Attentional, hedonic and interoceptive correlates of implicit processes in addiction: a learning perspective.

MATEO LEGANES FONTENEAU

Thesis submitted for the degree of Doctor of Philosophy

UNIVERSITY OF SUSSEX

September 2018

Thanks to:

My parents for their unconditional support through all my education.

Zofia for always being there for me.

Dora who shared with me her mentorship and knowledge in this incredible journey and turned it into one of the best experiences of my life.

Ryan for his guidance and support whenever I needed him.

My office mates at 4B9, particularly to Becks for brightening up my days.

The Avocados Gone Wild crew (Becks, Dr Ntontis, James, Meike, Mikey, Tasmin, Callum, Heather and Abi). So much avocado and fun was had over three years!

Our lab group, in particular to Kiki for her dedication and advice.

Sarah and Zoltan for sharing with me their invaluable knowledge and opening up my perspectives.

Pilar Tejero, Pedro Valero and Enrique Berjano. Without them I would I have never taken this amazing path.

My school and high-school teachers, in particular Jose Luis Gonzalez “el Chino”.

Summary

UNIVERSITY OF SUSSEX

Mateo Leganes Fonteneau

PhD in Psychology

Title: Attentional, hedonic and interoceptive correlates of implicit processes in addiction: a learning perspective.

Addiction is characterised by maladaptive drug-approach behaviours, some of which may take place without conscious cognitive control. Through repeated associations with a substance, drug related stimuli acquire incentive salience properties via Pavlovian reward learning, triggering these responses.

In order to better understand implicit processes in addiction it is crucial to increase our knowledge about unconscious reward mechanisms. Previous literature has failed to thoroughly demonstrate the ability of Pavlovian conditioned stimuli to generate responses in the absence of stimulus-outcome contingency awareness. Therefore, an effort was put to develop novel techniques measuring different aspects of conditioning.

Using an Emotional Attentional Blink, we proved stimuli associated with high probabilities of reward (HR) generated increased attentional responses in participants Unaware of contingencies. Integrating Conditioned Stimuli (CS) as task-irrelevant distractors in a Flanker task we found HR stimuli interfered with cognitive control,

again implicitly. A novel methodology, based on Bayesian analyses, allowed us determining the unconscious nature of learning, strengthening our findings. Conversely, subjective hedonic responses were not modulated by implicit learning, highlighting the inadequacy of such measures for the study of implicit conditioning.

In order to further understand individual differences in the development of conditioned responses, we examined the role of interoception, the mental representation of internal bodily sensations, in this matter. It was shown that interoceptive awareness modulates the development of reward prediction and hedonic responses in Pavlovian conditioning. We also examined the effect of a natural reward, alcohol, on interoceptive awareness, and found that under acute alcohol administration interoceptive awareness facilitates the perception of subjective substance effects.

These findings have important implications for our understanding of basic addictive processes. The ability of implicit CS to generate responses supports the existence of drug-approach behaviours devoid of conscious awareness. The role of interoception in Pavlovian conditioning provides the basis for its integration in classic learning theories.

Statement

All work presented in this thesis is original research. The chapters included reflect the collaboration with different authors as well as my supervisors and are formatted according to the latest submission for publication.

Chapter 2 was published in a peer-reviewed journal (*Behavioural Brain Research)* and is included in the thesis as it was published[1]. Prof. Theodora Duka and Dr. Ryan Scott participated in the design of the experimental paradigm, data analysis and theoretical conceptualisation, and provided corrections to the manuscript. Prof. Zoltan Dienes also collaborated with the data analysis. Experiment and material design, data collection, analyses and write up were carried out by Mateo Leganes Fonteneau.

Chapter 3 has been accepted for publication in the peer-reviewed journal *Learning and Memory*. Prof. Theodora Duka, Dr. Ryan Scott and Dr. Kyriaki Nikolaou participated in the experimental design as well as the theoretical conceptualisation, and provided corrections to the manuscript. Dr. Kyriaki Nikolaou provided the experimental paradigms as well as support with data analysis. Experiments and material design, data collection, analyses and write up were carried out by Mateo Leganes Fonteneau.

---

[1] Reference for Chapter 2:

Leganes-Fonteneau, M., Scott, R., & Duka, T. (2018). Attentional responses to stimuli associated with a reward can occur in the absence of knowledge of their predictive values. *Behavioural brain research*, *341*, 26-36.

