount of regularisation in neural networks leads to abrupt learning dynamics that characterise insight, either too little or too much regularisation caused the network to behave less insight-like. In the present study we only found a one directional relation, where deeper sleep and thus possibly more regularisation predicted insight. A speculative explanation for this might be that downscaling during N2 sleep of the nap led to a sort of reset of the previously learned synaptic weights which led participants to have a 'clean slate' after the nap, enabling them to restart the task with a fresh mind and discover the hidden rule more easily.

Lastly, why our findings diverge from what was reported by Lacaux et al. (2021) is unclear. A major difference between our studies is that we used the PSSST, while they used the NRT. The PSSST has crucial analogies in task structure to the NRT. Both tasks measure 'intrinsic' insight where the hidden rule as a potential for strategy improvement is never mentioned to participants, and both tasks can be solved in principle even if the hidden rule is not discovered, by using the initially learned rule. Besides the fact that our rule was much simpler, there are two major differences between these two insight tasks: first, the initial rule was learned via feedback in the PSSST, while it was instructed in the NRT; second, in our study the hidden rule became possible only after 350 trials, while for the NRT it is present from the start. This could imply potentially different learning mechanisms that could be affected differently by the respective sleep stages. Further, Lacaux et al. use occipital electrodes for oscillatory analyses, but our spectral slope results find an effect of aperiodic activity predicting insight in fronto-central electrodes (Fig. 22C).

While such differences do not allow inferences about the original finding, conceptual replications are important for validating broader scientific implications. How theoretical constructs such as insight are mapped onto specific tasks needs to be carefully evaluated, if one seeks to test the theoretical construct of interest. Further studies on the relationship between sleep and insight should therefore continue to evaluate different tasks, for instance one that is neither mathematical nor perceptual. Additionally, future work could also investigate the effect of a full night of sleep, rather than brief naps.

To conclude, the present study presents evidence of N2 sleep increasing insight likelihood, with the EEG spectral slope predicting insight beyond sleep stages. An exciting avenue for future studies will be to investigate the mapping between on-task EEG activity during insight moments to EEG activity during sleep and further examine potential relationships between the EEG spectral slope and regularisation in neural networks.

4.4 Methods

4.4.1 Behavioural Task

Participants

Participants between eighteen and 35 years of age were recruited via internal mailing lists as well as the research participation platform Castellum. Participation in the study was contingent on not having any learning difficulty nor colour blindness. Further, participants needed to report a normal sleep-wake cycle and no history of sleep disorders. Participants were excluded if they switched to the colour strategy immediately after the correlation onset, before the nap. All participants gave informed consent prior to beginning the experiment. The study protocol was approved by the local ethics committee of the Max Planck Institute for Human Development. Participants received 56€ for completing the entire experimental procedure.

Data inclusion was contingent on participants' showing learning of the stimulus classification. As in Chapter 2, we probed participants' accuracy on the three easiest, least noisiest coherence levels in the last block of the uncorrelated task phase. 30 subjects did not reach an accuracy level of at least 80% in those trials and were thus excluded from further analyses. Fifteen subjects were excluded, because the gained insight before the nap and further 7 subjects were excluded due to insufficient EEG data quality. The final sample included in all analyses thus contains 68 datasets.

Stimuli

We employed the PSSST described in Chapter 1, but adapted the motion coherence levels slightly. Dots were characterised by two features, (1) a motion direction (four possible orthogonal directions: NW, NE, SW, SE) and (2) a colour (orange or purple). The noise level of the motion feature was varied in 5 steps (5%, 23%, 41%, 59% or 76% coherent motion), making motion judgement relatively harder or easier. Colour difficulty was constant, thus consistently allowing easy identification of the stimulus colour. The condition with most noise (5% coherence) occurred slightly more frequently than the other conditions (30 trial per 100, vs 10, 20, 20, 20 for the other conditions).

The task was coded in JavaScript and made use of the jsPsych 6.1.0 plugins. Stimuli were presented on a 24 inch screen with a resolution of 1920 x 1200 pixel and a refresh rate of 59 Hz. On every trial, participants were presented a cloud of 200 moving dots with a radius of 7 pixels each. In order to avoid tracking of individual dots, dots had a lifetime of 10 frames before they were replaced. Within the circle shape of 400 pixel width, a single dot moved 6 pixel lengths in a given frame. Each dot was either designated to be coherent or incoherent and remained so throughout all frames in the display, whereby each incoherent dot followed a randomly designated alternative direction of motion.

The trial duration was 2000 ms and a response could be made at any point during that time window. After a response had been made via one of the two button presses, the white fixation cross at the centre of the stimulus turned into a binary feedback symbol (happy or sad smiley) that was displayed until the end of the trial. An inter trial interval (ITI) of either 400, 600, 800 or 1000 ms was randomly selected. If no response was made, a ”TOO SLOW” feedback was displayed for 300 ms before being replaced by the fixation cross for the remaining time of the ITI.

RDK Task Design

For the first 350 trials, the *motion phase*, the correct binary choice was only related to stimulus motion (two directions each on a diagonal were mapped onto one choice), while the colour changed randomly from trial to trial. For the binary choice, participants were given two response keys, ”X” and ”M”. The NW and SE motion directions corresponded to a left key press (”X”), while NE and SW corresponded to a right key press (”M”). Participants received trial-wise binary feedback (correct or incorrect), and therefore could learn which choice they had to make in response to which motion direction.

We did not specifically instruct participants to pay attention to the motion direction. Instead, we instructed them to learn how to classify the moving dot clouds using the two response keys, so that they would maximise their number of correct choices. To ensure that participants pick up on the motion relevance and the correct stimulus-response mapping, motion coherence was set to be at 100% in the first block (100 trials), meaning that all dots moved towards one coherent direction. In the second task block, we introduced the lowest, and therefore easiest, three levels of motion noise (41%, 59% and 76% coherent motion), before starting to use all five noise levels in block 3. Since choices during this phase should become solely dependent on motion, they should be affected by the level of motion noise.

After the *motion phase*, in the *motion and colour phase*, the colour feature became predictive of the correct choice in addition to the motion feature. This means that each response key, and thus motion direction diagonal, was consistently paired with one colour, and that colour was fully predictive of the required choice. Orange henceforth corresponded to a correct ”X” key press and a NW/SE motion direction, while purple was predictive of a correct ”M” key press and NE/SW motion direction. This change in feature relevance was not announced to participants, and the task continued for another 550 trials as before - the only change being the predictiveness of colour.

Before the last task block we asked participants whether they 1) noticed the colour rule in the experiment, 2) how long it took until they noticed it, 3) whether they used the colour feature to make their choices and 4) to replicate the mapping between stimulus colour and motion directions. We then instructed them about the correct colour mapping and asked them to rely on colour for the last task block. This served as a proof that subjects were in principle able to do the task based on the colour feature and to show that, based on this easier task strategy, accuracy should be near ceiling for all participants in the last instructed block.

Psychomotor Vigilance Task (PVT)

During the PVT, a white fixation cross was presented in the middle of the screen. After a delay (jittered with 4000±2000ms), the fixation cross changed its colour to red. The change in colour prompted participants to press the space key as fast as possible. On key press, participants received feedback about their reaction time for 2.5 sec. Overall, the PVT comprised 25 trials, corresponding to approximately 3 min. For results of the PVT see Supplemental Information, Fig. 27.

4.4.2 Experimental Procedure

The experimental procedure consisted of 3 parts: (1) a first behavioural session of about 25 minutes, including the PVT and 400 trials of the RDK task, followed by (2) a nap of 20 minutes and (3) a second behavioural session of about 30 minutes, including the PVT and 500 more trials of the RDK task.

(1) The experimental procedure began with the Pittsburgh Sleep Quality Index (PSQI) questionnaire. Participants then first completed the PVT and concluded with the first part of the RDK task of which the last 50 trials contained the hidden, easier strategy.

(2) Subsequently, participants were given time to rest and nap for 20 minutes. The EEG cabin was a completely dark and noise shielded room without sensory stimulation. During the nap break, participants were positioned in a semi-reclined position on an armchair with their legs resting on a foot piece, holding a light plastic cup in one hand. With the onset of N2-sleep this cup likely falls, waking participants up (see (Lacaux et al., 2021)). EEG recordings were exclusively recorded during this period and were used to identify different sleep stages. To increase the probability that people would fall asleep during the nap, sleep in the night before the experiment was reduced by 30% and participants were additionally asked to refrain from consuming caffeine prior to the session. All participants started the session at the same time of day at 1 pm.

(3) After the nap, participants resumed the behavioural testing and first performed a second PVT, followed by 500 more trials of the RDK task.

4.4.3 Modelling of Insight-like Switches

To investigate insight based strategy adaptations, we modelled participants' data using individually fitted sigmoid functions (for details see (A. T. Löwe et al., 2024)).

$$y = \frac{y_{max} - y_{min}}{1 + e^{-m(t-t_s)}} + y_{min}$$

The criterion defined in order to assess whether a subject switched to the colour strategy, is the accuracy in the highest noise level (5% coherence) in the last task block before the colour rule was explicitly instructed. Insight subjects are classified as those participants whose performance on those trials was above 85%. The individual insight moments t_s were derived from the individually fitted sigmoid functions.

4.4.4 EEG Analysis

EEG Recordings

During the nap period, EEG and electrooculography (EOG) data were recorded using a Brain Products 64-channel EEG system with a sampling rate of 1000 Hz. All electrodes were referenced online to A2 (right mastoid) and AFz was used as the ground electrode. Two external electrodes (biploar reference and ground electrode on the forehead) were placed on the chin to record muscle activity (electromyography, EMG). Impedances were kept below 20 kΩ.

Sleep Scoring

EEG and EOG data were re-referenced offline to linked mastoids and band pass filtered between 0.3 and 35 Hz (high pass filter: 0.3 Hz, two-pass butterworth filter, 3rd order; low pass filter: 35 Hz, two-pass butterworth filter, 5th order). EMG data were high pass filtered at 5 Hz (two-pass butterworth filter, 3rd order). Lastly, all data were downsampled to 200 Hz.

To identify different sleep stages, sleep was scored according to the guidelines from the American Academy of Sleep Medicine (AASM, (Berry et al., 2016)) based on EEG (O2, O1, Pz, Cz, C3, C4, F3 and F4), EOG and EMG data. Participants without any N1 or N2 period were assigned to the wake group. Participants who had at least 1 epoch (30 sec) of N1 and no signs of N2 (sleep spindles and/or K-complexes) were assigned to the N1 group. Participants with signs of N2 (sleep spindles and/or K-complexes) were assigned to the N2 group. For the AASM scoring, 30 sec epochs were used. Scoring was done by two scorers (ATL and MP), blind to the experimental condition. Additionally, we validated the scoring by a convolutional neural network trained on external polysomnography data (U-Sleep, Perslev et al. (2021), correlation with manual scoring: $r(66) = 0.82$, $p < 0.001$).