Chapter 4 was rejected by reviewers after submission to *Cognition and Emotion.* The corrections suggested by reviewers were addressed in the manuscript before the submission of the thesis. Prof. Theodora Duka, Dr. Ryan Scott and Dr. Sarah Garfinkel participated in the design of the experimental paradigm, data analysis and theoretical conceptualisation, and provided corrections to the manuscript. Experiment and material design, data collection, analyses and write up were carried out by Mateo Leganes Fonteneau.

Chapter 5 is currently under review (paper modified after the first round of comments by reviewers) in the journal *Biochemistry Pharmacology and Behaviour* for a special issue on interoception. Prof. Theodora Duka and Dr. Sarah Garfinkel participated in the design of the experimental paradigm, data analysis, and theoretical conceptualisation and provided corrections to the manuscript. Data collection was carried out by Yan Lam and Yun Cheang. Mateo Leganes Fonteneau supervised laboratory work, prepared alcohol solutions, analysed the data and wrote the manuscript.

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Abbreviations in order of appearance:

HR – High Reward (stimulus)

LR – Low Reward (stimulus

CS – Conditioned Stimulus (stimuli)

US – Unconditioned Stimulus (stimuli)

PC – Pavlovian Conditioning

R – Response

S – Stimulus

CA – Contingency Awareness

IAT – Implicit Association Test

EC – Evaluative Conditioning

CResp – Conditioned Response

SDT – Signal Detection Theory

ROC – Receiver Operating Characteristic

RSVP – Rapid Serial Visual Presentation

EAB – Emotional Attentional Blink

IAPS – International Affective Picture System

AUDIT – Alcohol Use Disorders Identification Test

AUQ – Alcohol Use Questionnaire

BAS / BIS – Behavioural Avoidance/Inhibition System

BIS-11 – Barratt Impulsiveness Scale

PANAS – Positive And Negative Affect Scale

BPQ – Bodily Perception Questionnaire

EA – Expectancy Awareness

CR – Correct Rejection

H – Hit

FA – False Alarm

M - Miss

OR – Odds Ratio

B – Bayes Factor

RT – Reaction Times

LMT – Low Meta Tracker

HMT – High Meta Tracker

LMD – Low Meta Discriminator

HMD – High Meta Discriminator

BMI – Body Mass Index

TAS-20 – Toronto Alexithymia Scale

VAS – Visual Analogue Scale

BAC – Blood Alcohol Content

1 Theoretical overview

1.1 Introduction

Substance use disorders and other addictions are defined as 'a chronic disorder characterised by a compulsion to seek and take a drug, loss of control in limiting the intake and emergence of a negative emotional state […] reflecting a motivational withdrawal syndrome when access to the drug is prevented' (Koob & Volkow, 2010). This pathological process entails aberrant cognitive responses generated by dysregulated reward learning mechanisms (Koob & Volkow, 2016), generating long term neural adaptations and cognitive impairments (Bates, Pawlak, Tonigan, & Buckman, 2006; Goldstein et al., 2009), particularly in executive control (Verdejo-García, Bechara, Recknor, & Pérez-García, 2006).

Although the existing cognitive behavioural therapies for addiction provide increasing success rates (Hendershot, Witkiewitz, George, & Marlatt, 2011), approaches targeting maladaptive cognitive processes could provide novel therapeutic solutions (Cristea, Kok, & Cuijpers, 2016; Wiers, Gladwin, Hofmann, Salemink, & Ridderinkhof, 2013).

Crucial to the understanding of addiction is the clarification of the role of implicit responses (Bechara, 2005; Belin, Belin-Rauscent, Murray, & Everitt, 2013) interacting with executive functioning as part of the generation of drug-approach behaviours (Stacy & Wiers, 2010).

Given the relevance of reward processing, and particularly of Pavlovian Conditioning (PC), in addictive behaviours (Berridge, 2000), the purpose of this thesis is to further

examine the implicit components of addiction through the study of appetitive learning in the absence of conscious awareness.

1.2 Reward learning theories

1.2.1 A historical account of Reinforcement learning

Early theories of reward learning attribute the formation of long lasting behaviours to the direct association between a stimulus and a response (S-R) (Thorndike, 1898; Watson, 1913). According to this view, actions followed by positive outcomes have more chances to be repeated in the future, whereas those followed by negative outcomes are less likely to recur. For example, an animal performing an action (i.e. pressing a lever) and subsequently obtaining a reward (i.e. food) would be more likely to press the lever again due to the positive characteristics of the reward.