In addition to sleep stages, Lacaux et al. (2021) reported a modulation of insight by alpha and delta power across the whole nap period. To test for an additional modulation of insight by power of different frequency ranges, we used a data driven approach across the frequency spectrum of 1-20Hz (see section Spectral Analysis).

EEG Data Analysis

EEG analyses were conducted using the FieldTrip toolbox ((Oostenveld, Fries, Maris, & Schoffelen, 2011)) and custom scripts written in MATLAB. Independent component analysis (ICA) was applied to remove eye movement artifacts from the data. For that, data were re-referenced offline to linked mastoids, filtered (two-pass butterworth filter: high-pass: 1Hz, low-pass: 100Hz, bandstop: 48-52Hz) and down-sampled (200 Hz). Bad channels were removed and coarse artifacts were discarded based on outliers regarding amplitude and variance (implemented in $ftrejectvisual$). ICA was applied to identify components reflecting eye movements (saved together with the unmixing matrix). The raw data were then pre-processed again since previous pre-processing was optimised for ICA. Data were re-referenced to linked mastoids, filtered (two-pass butterworth filter: high-pass: 0.1Hz, low-pass: 48Hz) and down-sampled (200Hz). Bad channels were removed and the previously obtained unmixing matrix was applied to the data, components reflecting eye movements were removed and data were demeaned. Finally, bad channels were interpolated (spherical spline interpolation) and artifacts were visually identified.

Spectral Slope Analysis

To obtain estimates of aperiodic activity, the spectral slope parameter x (reflecting the slope of the power spectrum) was used. Data were segmented into 4 second epochs with an overlap of 50%. For these segments, power spectra were obtained by applying a Hanning window and transforming data from time to frequency domain using Fast Fourier Transformation. Power spectra were calculated for 1-45Hz with a frequency resolution of 0.2Hz. The FOOOF algorithm (Donoghue et al., 2020) was then applied to obtain the spectral slope. Aperiodic activity a is defined by:

$$a = 10^b * \frac{1}{(k + f^{\frac{1}{x}}})$$

where b is the y intercept, k is the knee parameter and x is the slope parameter.

4.4.5 Statistical Analyses

Fisher's exact tests were used in the analysis of contingency tables. All tests were two-tailed with a significance level of less than 0.05. All computations were performed using R version 4.3.1. For comparisons of spectral slopes between Wake, N1 and N2 groups or between participants with vs. without insight across all channels, a cluster-based permutation test was used (F-statistics for comparison between Wake, N1 and N2: 1000 permutation, alpha = 0.05 , clusteralpha = 0.05; t-statistics for comparison between Insight vs. No insight: 1000 permutation, alpha = 0.025 , clusteralpha = 0.05). For post-hoc comparisons, t-tests were applied. For model comparisons, we used the following logistic regression models for each EEG channel:

Model$_{\mathrm{baseline}}$: Insight $\sim$ 1 + sleep stage

Model$_1$: Insight $\sim$ 1 + sleep stage + slope

Model$_2$: Insight $\sim$ 1 + slope

AIC scores were used to assess the best model fit.

4.5 Supplementary Information

Self Reported Sleep Stages and U-Sleep

After participants completed the 20-minute nap break, we asked them whether they fell asleep (N2), were between sleep and wake (N1) or stayed awake (Wake) during that time. These ratings differed from the EEG based sleep scoring as only 13 participants indicated to have fallen asleep (N2), 32 reported to have stayed between sleep and wake, and 23 subjects reported to have stayed awake. We then assessed insight differences based on these sleep self reports. The insight proportions yield similar results as

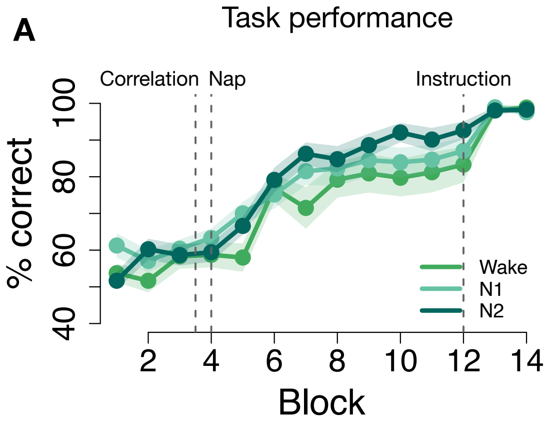


Figure 23: Accuracy and reaction times across sleep groups.

A: Accuracy and **B:** reaction times on the lowest motion coherence level for insight subjects of the respective sleep groups. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM.

the sleep staging based on EEG data: within the self reported N2 group, 92.3% (12/13) gained insight into the hidden strategy, while only 75% (24/32) of participants in the self reported N1 group and 52.2% (12/23) of the reported Wake group gained an insight in our task (Fig. 24A). When sleep stages were automatically defined with U-Sleep Perslev et al. (2021), 27 participants were categorised as Wake, 21 participants as N1 and 20 as N2. The insight proportions and statistical comparisons revealed similar results as the manually scored and subjective data: In the N2 group, 90% (18/20) gained insight but only 66.66% (14/21) and 59.26% (16/27) gained insight in the N1 and wake group respectively (Fisher's exact test N1 vs. Wake: p = 0.77; N1 vs. N2: p = 0.13; N2 vs. Wake: p = 0.025).

As for the EEG based results, the individually defined switch points in high noise trials (Fig. 20G,F; details see Methods), do not differ across reported sleep groups ($M_{\text{N2}} = 5.17 \pm 0.3; M_{\text{N1}} = 5.36 \pm 0.2; M_{\text{Wake}} = 4.84 \pm 0.2$, see Fig. 24B; all ts < 0.5, ps $> .11$).

After the nap, participants reporting to have slept (N2) perform significantly better than participants indicating to have stayed awake or to have been between sleep and wake ($M_{\mathrm{N2}} = 81 \pm 1.5\%; M_{\mathrm{N1}} = 76.9 \pm 2\%; M_{\mathrm{Wake}} = 73.6 \pm 2.5\%$, see Fig. 24C; N2 vs. W: $t(25.4) = 2.63$ $p = 0.014$, N2 vs. N1: $t(40.8) = 1.89$ $p = 0.07$). There was no such difference for reaction times ($M_{\mathrm{N2}} = 784.92 \pm 36.6; M_{\mathrm{N1}} = 816.54 \pm 29; M_{\mathrm{Wake}} = 861 \pm 35.3$, see Fig. 24C; all ts < 0.67, ps > 0.15).

This reported sleep effect again does not hold when considering data of insight subjects only. Accuracy does not differ between self reported sleep groups after the individually fitted switch points ($M_{\mathrm{N2}} = 91.3 \pm 3.3\%; M_{\mathrm{N1}} = 92.2 \pm 2.4\%; M_{\mathrm{Wake}} = 91.5 \pm 3.4\%$, see Fig. 24E; all ts < -0.05, ps $> .8$). Again, there was also no difference between reaction times after the insight ($M_{\mathrm{N2}} = 676.07 \pm 58.5; M_{\mathrm{N1}} = 668.06 \pm 46.2; M_{\mathrm{Wake}} = 665.33 \pm 58.4$, see Fig. 24F; all ts < 0.12, ps $> .9$).

We thus find the same results for self reported sleep groups as we do using the EEG based sleep staging: (1) N2 sleep significantly increases insight compared to Wake and (2) insight characteristics do not differ between subjects once insight has occurred.

No Evidence for Oscillatory Activity Predicting Insight

Additionally to sleep stages, Lacaux et al. (2021) found an association between insight and alpha and delta power. We pre-registered a data-driven analysis approach (including frequencies from 1-20 Hz) to test for a modulation of insight by power. To this end, we contrasted spectral slope corrected power spectra (FOOOF algorithm (Donoghue et al., 2020), 4 sec epochs, 1-20Hz, 0.2Hz frequency resolution, 50% overlap) between Wake, N1 and N2. Power spectra were calculated as described in the Methods section (Spectral Slope Analysis).

As expected, oscillatory power in the frequency range of 6-16 Hz significantly differed across all channels between Wake, N1 and N2 (cluster-based permutation test, F-statistics, $p_{\text{cluster}} = 0.005$). Post hoc cluster-based permutation tests revealed a positive and negative cluster in the alpha (5.8-11.3Hz) and sleep spindle frequency range (11.5-15.2Hz), respectively (post-hoc cluster-based permutation test, t-statistics, Wake $>$ N1: negative cluster, $p_{\text{cluster}} = 0.02$, 10.5-14Hz; Wake $>$ N2: positive cluster, $p_{\text{cluster}} = 0.07$, 6-9Hz; negative cluster, $p_{\text{cluster}} = 0.05$, 11.5-15.2Hz; N1 $>$ N2: positive