Stimulus-response and associative learning theories simplified however the characterisation of the mechanisms by which learning occurs, disregarding the role of hedonic responses and expectancies and assuming that the nature of a reward is irrelevant to S-R habituation (Berridge, 2000). For example, it was not specified how individuals can engage in specific responses (eating instead of drinking) depending on the reward (i.e. food instead of a sweet solution) if the nature of the reward itself is irrelevant to the generation of learnt responses. Moreover, the ability of rewards pertaining to different categories (i.e. sex, drugs or food) to equally act as reinforcers was not explained either, and neither was the generation of complex behavioural sequences. Finally, the simple association of a stimulus with an action could not explain the adaptation of responses to different situations or the generation of new

behaviours (Bindra, 1978). For example, if someone feels cold, they may put an extra layer of clothes on, but also close the window or turn the heater on. This is at odds with S-R theories of learning indicating that learning occurs as the consolidation or fading of specific sensory-motor connections (Bindra, 1978). This would imply that there is a common denominator to rewards that can drive and direct approach behaviours.

These theories were followed by simplified conceptualisations of associative learning, reducing them to mere observational accounts (Skinner, 2011). Skinner devised a series of valuable procedures for the study of reward learning, however, his refusal to further investigate the mechanisms by which learning occurs (perfectly illustrated by the title of one of his papers: *Are theories of learning necessary?*, 1950) gave rise to more complex interpretations of learning and behaviour, such as Hullian theories, attempting to disentangle the mechanisms underlying response learning.

Drive reduction theories of learning (Hull, 1943) attempted to explain the stimulus specificity on influencing behaviours and the source of reward as a mere reduction of drives. When a subject is in a state of hunger, that drives behaviours towards reducing it, for example eating. The obtainment of food would decrease the drive, and it is that drive reduction that ultimately generates a reinforcement.

1.2.2 Incentive theories of learning: Bolles – Bindra – Toates

A series of experimental incongruences, in which responses not directly reinforced produced behaviours, constituted the basis for criticisms directed to response reinforcement theories of learning. For example, the experimental observation of

stereotypies (Breland & Breland, 1961), such as a raccoon washing a token associated with food in a similar manner as it would with a fruit, meant that stimuli unable to directly decrease drives could still generate responses. This contributed to the rise of incentive theories of learning, by which Conditioned Stimuli (CS) acquire the motivational properties of rewards (US), dragging attention and orienting or generating goal directed behaviours (Berridge, 2000).

On that same line, Bolles's (1972) view on learning mechanisms was that reinforcement is not established through S-R associations, but by stimulus-stimulus and response-stimulus expectancies.

Bolles argued that animals learn to predict the occurrence of an US (i.e. food pellets) after a CS (i.e. a light) due to the concomitant presentation of both. This generates in animals the expectancy of a stimulus, be it pleasurable or aversive, in the presence of another one. A secondary mechanism extended these associations to the existence of R-S contingencies. Animals learn that certain responses in the environment generate an outcome, in that case driven by the expectancy of a stimulation after their own response (i.e. pressing a lever to obtain food pellets).

According to Bolles, it is the strength of the CS-US or R-S association that drives the generation of increased expectancies, augmenting response probabilities. The value of US should as well modulate responses, in part as a function of the internal homeostatic state of the subject, reflecting on Hull's drive theories (Hull, 1943).

Bindra's (1978) criticism to response-reinforcement theories (Thorndike, 1911) crystallised in an incentive-motivational model of learning, integrating within early

drive theories based on homeostatic processes the role of incentive stimuli. However, in opposition to Bolles's idea, he claimed that learning is not based solely on expectancies, but rather that a motivational transfer between US and CS drives the generation of learning.

Within this model, the central motive state (Bindra, 1968) constitutes the main entity under which goal-directed actions are processed. When an organismic state, such as hunger, coincides with the presence of an incentive stimulus (i.e. food), this generates an appetitive central motive state associated with the reward, resulting in approach or consummatory behaviours. The central motive state does not only affect behavioural outcomes, but also viscero-somatic reactions, producing preparatory responses such as salivation. It is also important to note the posited role of viscero-afferent responses (termed "sensory inflow" by Bindra) in the perception of organismic states within this model. These signals would comprise both exteroceptive and interoceptive information.