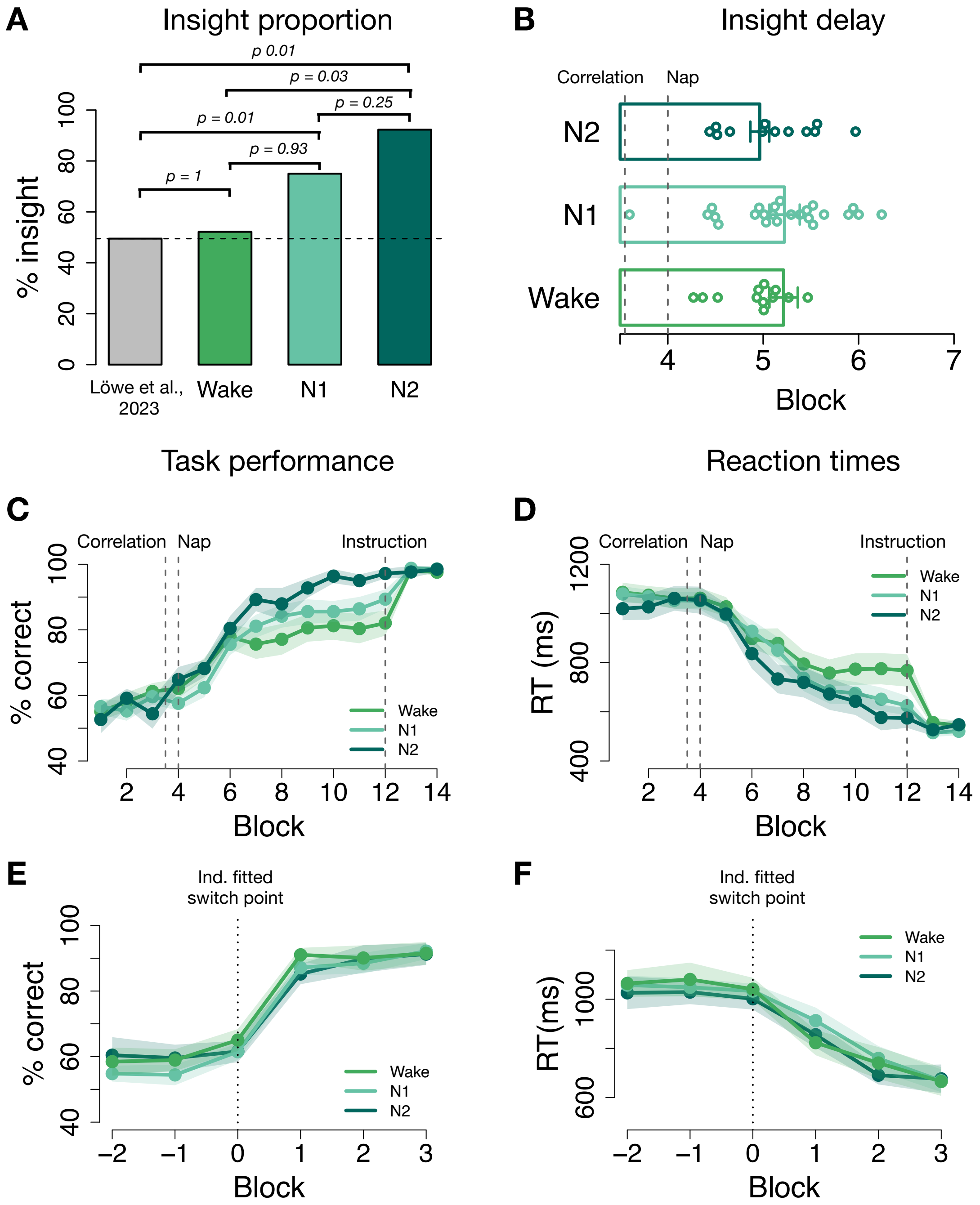


Figure 24: Behavioural results of self reported sleep groups.

Caption on the next page.

Figure 24: **A:** Insight proportion among the reported sleep groups. The insight ratio was significantly higher for people that reported to have slept deeply (N2) (92.3%) than for the reported Wake group (52.2%). The reported N1 sleep group ratio (75%) did not differ significantly from the other two groups. The insight baseline ratio of 49.5% was derived from our previous work using the same task without a nap period. **B:** Distribution of switch points for the self reported sleep stages. One beeswarm point is one insight participant. Barplots show the mean, error bars signify SEM. **C:** Accuracy and **D:** reaction times on the lowest motion coherence level for all subjects based on self reported sleep stages. Blocks shown are halved task blocks (50 trials each). Error shadow signifies SEM. **E** and **F** show data from **C** and **D** aligned to the individually fitted switch points for insight subjects only.

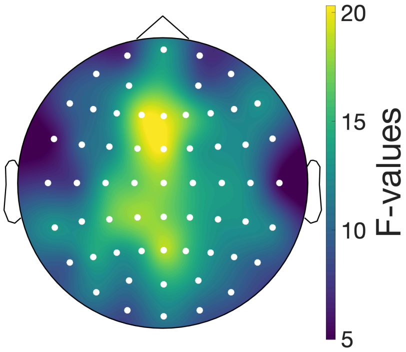


Figure 25: F-values of the comparison of the spectral slope between Wake, N1 and N2. The spectral slope significantly differs between Wake, N1 and N2 across all channels ($p_{\text{cluster}} = 0.003$).

cluster, $p_{\text{cluster}} = 0.007$, 5.8-12.3Hz). Neither averaged power in the alpha nor in the spindle cluster explained insight beyond sleep stages (AIC for model containing only sleep stages = 82.5; AIC for model with sleep stages + alpha power at channel C4 = 84.5; AIC for model with sleep stages + spindle power at channel C4 = 84.1, Fig. 26B). In line with the spectral slope analyses, we also removed sleep stages from both models. Removing sleep stages from both models resulted in a worse model fit (AIC for model with spindle power at channel C4 = 85.9; AIC for model with spindle power at channel C4 = 86.3, Fig. 26C). A complementary pattern emerges when directly contrasting participants with and without insight across the whole frequency range. No significant differences were observed (cluster-based permutation test, $p_{\text{cluster}} = 0.31$).

Together, these results suggest that oscillatory activity does not explain insight, neither alone nor in combination with sleep stages.

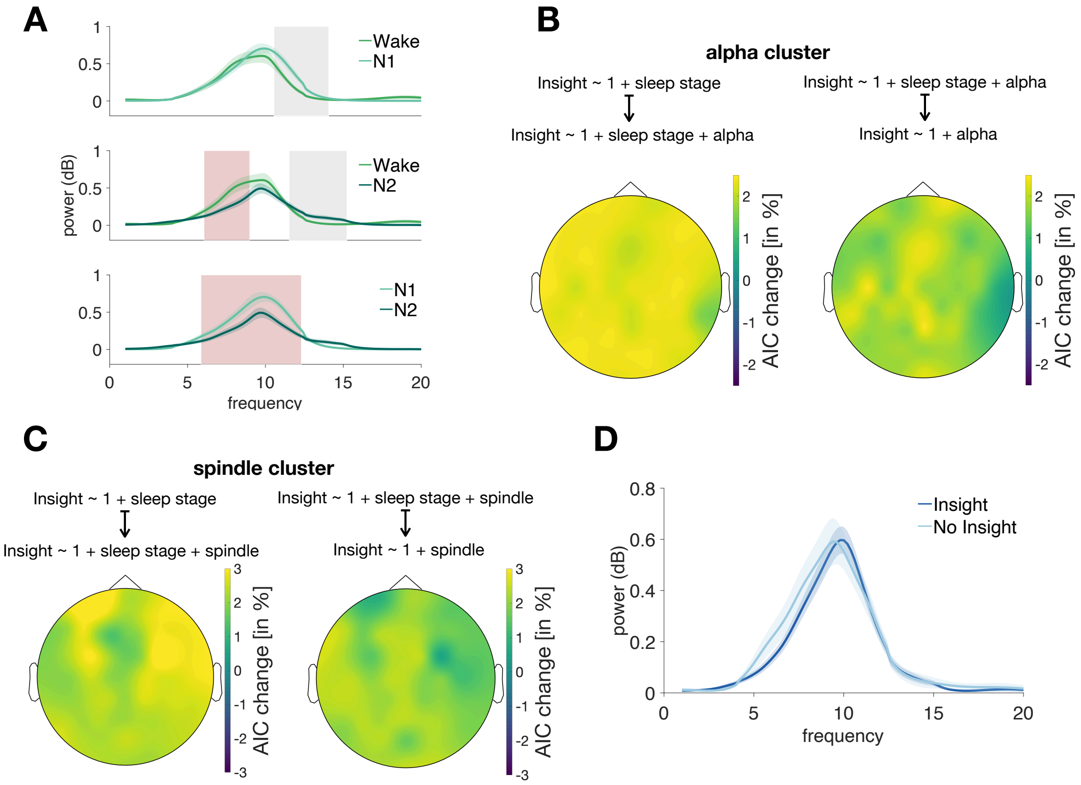


Figure 26: Oscillatory activity.

A: Overall, power significantly differs between Wake, N1 and N2. In grey, negative cluster are highlighted, in red positive cluster. **B:** Topographies of model comparison results testing a model of interest that include sleep stage and alpha power (left) or only alpha power (right) against a baseline model (left: insight $\sim$ 1 + sleep stage, right: insight $\sim$ 1 + sleep stage + alpha power). Shown are channel-wise model fit improvements obtained by including alpha power (left) or removing sleep stage (right; AIC in percentage). **C:** Topographies of model comparison results which can be interpreted as in B. Here, spindle power instead of alpha power is shown. **D:** There was no difference in power between participants with vs. without insight.

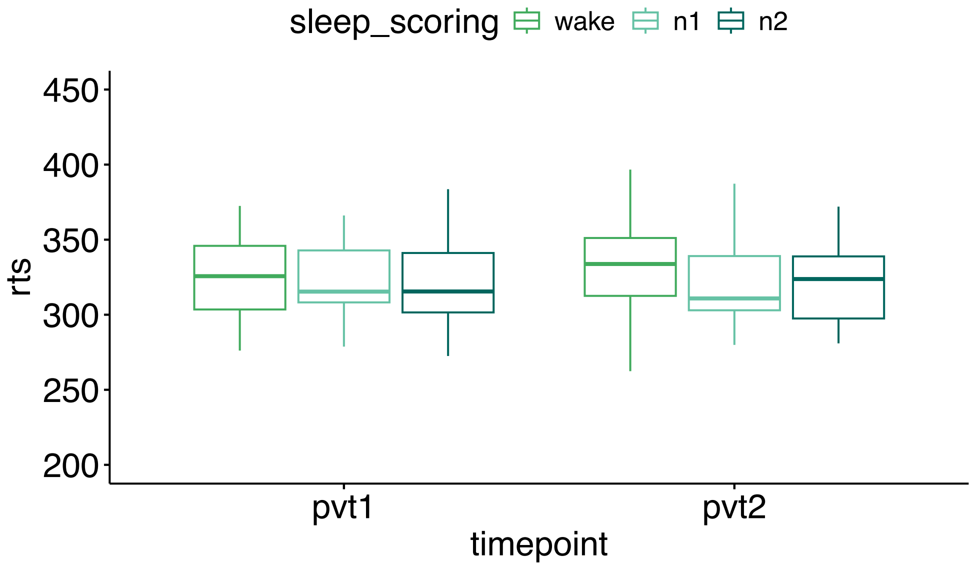


Figure 27: PVT results.

Before as well as after the nap period, participants' vigilance was assessed via a 3 min Psychomotor Vigilance Task (pvt1 = before, pvt2 = after nap period). Comparing reaction times between the Wake, N1 and N2 group before ($M_{\mathrm{Wake}} = 314.51 \pm 7.06ms; M_{\mathrm{N1}} = 317.92 \pm 4.55ms; M_{\mathrm{N2}} = 317.51 \pm 4.82ms$) and after ($M_{\mathrm{Wake}} = 324.71 \pm 7.44ms; M_{\mathrm{N1}} = 314.88 \pm 5.95ms; M_{\mathrm{N2}} = 312.93 \pm 4.18ms$) the nap period did not reveal any significant differences (linear model: rt$_{\mathrm{log}}$ $\sim$ 1 + sleep stage + time point; for all β: -0.045 $< \beta <$ 0.031, all $p >$ 0.2.)