A key aspect in this model is that CS acquire the motivational properties of rewards, eliciting goal-directed behaviours. Through concomitant CS-US presentations, positive contingencies are established, and the occurrence of CS can then excite the central representation of US.

When CS are presented under extinction, that is, once contingencies are established but in the absence of reward, CS will also generate a central motive state leading to approach behaviours compatible with its own nature, but also to physiological reactions congruent with those of the US originally matched with the CS.

Viscerosomatic and instrumental responses in PC appear to be independent from one another, particularly when considering CS and US separately (Bindra, 1974). For Bindra, responses necessary to approach a stimulus (e.g. walking towards it) and preparatory viscerosomatic reactions (e.g. salivation) might be the same for CS and US, whereas consummatory responses (e.g. biting or chewing) would not appear in relation to the CS (obviating however different instances of sign-tracking in which animals attempt for example to bite a light signalling a reward). This indicates that CS are not strictly linked to the responses associated with a US, but rather that they can act as a source of motivational arousal leading to behaviours adapted to the nature of the CS (in the case of humans for example, attentional orientation).

Drive states regained relevance due to the work of Cabanac (1971, 1979) on alliesthesia, linking the perception of the hedonic value of an external stimulus to the internal state of a subject. According to this concept, an equally tasty food would elicit different subjective sensations or experiences depending on the internal milieu of a subject. For example, the consumption of sugar would elicit more pleasurable responses if an animal is hungry. However, as satiety is achieved, hedonic responses would become neutral, and finally negative (Berridge, 1991).

A crucial distinction between drive reduction theories (Hull, 1943) and alliesthesia however is that rewards are not posited to produce reinforcement because they satiate or fulfil a drive, but rather that the consumption of a reward under the right physiological conditions increases its perceived attractiveness, impacting thereon behavioural responses. Toates (1986) based on this perspective his theory of learning,

reconciling the expectancy accounts of Bolles and the role of hedonic responses of Bindra.

According to Toates, both the hedonic value of US and the expectancies generated by associations can modulate behaviour. By introducing the alliesthetic component of reward appraisal, the value of a reward is not anymore fixed, and changes of alliesthetic parameters can determine the generation and retrieval of learnt responses.

For example, the presentation of rewards in small quantities can awaken physiological drives that can then cascade to increase the motivational salience of that same reward (Cornell, Rodin, & Weingarten, 1989). Such an effect, known as priming, has been studied in rats (i.e. Lê et al., 1998) and humans (i.e. Duka & Townshend, 2004) and can generate or reinstate responses towards alcohol cues.

Furthermore, as CS acquire incentive properties associated with rewards, their mere presentation can elicit priming mechanisms similar to those exerted by a small dose of alcohol or food (Weingarten, 1983). CS (visual, olfactory or auditory) can therefore trigger alliesthetic responses generating physiological states compatible with reward consumption, increasing the hedonic value of stimuli. Finally, the effect of CS on behavioural outcomes is also dependent on alliesthetic mechanisms, such that the incentive value of a stimulus paired with an appetitive reward will decrease under conditions of satiety.

1.2.3 Cognitive expectancy theories

According to Dickinson's view (1989), responses to stimuli take place due to the expectancy of a reward to occur in its presence. Therefore, knowledge about the

causal relationship between R-S contingencies drives and directs behaviours, a concept known as outcome-representation. A subject would thus perform an action with the conscious expectancy of obtaining a certain reward, and this outcome awareness is not driven solely by contingency knowledge, but also by the representation of the hedonic value of a reward. Although initial research in this sense studied the role of outcome expectancies in animal models (Dickinson & Balleine, 1995), the role of expectancies in human behaviour (Shanks, 2007; Shanks & Dickinson, 1988) will be discussed later on.

1.3 Theories of addiction

1.3.1 Positive and negative reinforcement

Reflecting on classifications similar to those of early theories of learning, addictive processes, and more specifically drug urges, were originally based either on positive or negative reinforcement approaches.

Withdrawal theories of addiction explain drug cravings and urges as a drive to reduce the effects of withdrawal states (Jellinek, 1955). A subject would therefore engage in drug consumption to alleviate the negative state generated by the absence of a substance. However, drug cravings can appear in subjects long after withdrawal effects have faded out (e.g. Mathew, Claghorn, & Largen, 1979), compromising such theory. That was accounted for by Wikler (1948), explaining that contexts or situations conditioned with withdrawal symptoms can elicit drug cravings and urges long after the fading of withdrawal. Another view (Siegel, 1975) held that the development of drug cravings occurred through conditioned compensatory mechanisms. Stimuli related with substance administration generate preparatory responses opposite to

those of the substance (i.e. excitability when facing stimuli related with an opioid, sleepiness when observing a stimulus related with caffeine) explaining the generation of substance tolerance. Those same compensatory responses could trigger as well withdrawal-like effects, generating craving.

Positive reinforcement theories on the other hand base their explanation of drug urges on the hedonic properties of stimulants (Stewart, de Wit, & Eikelboom, 1984; Wise, 1988). By representing the anticipation of a positive outcome, drug related stimuli elicit motivational states compatible with craving, ultimately driving relapses or consumption. This positive reinforcement mechanism is not at odds with negative reinforcement theories of withdrawal, and both phenomena could interplay in order to drive consumption (Baker, Morse, & Sherman, 1986; Roy A Wise & Koob, 2014).

1.3.2 Automatic and habit theories of addiction

The occurrence of cravings or urges does not always correlate with behavioural outcomes and drug consumption behaviours or with physiological responses to drug related stimuli (Tiffany, 1990). Tiffany described the necessity to explore separately non-automatic accounts of drug behaviour (such as urge responding) from automatic reactions to drug related stimuli (i.e. physiological reactivity). Those automatic mechanisms should reflect fast, autonomous, effortless, uncontrollable and unconscious processes.

Through years of drug consumption, addictive responses become engrained in learned schemata, reflecting automated patterns. Internal (i.e. emotional, physiological or

withdrawal states) or external drug related cues (i.e. a place, smell or visual stimulus) are thought to trigger or initiate “drug-use actions plans”.

Another take on the automaticity of drug-use disorders is that of habitual responses (Everitt & Robbins, 2005). The distinction between stimulus driven and goal directed responses based on expectancies was illustrated by a series of devaluation experiments in rats (Miles, Everitt, & Dickinson, 2003) in which instrumental responses were paired with solutions containing either cocaine or sucrose. After outcome devaluation, they found that food paired responses were extinguished, whereas cocaine related responses resisted outcome devaluation, thus indicating that responses towards substances reflect habitual pathways rather than outcome-contingency expectancies.

In humans however, evidence for drug approach behaviours being enacted through goal-directed mechanisms (Brown, Duka, & Forster, 2018) and the finding that drug cues can generate explicit craving responses (Tiffany, Warthen, & Goedeker, 2009) has impeded an accurate translation of findings obtained in animals (Hogarth, Dickinson, & Duka, 2010). On a devaluation experiment equivalent to that of Miles et al. (2003), Hogarth and Chase (2011) paired button presses with either chocolate or tobacco outcomes. After devaluating the outcomes via satiety or aversive related messages, they found a decrease in responses associated with the devalued outcomes, thus implying that traditional mechanisms of S-R associations based on habitual responses may not be able to explain drug-related behaviours in humans. The finding that stimulus-outcome expectancies can drive responses towards drug-related stimuli,

overcoming the effect of habitual responses, favoured a shift towards the relevance of outcome expectancies and their conscious appraisal.

1.3.3 Expectancy theories of addiction

Expectancy theories of addiction posit that the knowledge of outcomes following stimulation, such as the prediction of stimulant effects after cocaine consumption, are the main drive of drug-approach behaviours (Bolles & Fanselow, 1980; Brandon, Herzog, Irvin, & Gwaltney, 2004). From a appetitive conditioning point of view (Hogarth, Dickinson, Wright, Kouvaraki, & Duka, 2007; Hogarth & Duka, 2006), explicit knowledge about stimulus-outcome contingencies seems necessary, for example, for nicotine paired stimuli to affect instrumental, attentional and hedonic responses. In a series of experiments, Hogarth et al. (2007) paired a geometrical stimulus with probabilities of earning tobacco and another stimulus with probabilities of losing it. On a second phase, they paired a behavioural response (i.e. key press) with tobacco wins and another with loses and finally assessed the motivational influence of CS on behavioural outcomes in a process known as Pavlovian to Instrumental Transfer (PIT). Importantly, they found that stimuli initially associated with tobacco outcomes potentiated behavioural transfer effects, demonstrating the relevance of outcome-expectancies in reward seeking behaviours.