5 General Discussion

Insight is considered to be a unique aspect of human cognition tied to creativity or meta-cognitive reasoning and is expressed in sudden understanding and drastic performance improvements. Although insightful problem solving has been a topic of investigation for over a century, the cognitive and neural mechanisms underlying insight are still debated.

In this thesis, I investigated insight as a learning phenomenon in humans as well as artificial neural networks. I further probed the role of sleep as an incubator for insight and the relation to computational mechanisms of insight-like learning.

The first section of **Chapter 1** offers an introduction into the rich history of insight research and besides introducing paradigms to study insight, describes the related key concepts of impasse, incubation, suddenness and affective components, as well as neural correlates and computational models of insight. Section 2 of the chapter introduces the PSSST as a hidden rule insight task to study learning of both a hidden and explicit strategy with high temporal resolution. Our paradigm offers a fine grained temporal window into the learning dynamics leading up to insight.

Chapter 2 uses the PSSST to characterise insight in human participants based on three attributes: suddenness, selectivity and delay. These behavioural characteristics permit to assess insight-like behaviour across natural and artificial intelligence and therefore gain a deeper understanding of the computational mechanisms involved in insight learning dynamics.

Chapter 3 compares these learning dynamics between humans and artificial neural networks. Our simple neural networks, which are devoid of complex cognitive processes, qualitatively and quantitatively align with human behavioural characteristics of insight. Analysing different network architectures and learning dynamics revealed that a regularised gating mechanism and noise added to gradient updates are crucial for eliciting insight-like behaviour. These mechanisms in combination allow the networks to accumulate “silent knowledge” that is initially suppressed by regularised gating. Our computational results hint at a link between behavioural markers of insight and measurements of noise and regularisation in the brain, therefore opening new avenues for investigating insight in biological agents.

Chapter 4 identifies a link between insight and N2 sleep as well as aperiodic neural activity. N2, but not N1 sleep increased insight likelihood on the PSSST after a daytime nap, implying a role of deeper sleep for insight. This is further supported by EEG power spectra showing that spectral slopes best predict insight, beyond sleep stages alone. Moreover, regularistaion in the form of synaptic downscaling has been associated with deeper sleep, further pointing towards a link between insight and regularisation.

5.1 Noise and Insight

Chapter 3 painted a clear image regarding the importance of noise for evoking insight: if no noise was added, not a single neural network would show insight-like behaviour. Gaussian – or white – noise was added during the gradient computations and in some cases accumulated to non-zero weights for the learnt irrelevant feature. These weights then aid in triggering non-linear learning dynamics once these initially irrelevant features become relevant. Importantly, insight does not arise if noise was added only to the learnt *relevant* feature, while adding it only to the *irrelevant* one quickly induced substantial amounts of insight.

Cumulative activity noise added to the networks' output did not have an effect on insight-like behaviour. This thus suggests that insight-like learning dynamics are unrelated to perception or outcome noise, but rather specific to internal processing as simulated through the gradient noise. These findings pose two important questions: (i) what might this simulated noise represent in biological brains and (ii) are individual differences in stimulus processing of a given insight task trait-like or random.

Previous studies have pointed towards stochastic brain dynamics being advantageous, allowing e.g. for creative problem solving and exploratory behaviour (Rolls & Deco, 2012; Faisal et al., 2008; Garrett et al., 2013; Waschke et al., 2021). Two omnipresent sources of noise in brain activity on the molecular level are cellular and synaptic noise (Faisal et al., 2008). Known benefits on the synaptic level are increased signal detectability and robustness which are both caused by probabilistic firing properties of neurons based on Poissonian spiking dynamics (Rolls et al., 2008). On a more conceptual level, noise has been implied in playing an important role in triggering spontaneous transitions between attractor states (Rolls & Deco, 2012).

On the macroscopic level of the brain, neural noise can describe trial-to-trial response variability (He, 2013). On the other hand, ”noise” in terms of variability of the entire signal reflects a direct measure of the moment to moment variability magnitude in a neuroimaging time series (Garrett et al., 2013). The only study to date investigating insight and neural variability, is a within-subject comparison of spectral power volatility, i.e. the amount of fluctuation between synchronised and desynchronised neural oscillations, which they found to be negatively associated with insight (Yu, Oh, Kounios, & Beeman, 2024). Solving CRA puzzles with insight was linked to significantly lower neural volatility during and before those trials as opposed to analytically solved trials that were linked to rapidly changing oscillatory synchronisation and desynchronisation (Yu et al., 2024). While this, contrary to our findings, would plead for less ”noise” being conducive to insight, it is not yet understood how different variability measures such as standard deviation and volatility relate to each other (Garrett et al., 2013). Further, they may not necessarily rely on the same generating neural mechanism (Waschke et al., 2021).

Selective attention has been found to result in a topographically specific reduction of noise correlations in non-human animals, while human performance similarly improves with attention-related decreases in low-frequency power (Waschke et al., 2021; Cohen & Maunsell, 2009; J. F. Mitchell, Sundberg, & Reynolds, 2009; Rabinowitz, Goris, Cohen, & Simoncelli, 2015). This would be in line with our findings that identified the interplay of gated regularisation and noise as necessary conditions for evoking insight-like behaviour. If the attention-like mechanism of gated regularisation is (too) strong, this suppresses the effect of noise that accumulates to non-zero weights in the unattended feature dimension. If attention would be lower and thus wider spread, including also task-irrelevant features, this can lead to insight about the changed feature relevance.

Besides affecting behaviour on an intra-individual level, neural variability has been suggested to explain behavioural differences in a ”trait-like” manner, since neural variability correlates across tasks and appears stable within individuals (Waschke et al., 2021; Grossman et al., 2019; Wehrheim, Faskowitz, Schubert, & Fiebach, 2024). The question whether there could be a ”noise-trait”, marked by specific inter-individual differences in neural variability, that could predict insight, remains an interesting topic for future research. It could be advantageous to correlate different attentional traits across tasks, including task switching and cognitive control paradigms, to insight-like problem solving propensities. Further, since different insight tasks rely on vastly different cognitive processes, it is important to also probe the link between insight and neural variability parameters with regard to different cognitive domains. Understanding inter-individual differences in insight behaviour and their underlying neural bases and computations will be a crucial advance to understanding the insight phenomenon.

5.2 Regularisation and Insight

Besides noise, regularisation was another factor necessary for insight-like behaviour in our neural network simulations: the number of insight networks increased linearly with the regularisation parameter λ. Furthermore, since regularisation forced the gates of the irrelevant feature towards 0, this caused a sustained suppression period before the switch to the hidden rule, where the networks were ”blind” to the solution, similar to an impasse observed in humans.

Regularisation as a simplification process is commonly used in machine learning to improve performance and avoid overfitting (Liu et al., 2020). Generally, two classes of regularisation are distinguished: explicit and implicit regularisation. While explicit regularisation adds a penalty term to the optimisation process, implicit regularisation happens passively through mechanisms such as dropout, weight sharing or early stopping. It should be pointed out that albeit we used a specific case of explicit (L1-)regularisation, different implicit regularisation forms might exist and be possibly implemented in the brain. While it is unknown to what extent our results generalise to other regularisation techniques, I am going to explore potential implementations of regularisation mechanisms in the brain below.

On a molecular level, regularisation has been linked to synaptic downscaling during sleep, a process that maintains a synaptic firing homeostasis by adjusting synaptic weights (Lee & Kirkwood, 2019). This scaling process aids stable energy requirements and may avoid memory interference (De Vivo et al., 2017). As a sliding threshold mechanism, it modulates the threshold for inducing long-term potentiation (Lee & Kirkwood, 2019). The selective renormalisation of synapses through regularisation during sleep might thus be an ideal candidate for gist extraction and therefore contribute to insight-like learning phenomena.

On a higher cognitive level, regularisation has been implied in heuristics. It has been proposed in the context of an infinitely strong prior in a Bayesian inference framework (Parpart et al., 2018), working as a sort of attention mechanism that regularises inputs and information in a way that is congruent with the specific prior (Parpart et al., 2018). Since a prior functions as an explicit term, this heuristic model would be classified as explicit regularisation. Another idea of an explicit mechanism is that cognitive control could be regarded as regularised optimisation in a multi-tasking paradigm (Ritz et al., 2022). Better transfer learning between tasks would be enabled by effort costs regularising towards more task-general policies.

Cognitively, regularisation thus seems to be generally implied in attention dynamics. Both these cognitive mechanisms of heuristics and cognitive control seem theoretically reconcilable with our results of regularisation working as an attention mechanism that leads to suppression of initially irrelevant inputs and therefore a prolonged impasse period. Taken together with the molecular mechanisms of avoiding overfitting and downscaling to extract informational gist, this suggests specific neural states that might be beneficial for insight.

How exactly these states are governed though and in what precise way regularisation relates to insight on a systems level remains to be understood. It would be particularly interesting to compare different implementations of regularisation mechanisms on a computational level and test their translation to biological models. Since regularisation is so ubiquitously used in artificial neural networks with great success for learning, it seems important to understand the embedding of regularisation in biological neural networks as well.

5.3 Sleep and Insight

Ever since Wagner et al. (2004) published their seminal paper ”Sleep Inspires Insight”, researchers sought to investigate this relationship further. Since sleep is linked to memory consolidation (Rasch & Born, 2013) and restructuring (Cowan et al., 2020), it would represent an ideal candidate for the representational restructuring postulated to occur during incubation. While some researchers have failed to replicate beneficial effects of sleep on insight problem solving (Cordi & Rasch, 2021; Schönauer et al., 2018; Brodt, Inostroza, Niethard, & Born, 2023), others found sleep effects to be specific to certain sleep phases (Lacaux et al., 2021; Vickrey & Lerner, 2023; Verleger et al., 2013).

The human sleep cycle oscillates between rapid eye movement (REM) and non REM (NREM) sleep, both marked by characteristic neurophysiological signatures. The cycle begins with the first NREM stage, N1 sleep, and gradually progresses deeper into sleep until reaching deep, slow wave sleep (SWS) in N3. Light N1 sleep shows reduced alpha band activity, while deeper N2 sleep is marked by sleep spindles and k-complexes and the deepest N3 sleep by slow delta waves. REM sleep, as the last stage of the cycle, is characterised by EEG activity that is similar to wakefulness and is usually associated with vivid dreaming.

One study exposing participants to pink noise during sleep found that this procedure altered the sleep architecture by limiting N1 sleep after sleep onset and eliminated the sleep-dependent benefit for insight (Vickrey & Lerner, 2023), similar to effects found by Lacaux et al (2021). In our preregistered nap study however, we found that N2 sleep significantly increased insight. Further, the spectral slope of the aperiodic EEG activity predicted insight above sleep stage alone, whereas neither alpha power nor the spindle frequency range were predictive. Compared to ordinal sleep staging, the 1/f slope offers a continuous way to measure sleep depth. A steeper spectral slope predicting insight thus implied that deeper sleep is needed for insight.

While our nap intervention was not sufficiently long for participants to reach N3 sleep, several findings suggest a special role of particularly SWS for insight. Increased alpha band activity in SWS, although not significant in our study, has been found to be predictive of post sleep insight, potentially reflecting representational restructuring happening during deeper sleep (Yordanova et al., 2012). Another study found the same SWS specific effect additionally in the beta frequency range, but interpreted particularly the oscillatory 10 Hz patterns during SWS to imply neocortical read-out of implicitly learned information stored in the hippocampus (Verleger et al., 2013). This SWS specific gist extraction might work through means of overlapping memories being replayed close together in time. Replay describes the process of reactivating sequential experiences on a compressed time scale either during sleep or wake rest (Foster, 2017). The temporally proximal replay mentioned above could thus lead to overlapping areas being strengthened through Hebbian plasticity (Lewis & Durrant, 2011). Besides SWS, REM sleep has also been implied as being critical for representational restructuring (Lewis, Knoblich, & Poe, 2018). It has been proposed that particularly the combination of SWS and REM over several nights might be beneficial for creativity, through alternating iteratively between generating cortically represented schemas in NREM, and forming links between these and other information during REM (Lewis et al., 2018).

The question thus arises whether insight might be a product of memory replay. Indeed, replay during SWS could be beneficial for gist extraction, because the neocortex is biased by the hippocampus to replay thematically linked memories in this phase (Lewis et al., 2018). REM replay on the other hand might aid in removing unnecessary constraints by detecting and reducing overlap in information structures (Lewis et al., 2018). Replay in SWS and REM could hence be characterised by differential signatures that could work either alone or together to promote insight. While REM with its high plasticity setting would be ideal for forming new connections, SWS would complement this by amplifying learnt information and extracting the relevant bits (Lewis et al., 2018).

Another candidate mechanism for gist extraction however might be regularisation (Nere, Hashmi, Cirelli, & Tononi, 2013). Synaptic downscaling, the process regulating synaptic strength based on firing rates during wake, has been related to regularisation by proponents of the synaptic homeostasis hypothesis (Tononi & Cirelli, 2006, 2003, 2014; Hoel, 2021). This firing rate balance, often referred to as excitation-inhibition (E:I) balance, has on a computational level further been linked to the spectral slope of aperiodic EEG activity, where a reduced E:I balance is reflected in a steeper spectral slope (Gao et al., 2017). This implies that beyond offering a continuous, fine-grained measure of sleep depth and consciousness, the spectral slope might be linked to insight by reflecting regularisation, which has been found to be beneficial for insight in our neural networks.

Several speculative mechanisms of how regularisation might promote gist extraction are imaginable. On the one hand, the regularising mechanism could downscale redundant synapses, therefore aiding gist extraction as only the relevant connections are maintained. On the other hand, previously learned synaptic weights could be reset completely, leading to a sort of 'clean slate' upon waking up, allowing to face problems with a fresh mind which could then lead to insight. A third hypothesis might be that the synaptic downscaling must not affect the information structure directly, but rather lead to a tagging of learnt relevant information to guide prioritised replay.

Another possibility is that regularisation primarily promotes gist extraction while replay promotes generation of novel inferences. Different cognitive tasks might thus benefit from these mechanisms in different ways – and potentially explain divergent findings regarding the relationship of sleep and insight. In support of this, a review on the relationship of sleep and extracting hidden regularities found a strong task dependence (Lerner & Gluck, 2019). Tasks like the PSSST or NRT can lead to insight if participants overcome the initially learned strategy. Both downscaling of redundant information, as well as downcaling to a 'clean slate' could account for the post sleep insight. In favour of the clean slate hypothesis would be that subjects often take a while after waking up to discover the hidden rule (Wagner et al., 2004; Lacaux et al., 2021; A. Löwe et al., 2024), i.e. they don't wake up with the insight solution. Replay on the other hand, might be beneficial for generating new connections, such as would be required for insight puzzles. It would also be expected that based on this mechanism, subjects should know the solution upon waking and restarting the task.

To further study the relationship between sleep and insight, it is therefore important to continue to evaluate different tasks and carefully scrutinise the different sleep stages. Since both replay and regularisation in humans are often only inferred or implicitly measured, it would be beneficial to try to test these mechanisms in animal models as well, potentially even in combination with insight paradigms.

The above further evokes the question on how sleep, regularisation and noise relate. While Chapter 3 suggested that regularisation in neural networks leads to insight-like learning dynamics, either too little or too much regularisation caused the network to behave less insight-like (A. T. Löwe et al., 2024). Chapter 4 however pointed towards a one directional relation of regularisation and insight, where deeper sleep and thus possibly more regularisation were predictive of insight. Speculatively, it could be conceivable that regularisation plays different roles during wake and sleep. While during sleep regularisation might aid gist extraction by pruning irrelevant connections and therefore aiding insight, during wake it might function as an attention mechanism that would make it impossible to overcome the blindness to the alternative solution if the regularisation was too constraining. In the latter hypothetical case, noise would be needed to aid in jumping out of current attractor state, but would not necessarily play a role for insight during sleep.

5.4 Is Insight Special?

Although insight has been extensively researched for more than a century, one is left with a picture of a highly heterogeneous, task-dependent phenomenon. This begs the question whether we are all talking about the same insight.

Most insight papers will motivate their research by trying to understand creative processes, breakthrough moments of scientific discovery or deep understanding. It is obvious that the way insight is investigated in the lab, trying to solve puzzles or word riddles, is far from these kind of cognitive phenomena. A crucial difference between the classic insight fables of Archimedes' primal Eureka moment or August Kekulé seeing the structure of benzene before his eyes and laboratory insight tasks is the *goal*. Most insight problems measure a binary outcome: solving with insight vs solving analytically. It thus seems that the primary goal here is to reach insight, not performance. In real life though, we are never set on reaching a solution to a problem through insight – we merely want to find the solution. Our goal is performance.

Often, we are not even aware of solutions we could have an insight about, because our focus lies on our current representational space. Some researchers distinguish the moment of insight, the finding of the problem solution, from the Aha! experience with its physical components of relief and pleasure (Danek et al., 2014; Gick & Lockhart, 1996). It is possible that sudden behavioural changes occurring during learning could lead to the subjective effects accompanying insight. We showed that even simple gradient-based neural networks, devoid of any complex cognitive mechanisms, exhibited behavioural signatures of insight. Further, fast neural transitions accompanying abrupt behavioural changes have even been shown in rodents (Durstewitz et al., 2010). This implies that insight is maybe not a mysterious, unique cognitive process after all, but a ubiquitous learning phenomenon.

The question emerges whether insight is not special after all. I would make a claim for a *spectrum* of insight that can take on various forms, spanning all cognitive realms and ranging from small 'Aha!'s to profound 'Eureka!'s, depending on the structures implied. At the base of it all lies a learning process, leading to abrupt, sudden performance improvements, that are delayed and happen selectively. This learning process will differ based on the task used – because learning is different with respect to the cognitive process involved. How this learning process maps onto neural correlates and task differences remains to be fully understood, just like the base cognitive processes. Insight might not involve magic, but it is a fundamental learning and memory phenomenon that lies at the root of big discoveries as well as insignificant mental processes. This, seems quite special to me.

5.5 Concluding Remarks

In this thesis, I present a new method to behaviourally measure and characterise insight. I further provide evidence towards understanding the computational mechanisms underlying insight. Moreover, I identified a link between insight and N2 sleep as well as the spectral slope of EEG activity. It remains to be understood how exactly noise and regularisation, implied to cause insight in neural networks, map onto the human brain and what precisely the underlying neural correlates of insight are in general. Another exciting avenue for future research is to probe different sleep phases and mechanisms such as synaptic downscaling and replay with respect to the generation of insight. All in all, the findings presented here begin to elucidate the insight process as a learning phenomenon and how deep sleep relates to this.

Glossary

AASM American Academy of Sleep Medicine. 105

AIC Akaike Information Criterion. 92, 96, 97, 107, 111, 112

aSTG anterior superior temporal gyrus. 14

BF Bayes Factor. 92

BIC Bayesian Information Criterion. 32, 34, 56, 59, 82

CRA compound remote associates. 11, 13, 116

EEG electroencephalography. 14, 16, 18, 19, 85, 87, 89, 91, 95, 99–101, 104–109, 115, 119, 121, 124

EII explicit-implicit interaction theory. 16, 67, 68

EMG electromyography. 91, 105

EOG electrooculography. 91, 105

fMRI functional magnetic resonance imaging. 15, 21

ICA Independent Component Analysis. 106

IFG inferior frontal gyrus. 15

ITI inter trial interval. 22, 37, 102

NAcc nucleus accumbens. 15

NREM non rapid eye movement. 8, 99, 119, 120

NRT number reduction task. 14, 15, 25, 35, 87, 99, 100, 121

PSSST Perceptual Spontaneous Strategy Switch Task. 18, 20, 25, 28, 35, 36, 51, 88, 91, 95, 99, 101, 114, 115, 121

PVT Psycho motor vigilance task. 1, 103, 104, 113

RAT remote associates task. 11, 13, 14

RDK random dot kinematogram. 103, 104

REM rapid eye movement. 99, 119, 120

SEM standard error of the mean. 34, 45, 59, 64, 65, 78, 80, 82, 83, 91, 95, 108, 111

SGD stochastic gradient descent. 48, 53

SWS slow wave sleep. 120

References

Allegra, M., Seyed-Allaei, S., Schuck, N. W., Amati, D., Laio, A., & Reverberi, C. (2020). Brain network dynamics during spontaneous strategy shifts and incremental task optimization. *NeuroImage*, *217*, 116854. doi: 10.1016/j.neuroimage.2020.116854

Ameen, M. S., Jacobs, J., Schabus, M., Hoedlmoser, K., & Donoghue, T. (2024). The Temporal Dynamics of Aperiodic Neural Activity Track Changes in Sleep Architecture. *bioRxiv*. doi: 10.1101/2024.01.25.577204

Aziz-Zadeh, L., Kaplan, J. T., & Iacoboni, M. (2009). ”Aha!”: The neural correlates of verbal insight solutions. *Human Brain Mapping*, *30*(3), 908–916. doi: 10.1002/hbm.20554

Becker, M., Yu, Y., & Cabeza, R. (2023). The influence of insight on risky decision making and nucleus accumbens activation. *Scientific Reports*, *13*(1), 1–14. doi: 10.1038/s41598-023-44293-2