In line with that perspective, the dual-process theory of motivation (Dickinson & Balleine, 1994) holds that reward approach behaviours can be explained both by stimulus-outcome expectancies, resistant to stimulus revaluations, and response-outcome associations updated by momentary reward values. The results of the

aforementioned experiment by Hogarth and Chase (2011) on tobacco and chocolate devaluation show the importance of expected values on the generation of responses. However, they also found that cues associated with rewards elicited responding even after devaluation, pointing towards the independence of goal directed responses susceptible to devaluation, and more automatic or habitual responses elicited by S-R schemes.

In favour of expectancy accounts of drug-approach behaviours in humans, there is little evidence for the occurrence of conditioned responses without contingency awareness (CA), and the existing evidence does not necessarily overcome methodological criticisms (Lovibond & Shanks, 2002, discussed later on). CA refers to the ability to predict a stimulus-outcome contingency, for example, knowing that a certain reward is preceded by a certain CS. However, some instances of implicit PC effects (i.e. Clark, Manns, & Squire, 2001; Perruchet, 1985, discussed later on) opened the possibility for implicitly CS to generate some sort of response (Hogarth & Duka, 2006; Robinson & Berridge, 2003). Two streams of conditioning, implicit and explicit, may mediate drug-cue enhancement of conditioned responses through different pathways independent of conscious expectancies.

1.3.4 Incentive sensitization

A further take on automatic theories of addiction (Tiffany, 1990), indicates that addictive processes are characterised by an over-representation of the rewarding properties of an outcome or substance (Robinson & Berridge, 1993), increasing their

salience, attractiveness and motivational properties. These incentive properties are then transferred to CS associated with it.

The most characteristic trait of Robinson and Berridge's theory is the dissociation between hedonic ("liking") and motivational ("wanting)" aspects of reward learning (Berridge & Robinson, 2016).

Laboratory experiments with dopamine depleted rodents showed for example that sweet substances generated hedonic responses, assessed via orofacial expressions, but no incentive motivational properties (Berridge & Robinson, 1998). Liking can therefore occur in some instances in the absence of motivational drives towards a reward.

On the other hand, evidence shows that animals can engage in consummatory behaviours without presenting the expected hedonic responses (Wyvell & Berridge, 2000). For example, hypothalamic electric stimulation in rats was seen to generate ingestion of a sweet solution in the absence of positive hedonic reactions towards it (Berridge & Valenstein, 1991).

In humans, this same distinction may occur between "liking" and "wanting". Dopamine systems may become sensitized both to substances and drug-related stimuli, displaying exaggerated responsiveness towards them. This increased incentive salience can then drag for example addicts to consume a substance (i.e. smoke a cigarette) even in the absence of positive hedonic responses to the substance itself. The distinction between "wanting" and "liking" neural circuits has also been observed in humans via imaging (e.g. Evans et al., 2006; Volkow et al., 2002) and pharmacological (Hardman, Herbert, Brunstrom, Munafò, & Rogers, 2012; Sienkiewicz-Jarosz et al.,

2013) studies, resulting in the consideration of dopaminergic pathways as the basis for motivation or desire rather than pleasure (Berridge & Robinson, 2016).

Importantly, the authors consider that some aspects of both “liking” and “wanting” can reflect unconscious processes (Berridge, 1999). Particularly, they claimed that emotional responses to rewards can occur in the absence of conscious emotional awareness (Berridge & Winkielman, 2003). Using a subliminal emotional priming paradigm Winkielman and colleagues (2005) found that happy faces increased the value and consumption of a drink in thirsty participants, whereas negative faces decreased it. This would bring evidence for an interaction between alliesthetic states (Cabanac, 1971) and subliminal or implicit emotional processes, such that under a bodily state (thirst) that would per-se increase the value of a reward (water) implicit processes can alter further this value.

1.3.5 Dual process model

Theories reviewed previously all refer, to a greater or lesser extent, to the existence of implicit or automatic processes driving drug approach behaviours or hedonic reactions (Berridge & Winkielman, 2003; Hogarth & Duka, 2006; Tiffany, 1990).

Dual process models of addiction (Bechara, 2005; Wiers & Stacy, 2006) explain the generation of drug related behaviours as a combination of an implicit or automatic system and explicit and controlled mechanisms.

The impulsive system, based on the automatic appraisal of the hedonic and incentive value of drug-related stimuli drives implicit approach tendencies, such as attentional biases or implicit memory associations.