Beeman, M., & Bowden, E. M. (2000). The right hemisphere maintains solution-related activation for yet-to-be-solved problems. *Memory & Cognition*, *28*(7), 28.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum Learning. In *In: Proceedings of international conference on machine learning* (pp. 41–48). Retrieved from http://arxiv.org/abs/1611.06204

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Robin, M. L., & Marcus, C. L. (2016). American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events : Rules, Terminology, and Technical Specifications, Version 2.2. *American Academy of Sleep*, *28*(3), 391–397.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer US. doi: 10.1007/978-3-030-57077-4{_}11

Bowden, E. M. (1997). The Effect of Reportable and Unreportable Hints on Anagram Solution and the Aha! Experience. *Consciousness and Cognition*, *6*(4), 545–573. doi: 10.1006/ccog.1997.0325

Bowden, E. M., & Jung-beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments & Computers*, *35*(4), 634–639.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, *9*(7), 322–328. doi: 10.1016/j.tics.2005.05.012

Brodt, S., Inostroza, M., Niethard, N., & Born, J. (2023). Sleep—A brain-state serving systems memory consolidation. *Neuron*, *111*(7), 1050–1075. doi: 10.1016/j.neuron.2023.03.005

Brodt, S., Pöhlchen, D., Täumer, E., Gais, S., & Schönauer, M. (2018). Incubation, not sleep, AIDS problem-solving. *Sleep*, *41*(10), 1–11. doi: 10.1093/sleep/zsy155

Cohen, M. R., & Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, *12*(12), 1594–1600. doi: 10.1038/nn.2439

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, *10*(3). doi: 10.1371/journal.pbio.1001293

Colombo, M. A., Napolitani, M., Boly, M., Gosseries, O., Casarotto, S., Rosanova, M., ... Sarasso, S. (2019). The spectral exponent of the resting EEG indexes the presence of consciousness during unresponsiveness induced by propofol, xenon, and ketamine. *NeuroImage*, *189*, 631–644. doi: 10.1016/j.neuroimage.2019.01.024

Cordi, M. J., & Rasch, B. (2021). How robust are sleep-mediated memory benefits? *Current Opinion in Neurobiology*, *67*, 1–7. doi: 10.1016/j.conb.2020.06.002

Costa, R. P., Assael, Y. M., Shillingford, B., De Freitas, N., & Vogels, T. P. (2017). Cortical microcircuits as gated-recurrent neural networks. *Advances in Neural Information Processing Systems*, 272–283.

Cowan, E., Liu, A., Henin, S., Kothare, S., Devinsky, O., & Davachi, X. L. (2020). Sleep spindles promote the restructuring of memory representations in ventromedial prefrontal cortex through enhanced hippocampal–cortical functional connectivity. *Journal of Neuroscience*, *40*(9), 1909–1919. doi: 10.1523/JNEUROSCI.1946-19.2020

Craig, M., Ottaway, G., & Dewar, M. (2018). Rest on it: Awake quiescence facilitates insight. *Cortex*, *109*, 205–214. doi: 10.1016/j.cortex.2018.09.009

Cranford, E. A., & Moss, J. (2012). Is Insight Always the Same? A Protocol Analysis of Insight in Compound Remote Associate Problems. *The Journal of Problem Solving*, *4*(2), 128–153. doi: 10.7771/1932-6246.1129

Danek, A. H., Fraps, T., von Müller, A., Grothe, B., & Öllinger, M. (2014). It's a kind of magic-what self-reports can reveal about the phenomenology of insight problem solving. *Frontiers in Psychology*, *5*, 1–11. doi: 10.3389/fpsyg.2014.01408

Darsaud, A., Wagner, U., Balteau, E., Desseilles, M., Sterpenich, V., Vandewalle, G., … Maquet, P. (2011). Neural precursors of delayed insight. *Journal of Cognitive Neuroscience*, *23*(8), 1900–1910. doi: 10.1162/jocn.2010.21550

De Vivo, L., Bellesi, M., Marshall, W., Bushong, E. A., Ellisman, M. H., Tononi, G., & Cirelli, C. (2017). Ultrastructural evidence for synaptic scaling across the wake/sleep cycle. *Science*, *355*(6324), 507–510. doi: 10.1126/science.aah5982

Dietrich, A., & Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin*, *136*(5), 822–848. doi: 10.1037/a0019749

Donoghue, T., Haller, M., Peterson, E. J., Varma, P., Sebastian, P., Gao, R., … Voytek, B. (2020). Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, *23*(12), 1655–1665. doi: 10.1038/s41593-020-00744-x

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486. doi: 10.1126/science.1252254

Dubey, R., Ho, M., Mehta, H., & Griffiths, T. L. (2021). Aha! moments correspond to metacognitive prediction errors. *PsyArxiv*.

Duncker, K. (1945). *On problem solving* (Vol. 58; J. F. Dashiell, Ed.) (No. 5). The American Psychological Association, Inc.

Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, *5*(2), 94–97. doi: 10.1111/j.1467-9280.1994.tb00637.x

Durstewitz, D., Vittoz, N. M., Floresco, S. B., & Seamans, J. K. (2010). Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, *66*(3), 438–448. doi: 10.1016/j.neuron.2010.03.029

Esser, S., Lustig, C., & Haider, H. (2022). What triggers explicit awareness in implicit sequence learning? Implications from theories of consciousness. *Psychological Research*, *86*(5), 1442–1457. doi: 10.1007/s00426-021-01594-3

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303. doi: 10.1038/nrn2258

Fleck, J. I., & Weisberg, R. W. (2013). Insight versus analysis: Evidence for diverse methods in problem solving. *Journal of Cognitive Psychology*, *25*(4), 436–463. doi: 10.1080/20445911.2013.779248

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(44). doi: 10.1073/pnas.1800755115

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258–1270. doi: 10.1016/j.neuron.2022.01.005

Foster, D. J. (2017). Replay Comes of Age. *Annual Review of Neuroscience*, *40*, 581–602. doi: 10.1146/annurev-neuro-072116-031538

Frensch, P. A., Haider, H., Rünger, D., Neugebauer, U., Voigt, S., & Werg, J. (2003). The route from implicit learning to verbal expression of what has been learned:

Verbal report of incidentally experienced environmental regularity. In L. Jimenez (Ed.), *Attention and implicit learning* (pp. 335–366). John Benjamins Publishing Company.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active Inference , Curiosity and Insight. *Neural Computation*, *29*, 2633–2683. doi: 10.1162/neco

Gao, R., Peterson, E. J., & Voytek, B. (2017). Inferring synaptic excitation/inhibition balance from field potentials. *NeuroImage*, *158*, 70–78. doi: 10.1016/j.neuroimage.2017.06.078

Garrett, D. D., Samanez-larkin, G. R., Macdonald, S. W. S., Lindenberger, U., Mcintosh, A. R., & Grady, C. L. (2013). Moment-to-moment brain signal variability : A next frontier in human brain mapping? *Neuroscience and Biobehavioral Reviews*, *37*(4), 610–624. doi: 10.1016/j.neubiorev.2013.02.015

Gaschler, R., Marewski, J. N., & Frensch, P. A. (2015). Once and for all—How people change strategy to ignore irrelevant information in visual tasks. *Quarterly Journal of Experimental Psychology*, *68*(3), 543–567. doi: 10.1080/17470218.2014.961933

Gaschler, R., Schuck, N. W., Reverberi, C., Frensch, P. A., & Wenke, D. (2019). *Incidental covariation learning leading to strategy change* (Vol. 14) (No. 1). doi: 10.1371/journal.pone.0210597

Gaschler, R., Vaterrodt, B., Frensch, P. A., Eichler, A., & Haider, H. (2013). Spontaneous Usage of Different Shortcuts Based on the Commutativity Principle. *PLoS ONE*, *8*(9), 1–13. doi: 10.1371/journal.pone.0074972

Gick, M. L., & Lockhart, R. S. (1996). Cognitive and Affective Components of Insight. In J. E. Davidson & R. J. Sternberg (Eds.), *The nature of insight* (pp. 197–228). MIT Press. doi: 10.7551/mitpress/4879.001.0001

Groschner, L. N., Malis, J. G., Zuidinga, B., & Borst, A. (2022). A biophysical account of multiplication by a single neuron. *Nature*, *603*(7899), 119–123. doi: 10.1038/s41586-022-04428-3

Grossman, S., Yeagle, E. M., Harel, M., Espinal, E., Harpaz, R., Noy, N., ... Malach, R. (2019). The Noisy Brain: Power of Resting-State Fluctuations Predicts Individual Recognition Performance. *Cell Reports*, *29*(12), 3775–3784. doi: 10.1016/j.celrep.2019.11.081

Haider, H., & Rose, M. (2007). How to investigate insight: A proposal. *Methods*, *42*(1), 49–57. doi: 10.1016/j.ymeth.2006.12.004

Harada, T. (2023). Q-learning model of insight problem solving and the effects of learning traits on creativity. *Frontiers in Psychology*, *14*, 1–10. doi: 10.3389/fpsyg.2023.1287624

Hashmi, A., Nere, A., & Tononi, G. (2013). Sleep-dependent synaptic down-selection (II): Single-neuron level benefits for matching, selectivity, and specificity. *Frontiers in Neurology*, 1–16. doi: 10.3389/fneur.2013.00148

He, B. J. (2013). Spontaneous and task-evoked brain activity negatively interact. *Journal of Neuroscience*, *33*(11), 4672–4682. doi: 10.1523/JNEUROSCI.2922-12.2013

Hedne, M. R., Norman, E., & Metcalfe, J. (2016). Intuitive feelings of warmth and confidence in insight and noninsight problem solving of magic tricks. *Frontiers in Psychology*, *7*, 1–13. doi: 10.3389/fpsyg.2016.01314

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: A unified theory and a connectionist model. *Psychological Review*, *117*(3), 994–1024. doi: 10.1037/a0019532

Hoel, E. (2021). The overfitted brain : Dreams evolved to assist generalization. *Patterns*, *2*(5), 100244. doi: 10.1016/j.patter.2021.100244

Horváth, C. G., Szalárdy, O., Ujma, P. P., Simor, P., Gombos, F., Kovács, I., ... Bódizs, R. (2022). Overnight dynamics in scale-free and oscillatory spectral parameters of NREM sleep EEG. *Scientific Reports*, *12*(1), 1–12. doi: 10.1038/s41598-022-23033-y

Imamoglu, F., Koch, C., & Haynes, J.-D. (2013). MoonBase: Generating a database of two-tone "Mooney" images. *Journal of Vision*, *13*(50).

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of Recurrent Network architectures. *32nd International Conference on Machine Learning, ICML 2015*, *3*, 2332–2340.

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., . . . Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology*, *2*(4), 500–510. doi: 10.1371/journal.pbio.0020097

Karlsson, M. P., Tervo, D. G., & Karpova, A. Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, *338*(6103), 135–139. doi: 10.1126/science.1226518

Kizilirmak, J. M., Galvao Gomes da Silva, J., Imamoglu, F., & Richardson-Klavehn, A. (2016). Generation and the subjective feeling of “aha!” are independently related to learning from insight. *Psychological Research*, *80*(6), 1059–1074. doi: 10.1007/s00426-015-0697-2

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint Relaxation and Chunk Decomposition in Insight Problem Solving. *Journal of Experimental Psychology: Learning Memory and Cognition*, *25*(6), 1534–1555. doi: 10.1037/0278-7393.25.6.1534

Köhler, W. (1925). *The Mentality of Apes*. Kegan Paul, Trench, Trubner & Co. ; Harcourt, Brace & Co.

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, *65*, 71–93. doi: 10.1146/annurev-psych-010213-115154

Kounios, J., & Beeman, M. (2015). *The eureka factor: Aha moments, creative insight, and the brain.* New York: Random House.

Kounios, J., Fleck, J. I., Green, D. L., Payne, L., Stevenson, J. L., Bowden, E. M., & Jung-Beeman, M. (2008). The origins of insight in resting-state brain activity. *Neuropsychologia*, *46*(1), 281–291. doi: 10.1016/j.neuropsychologia.2007.07.013

Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., & Jung-Beeman, M. (2006). Subsequent solution by sudden insight. *Psychologi-*

cal Science, *17*(10), 882–890.

Kralik, J. D., Mao, T., Cheng, Z., & Ray, L. E. (2016). Modeling Incubation and Restructuring for Creative Problem Solving in Robots. *Robotics and Autonomous Systems*, *86*, 162–173.

Krishnamurthy, K., Can, T., & Schwab, D. J. (2022). Theory of Gating in Recurrent Neural Networks. *Physical Review X*, *12*(1), 11011. doi: 10.1103/PhysRevX.12.011011

Lacaux, C., Andrillon, T., Bastoul, C., Idir, Y., Fonteix-galet, A., Arnulf, I., & Oudiette, D. (2021). Sleep onset is a creative sweet spot. *Science Advances*, *5866*(December), 1–10.

Langley, P., & Jones, R. (1986). A Computational Model of Scientific Insight. *Technical Report 87-01*, 313–337.

Lee, H.-K., & Kirkwood, A. (2019). Mechanisms of Homeostatic Synaptic Plasticity in vivo. *PNAS*, *13*, 1–7. doi: 10.3389/fncel.2019.00520

Lendner, J. D., Helfrich, R. F., Mander, B. A., Romundstad, L., Lin, J. J., Walker, M. P., . . . Knight, R. T. (2020). An electrophysiological marker of arousal level in humans. *eLife*, *9*, 1–29. doi: 10.7554/eLife.55092

Lendner, J. D., Niethard, N., Mander, B. A., van Schalkwijk, F. J., Schuh-Hofer, S., Schmidt, H., . . . Helfrich, R. F. (2023). Human REM sleep recalibrates neural activity in support of memory formation. *Science Advances*, *9*(34), 1–16. doi: 10.1126/sciadv.adj1895

Lerner, I., & Gluck, M. A. (2019). Sleep and the extraction of hidden regularities: A systematic review and the importance of temporal rules. *Sleep Medicine Reviews*, *47*, 39–50. doi: 10.1016/j.smrv.2019.05.004

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

Lewis, P. A., Knoblich, G., & Poe, G. (2018). How Memory Replay in Sleep Boosts Creative Problem-Solving. *Trends in Cognitive Sciences*, *22*(6), 491–503. doi:

10.1016/j.tics.2018.03.009

Liu, S., Papailiopoulos, D., & Achlioptas, D. (2020). Bad global minima exist and SGD can reach them. *Advances in Neural Information Processing Systems*, *2020-Decem*(NeurIPS).

Löwe, A., Petzka, M., Tzegka, M., & Schuck, N. (2024). N2 Sleep Inspires Insight. *bioRxiv*, 1–25.

Löwe, A. T., Touzo, L., Muhle-Karbe, P. S., Saxe, A. M., Summerfield, C., & Schuck, N. W. (2024). Abrupt and spontaneous strategy switches emerge in simple regularised neural networks. *PLoS Computational Biology (In Press)*, 1–22. doi: 10.32470/ccn.2023.1026-0

Ludmer, R., Dudai, Y., & Rubin, N. (2011). Uncovering Camouflage: Amygdala Activation Predicts Long-Term Memory of Induced Perceptual Insight. *Neuron*, *69*(5), 1002–1014. doi: 10.1016/j.neuron.2011.02.013.Uncovering

Maier, N. R. (1930). Reasoning in humans. I. On direction. *Journal of Comparative Psychology*, *10*(2), 115–143. doi: 10.1037/h0073232

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. doi: 10.1038/nrn1076

Mednick, S. (1968). The Remote Associates Test. *The Journal of Creative Behavior*, *2*(3), 213–214. doi: 10.1002/j.2162-6057.1968.tb00104.x

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(5725), 1–12. doi: 10.1038/s41467-020-19632-w

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, *15*(3), 238–246. doi: 10.3758/BF03197722

Miller, P., & Katz, D. B. (2010). Stochastic transitions between neural states in taste processing and decision-making. *Journal of Neuroscience*, *30*(7), 2559–2570. doi: 10.1523/JNEUROSCI.3047-09.2010

Miskovic, V., MacDonald, K. J., Rhodes, L. J., & Cote, K. A. (2019). Changes in

EEG multiscale entropy and power-law frequency scaling during the human sleep cycle. *Human Brain Mapping*, *40*(2), 538–551. doi: 10.1002/hbm.24393

Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2009). Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. *Neuron*, *63*(6), 879–888. doi: 10.1016/j.neuron.2009.09.013

Mitchell, S. J., & Silver, R. A. (2003). Shunting inhibition modulates neuronal gain during synaptic excitation. *Neuron*, *38*(3), 433–445. doi: 10.1016/S0896-6273(03)00200-9

Nere, A., Hashmi, A., Cirelli, C., & Tononi, G. (2013). Sleep-dependent synaptic down-selection (I): Modeling the benefits of sleep on memory consolidation and integration. *Frontiers in Neurology*(September), 1–17. doi: 10.3389/fneur.2013.00143

Nissen, M. J., & Bullemer, P. (1987). Attention Requirements of Learning Evidence from Performance Measures. *Cognitive Psychology*, *19*, 1–32.

Norimoto, H., Makino, K., Gao, M., Shikano, Y., Okamoto, K., Ishikawa, T., ... Ikegaya, Y. (2018). Hippocampal ripples down-regulate synapses. *Science*, *359*(6383), 1524–1527. doi: 10.1126/science.aao0702

Oh, Y., Chesebrough, C., Erickson, B., Zhang, F., & Kounios, J. (2020). An insight-related neural reward signal. *NeuroImage*, *214*, 116757. doi: 10.1016/j.neuroimage.2020.116757

Ohlsson, S. (1984). Restructuring revisited: II. An information processing theory of restructuring and insight. *Scandinavian Journal of Psychology*, *25*(2), 117–129. doi: 10.1111/j.1467-9450.1984.tb01005.x

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In *Advances in the psychology of thinking.* Harvester Wheatseaf.

Ohlsson, S. (2011). *Deep Learning*. Cambridge: Cambridge University Press.

Olcese, U., Esser, S. K., & Tononi, G. (2010). Sleep and synaptic renormalization: A computational study. *Journal of Neurophysiology*, *104*(6), 3476–3493. doi: 10.1152/jn.00593.2010

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *156869*. doi: https://doi.org/10.1155/2011/156869

Ormerod, T. C., MacGregor, J. N., & Chronicle, E. P. (2002). Dynamics and Constraints in Insight Problem Solving. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(4), 791–799. doi: 10.1037/0278-7393.28.4.791

Ovington, L. A., Saliba, A. J., & Goldring, J. (2016). Dispositional Insight Scale: Development and Validation of a Tool That Measures Propensity Toward Insight In Problem Solving. *Creativity Research Journal*, *28*(3), 342–347. doi: 10.1080/10400419.2016.1195641

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, *102*, 127–144. doi: 10.1016/j.cogpsych.2017.11.006

Perkins, D. (2000). *The Eureka effect. The art and logic of breakthrough thinking.* New York, NY: Norton.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*, *4*(1), 1–12. doi: 10.1038/s41746-021-00440-5

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, *317*(6035), 314–319. doi: 10.1038/317314a0

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. *arXiv*, 1–10.

Rabinowitz, N. C., Goris, R. L., Cohen, M., & Simoncelli, E. P. (2015). Attention stabilizes the shared gain of V4 populations. *eLife*, *4*, 1–24. doi: 10.7554/eLife.08998

Rajananda, S., Lau, H., & Odegaard, B. (2018). A random-dot kinematogram for web-based vision research. *Journal of Open Research Software*, *6*(1). doi: 10.5334/jors.194

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews*, *93*(2), 681–766. doi: 10.1152/physrev.00032.2012

Ritz, H., Leng, X., & Shenhav, A. (2022). Cognitive control as a multivariate optimization problem. *Journal of Cognitive Neuroscience*, *34*(4), 569–591.

Rolls, E. T., & Deco, G. (2012). *The Noisy Brain: Stochastic dynamics as a principle of brain function*. Oxford University Press. doi: 10.1093/acprof:oso/9780199587865.001.0001

Rolls, E. T., Tromans, J. M., & Stringer, S. M. (2008). Spatial scene representations formed by self-organizing learning in a hippocampal extension of the ventral visual system. *European Journal of Neuroscience*, *28*(10), 2116–2127. doi: 10.1111/j.1460-9568.2008.06486.x

Sandkühler, S., & Bhattacharya, J. (2008). Deconstructing Insight: EEG Correlates of Insightful Problem Solving. *PloS one*, *3*(1). doi: 10.1371/Citation

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., & Cox, D. D. (2019). On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, *2019*(12). doi: 10.1088/1742-5468/ab3985

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the international conference on learning representations 2014.* (pp. 1–22).

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *166*(23), 11537–11546. doi: 10.1073/pnas.1820226116

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, *110*(3), 395–411. doi: 10.1016/j.cognition.2008.11.017

Schneider, B., Szalárdy, O., Ujma, P. P., Simor, P., Gombos, F., Kovács, I., . . . Bódizs, R. (2022). Scale-free and oscillatory spectral measures of sleep stages in humans. *Frontiers in Neuroinformatics*, *16*. doi: 10.3389/fninf.2022.989262

Schönauer, M., Brodt, S., Pöhlchen, D., Breßmer, A., Danek, A. H., & Gais, S.