The reflective system on the other hand, is based on emotional regulations and explicit motivations, supposed to act as a stopper to the occurrence of pernicious drug consumption.

In adolescents for example, these automatic processes (studied using an alcohol approach-avoidance task which measures task-irrelevant behavioural tendencies towards stimuli) seem to initiate shortly after alcohol drinking starts (Peeters et al., 2012) and are seen to interact with cognitive functioning to predict the occurrence of alcohol consumption (Thush et al., 2008).

This inability to overcome or control automatic responses is therefore a definitional characteristic of addictive processes, and effective decision making and inhibitory control is meant to be impaired both by chronic (Bechara & Noel, 2006) and acute (Field, Wiers, Christiansen, Fillmore, & Verster, 2010) drug consumption, aggravating the pathological loop.

1.4 Measures of conditioned responses in humans

The theories presented before all rely their understanding of learning and addictive processes on the way individuals interact with reward related stimuli. This section presents an outline of the main methods by which hedonic, attentional, behavioural and physiological reactions towards appetitive CS are measured, whether associated with primary (i.e. food) or secondary (i.e. money) reinforcers, and whether explicitly (i.e. pictures of food) or implicitly (i.e. stimuli predicting food without awareness) related with the reward.

1.4.1 Hedonic responses

The measurement of hedonic responses independently from behavioural outcomes appears to be a relevant aspect of reward learning theories. Robinson and Berridge (2001) stress the need to separate motivational and hedonic aspects when examining reward related behaviours. From the point of view of expectancy theories of addiction, emotional responses are meant to be tied to conscious drug-expectancies as they seem to occur only in participants conscious of outcome-contingencies (Hogarth, Dickinson, Hutton, Elbers, & Duka, 2006).

Hedonic responses towards rewards or drug-related stimuli are typically assessed using Likert or visual analogue scales in which participants have to indicate their level of pleasantness, liking or appreciation towards stimuli (see Pool, Sennwald, Delplanque, Brosch, & Sander, 2016 for a detailed review).

Other measures of hedonic responsiveness towards cues are implicit association tests (IAT; Greenwald, Mcghee, & Schwartz, 1998). In this kind of task participants have to categorise a target stimulus or picture (e.g. an alcohol or soda picture) according first to one criterion (press left if the picture is a bottle of alcohol/right if it is soda). They then have to categorise other stimuli, such as valenced words, as positive or negative (press left if the word is positive/ right if negative). On the test phase, both categories are merged, and in the congruent condition the same response would classify a picture as alcohol/positive or soda/negative. In the incongruent condition, a given response would correspond to an incongruent category (alcohol/negative or soda/positive). Congruent trials should yield shorter reactions times than incongruent ones, and the strength of the association between stimulus category and their valence should be

reflected in increased differences in reaction times between both conditions. In case of studying hedonic responses, the valence categories could be replaced by “I like – I don’t like” (see Tibboel, De Houwer, & Van Bockstaele, 2015 for a review).

Importantly, hedonic responses have not only been measured towards naturalistic stimuli, but also towards abstract stimuli conditioned with other outcomes, such as tobacco (Austin & Duka, 2012) or monetary rewards (Austin & Duka, 2010; Jeffs & Duka, 2017) using Likert scales; and towards CS in evaluative tasks (Mitchell, Anderson, & Lovibond, 2003) using IAT.

Other techniques, such as effort mobilized or forced choice preferences between two cues have been employed in some cases, although they present several limitations in their application, mostly due to the difficulty to isolate motivational from hedonic components of incentive salience (Pool et al., 2016).

In that sense, attempts to measure emotional responses with subjective pleasantness ratings might not be the most valid approach as they rely on explicitly determined subjective ratings to measure an hedonic construct that might be automatic or stemming from implicit associations (Berridge, 1999). Criticisms have also been raised to the study of emotional responses using IAT (Houben & Wiers, 2006), as they can be biased by explicit knowledge about outcome-contingencies even if the corresponding stimulus and reward have never been explicitly paired (Jan De Houwer, 2006). Furthermore, semantic activation of positive and negative values is necessary for performance on the task, as well as a clear observation of the target stimulus and the representation of the semantic categories that define it.