(2018). Sleep does not promote solving classical insight problems and magic tricks. *Frontiers in Human Neuroscience*, *12*(February), 1–11. doi: 10.3389/fnhum.2018.00072

Schuck, N. W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P. A., Haynes, J. D., & Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, *86*(1), 331–340. doi: 10.1016/j.neuron.2015.03.015

Schuck, N. W., Li, A. X., Wenke, D., Ay-Bryson, D. S., Loewe, A. T., Gaschler, R., & Shing, Y. L. (2022). Spontaneous discovery of novel task solutions in children. *Plos One*, *17*(5), e0266253. doi: 10.1371/journal.pone.0266253

Shen, W., Tong, Y., Li, F., Yuan, Y., Hommel, B., Liu, C., & Luo, J. (2018). Tracking the neurodynamics of insight: A meta-analysis of neuroimaging studies. *Biological Psychology*, *138*, 189–198. doi: 10.1016/j.biopsycho.2018.08.018

Shen, W., Yuan, Y., Liu, C., & Luo, J. (2016). In search of the 'Aha!' experience: Elucidating the emotionality of insight problem-solving. *British Journal of Psychology*, *107*(2), 281–298. doi: 10.1111/bjop.12142

Sio, U. N., & Ormerod, T. C. (2009). Does Incubation Enhance Problem Solving? A Meta-Analytic Review. *Psychological Bulletin*, *135*(1), 94–120. doi: 10.1037/a0014212

Smith, R. W., & Kounios, J. (1996). Sudden insight: All-or-none processing revealed by speed-accuracy decomposition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*(6), 1443–1462. doi: 10.1037/0278-7393.22.6.1443

Sprugnoli, G., Rossi, S., Emmendorfer, A., Rossi, A., Liew, S. L., Tatti, E., . . . Santarnecchi, E. (2017). Neural correlates of Eureka moment. *Intelligence*, *62*, 99–118. doi: 10.1016/j.intell.2017.03.004

Stuyck, H., Aben, B., Cleeremans, A., & Van den Bussche, E. (2021). The Aha! moment: Is insight a different form of problem solving? *Consciousness and Cognition*, *90*, 103055. doi: 10.1016/j.concog.2020.103055

Suppermpool, A., Lyons, D. G., Broom, E., & Rihel, J. (2024). Sleep pressure modulates single-neuron synapse number in zebrafish. *Nature*, *629*(8012), 639–645.

doi: 10.1038/s41586-024-07367-3

Tan, T., Zou, H., Chen, C., & Luo, J. (2015). Mind Wandering and the Incubation Effect in Insight Problem Solving. *Creativity Research Journal*, *27*(4), 375–382. doi: 10.1080/10400419.2015.1088290

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. Macmillan Press.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, *2*(94).

Tik, M., Sladky, R., Luft, C. D. B., Willinger, D., Hoffmann, A., Banissy, M. J., ... Windischberger, C. (2018). Ultra-high-field fMRI insights on insight: Neural correlates of the Aha!-moment. *Human Brain Mapping*, *39*(8), 3241–3252. doi: 10.1002/hbm.24073

Tononi, G., & Cirelli, C. (2003). Sleep and synaptic homeostasis: A hypothesis. *Brain Research Bulletin*, *62*(2), 143–150. doi: 10.1016/j.brainresbull.2003.09.004

Tononi, G., & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Medicine Reviews*, *10*(1), 49–62. doi: 10.1016/j.smrv.2005.05.002

Tononi, G., & Cirelli, C. (2014). Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration. *Neuron*, *81*(1), 12–34. doi: 10.1016/j.neuron.2013.12.025

Trans. Grube, G. M. A. (1997). *Plato. The Republic.* Hackett Publishing Company.

Tulver, K., Kaup, K. K., Laukkonen, R., & Aru, J. (2023). Restructuring insight: An integrative review of insight in problem-solving, meditation, psychotherapy, delusions and psychedelics. *Consciousness and Cognition*, *110*, 103494. doi: 10.1016/j.concog.2023.103494

Turrigiano, G. G., & Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, *5*(2), 97–107. doi: 10.1038/nrn1327

Verleger, R., Rose, M., Wagner, U., Yordanova, J., & Kolev, V. (2013). Insights into sleep's role for insight: Studies with the number reduction task. *Advances in Cognitive Psychology*, *9*(4), 160–172. doi: 10.5709/acp-0143-8

Vickrey, B., & Lerner, I. (2023). Overnight exposure to pink noise could jeopardize sleep-dependent insight and pattern detection. *Frontiers in Human Neuroscience*, *17*, 1–9. doi: 10.3389/fnhum.2023.1302836

Vitruvius. (1567). *De Architetura libri decem*. Venice: Daniele Barbaro.

Voytek, B., Kramer, M. A., Case, J., Lepage, K. Q., Tempesta, Z. R., Knight, R. T., & Gazzaley, A. (2015). Age-Related Changes in 1/f neural ElectrophysiologicalNoise.pdf. *Journal of Neuroscience*, *35*(38), 13257–13265. doi: 10.1523/JNEUROSCI.2332-14.2015

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature*, *427*(6972), 352–355. doi: 10.1038/nature02223

Wallas, G. (1926). *The art of thought*. London: J.Cape.

Waschke, L., Kloosterman, N. A., Obleser, J., & Garrett, D. D. (2021). Behavior needs neural variability. *Neuron*, *109*(5), 751–766. doi: 10.1016/j.neuron.2021.01.023

Wehrheim, M. H., Faskowitz, J., Schubert, A. L., & Fiebach, C. J. (2024). Reliability of variability and complexity measures for task and task-free BOLD fMRI. *Human Brain Mapping*, *45*(10), 1–18. doi: 10.1002/hbm.26778

Weisberg, R. W. (2015). Toward an integrated theory of insight in problem solving. *Thinking and Reasoning*, *21*(1), 5–39. doi: 10.1080/13546783.2014.886625

Wertheimer, M. (1925). *Drei Abhandlungen zur Gestalttheorie*. Erlangen: Verlag der Philosophischen Akademie.

Wertheimer, M. (1959). *Productive thinking* (2nd ed. ed.). New York, NY: Harper.

Woltz, D., Bell, B., Kyllonen, P., & Gardner, M. (1996). Memory for order of operations in the acquisition and transfer of sequential cognitive skills. *Journal of Experimental Psychology: Learning, Menory and Cognition*, *22*(2), 438–457.

Yordanova, J., Kolev, V., Wagner, U., Born, J., & Verleger, R. (2012). Increased Alpha (8-12 Hz) activity during slow wave sleep as a marker for the transition from implicit knowledge to explicit insight. *Journal of Cognitive Neuroscience*, *24*(1), 119–132. doi: 10.1162/jocn{_}a{_}00097

Yu, Y., Oh, Y., Kounios, J., & Beeman, M. (2024). Electroencephalography Spectral-power Volatility Predicts Problem-solving Outcomes. *Journal of Cognitive Neuroscience*, *36*(5), 901–915. doi: 10.1162/jocn{_}a{_}02136

6 Acknowledgements

I am incredibly lucky to have not only had the opportunity to write this thesis, but also receive the support of truly wonderful people surrounding me. First and foremost, I want to thank Nico for hiring me as the Schuck lab's very first Research Assistant and getting me so hooked on my work on insight that I not only stayed for my Master's thesis, but also my PhD. Thank you for leaving me freedom to explore my own interests while gently guiding me to work on my weaknesses. Thank you for teaching me academic writing, analysis and to go lower when R2 goes low (just kidding). Thank you for the lasertag fun, the oatbrews and all the really hilarious jokes. A funnier supervisor is hard to come by.

I am incredibly grateful to Andrew and the entire Saxe Lab for taking me into their London HQ when my supervisor was busy changing diapers. I owe a huge deal of what I understand of neural networks and their analysis to the months spent in the more theoretical Neuroscience realm located in the country of pints and scones.

Thank you also to Chris and the Summerfield lab for support and exchanges over the years that never ceased to light my spark for science again when I needed it the most.

I would have made it nowhere near the end of this PhD without Marit, Shany and Ondrej. It is hard to explain to anyone outside of our group chat what place these conversations take in my life. Besides you three being the funniest people alive, you are more than therapy, a place where I am understood, supported and uplifted. A very special thank you goes to Marit in particular who is the dreamiest collaborator I could have ever imagined. Our date nights and office days will forever hold a special place in my heart.

I also want to thank Noa for the most fun Cosyne there ever was, Fabi for all the cakes, Elsa for everything and all the rest of the Schuck Lab for some incredible years that made me grow so much as a person and brought me endless wisdom, joy and life lessons.

This thesis would not exist without music. The list would extend for too long, but I'm deeply grateful to all the artists of the last decades and today who provided the music that fuelled me to code, read, write and *think* myself into different realms.

Thank you to my Dad who kept me from studying Pharmacology. Thank you for having high expectations that were not always easy to handle, but brought me to where I am.

I would be nowhere in life without my Mom. You never questioned any of my choices, believed in me every single day of my life and always made me feel home. Your support, trust and endless selfless love is the most precious gift I got in this life. We've been having conversations every day for the past 29 years and yet we never have enough time. Thank you for always showing me the way, with love and grace.

Thank you to my therapist Mrs. H. for leading me out of the darkness, towards life and light.

Thank you to Ernst for first being my date, then my business partner and always my best friend. Thank you for convincing me to start a record label together half way through my PhD, thank you for believing in me when I don't, thank you for bringing all the chaos and spice into my life while keeping me safe.

Thank you Mathilde for pushing me to strive for something more meaningful. Thank you for sharing your spirituality, wisdom and care.

Thank you to Angus for pushing my physical limits when I am reaching my mental limits. Thank you for all the music, butter and moon rises, thank you for being the literal strongest shoulder to lean on.

Thank you Lari for the last summers of our youth I'll never forget and growing into adulthood together.

Thank you Thuy for being by my side through it all, since long before I started being interested in brains.

Thank you Daniel for being the only person in my life to comprehend every angle of my life, thank you also for your unmatched wit and humour.

Thank you Caroline for being my room mate in Woods Hole and turning out to be

my long lost twin at heart.

Thank you to all my other friends – you know who you are, I appreciate you endlessly.

Thank you Basile for this perfect puzzle, for completing me, grounding me, making me feel less alien on this planet and for showing me love I never knew possible. Your quest for answers will forever inspire me.

I dedicate 10% of this thesis to my grandpa, the first Dr. Löwe, who would have been so proud to see a third generation carry on this name, and 90% to my Mom, the first Dr. A. Löwe, single-raising, kindest, toughest Dr. Mom I could have wished for in this life – my biggest supporter from day one whose primary goal was to teach me that women could and *should* do anything.