1.4.2 Attentional and behavioural measures

According to the incentive sensitisation theory (Robinson & Berridge, 1993) stimuli predictive of reward or associated with a substance are meant to generate preferential attentional responses (Field & Cox, 2008). This preferential reactivity has traditionally been studied via dot-probe paradigms (MacLeod, Mathews, & Tata, 1986) regarding a variety of substances, such as opiates (Lubman, Peters, Mogg, Bradley, & Deakin, 2000), alcohol (Townshend & Duka, 2001) or tobacco (Hogarth, Mogg, Bradley, Duka, & Dickinson, 2003; Waters, Heishman, Lerman, & Pickworth, 2007). In this kind of task, a relevant stimulus or cue appears either on the left or right side of the screen for a short period of time, matched on the other side of the screen by a control stimulus. Stimulus presentation is followed by a dot or "probe" either congruent or incongruent with the location of the prime. Shorter reaction times for trials in which the probe replaces the relevant cue indicate an attentional preference towards that stimulus. This task has however shown poor reliability in a series of validation experiments (Price et al., 2015; Schmukle, 2005).

Other classic cognitive tasks have been modified to incorporate rewarding stimuli or drug cues. An example of this is the Stroop task (1935) which studies the interference of semantically irrelevant information in the naming of coloured words. Such an effect has been studied using alcohol related words (Bauer & Cox, 1998), tobacco cues (Janes et al., 2010), but also reward related associations (Krebs, Boehler, & Woldorff, 2010), showing that stimuli paired with rewards increase attentional interference.

In the Flanker task (Eriksen, 1995) participants have to indicate the direction of an arrow embedded in congruent flanking arrows (pointing in the same direction) or

incongruent (pointing in opposite direction), providing a measure of executive control. By adding task irrelevant alcohol related pictures in the background, Nikolaou, Field and Duka (2013) found that such distractors generated increased interferences in cognitive control compared to control stimuli.

Attentional blink tasks consist of the rapid visual serial presentation of stimuli on screen, and participants have to detect the presence of a series of stimuli, a target and a probe, embedded within the stream of pictures (Raymond, Shapiro, & Arnell, 1992). In this original example, detection of an initial target impedes recognition of a subsequent probe presented later on in the stream. A modified version, in which emotional stimuli are embedded in the stream (McHugo, Olatunji, & Zald, 2013) shows that aversive stimuli (preceding the probe) can decrease accuracy in probe detection even though they are task irrelevant. Such tasks have also been used with aversively CS as distractors (Smith, Most, Newsome, & Zald, 2006); or proven that stimuli conditioned with monetary outcomes can overcome the interference generated by emotional distractors (Yokoyama, Padmala, & Pessoa, 2015).

Finally, eye-tracking techniques have been used with drug-related stimuli to study attentional allocations towards them (e.g. Kang et al., 2012; Mogg, Bradley, Field, & De Houwer, 2003). Similar experiments have been conducted with aversive (e.g. Hogarth, Dickinson, Austin, Brown, & Duka, 2008) and reward (e.g. Austin & Duka, 2010) CS.

The occurrence of preferential attentional responses is again hypothesized to be based on the explicit knowledge (mental representations) of stimulus-outcome contingencies (Field & Cox, 2008; Hogarth, Dickinson, Hutton, Bamborough, & Duka, 2006).

According to these authors, the conscious expectancy of drug availability signalled by these stimuli generates Conditioned Responses (CResp) compatible with the stimulus presented, in this case attentional reactivity. On top of that, such stimuli can interact with craving states. Attentional responses towards drug CS can either generate craving responses, or be modulated by subjective craving (Field, Munafò, & Franken, 2009) or deprivation states (Field, Mogg, & Bradley, 2004). On the other hand, Wiers and Stacy (2006) propose that such measures assess implicit processes reflecting *“"deeper" affective mechanisms that operate outside awareness*”. The existence and study of those implicit processes can be determinant for the understanding of addiction.

1.4.3 Physiological responses

Cue reactivity, understood as the automatic generation of bodily responses towards CS, has also been examined assessing a variety of physiological responses towards drug-related stimuli (Carter & Tiffany, 1999). In this type of procedure, participants are exposed to a cue (e.g. alcohol bottles) whilst they report different variables, such as craving states or pleasantness responses. Physiological recordings usually can consist of heart-rate responsiveness, showing, for example, that increased heart-rate variability in the presence of alcohol stimuli can predict clinical relapse in patients suffering from Alcohol use disorders (Garland, Franken, & Howard, 2012). Stimuli conditioned with cigarette puffs also were found to generate increased physiological reactivity, as measured via skin conductance response (Winkler et al., 2011). Furthermore, stimuli associated with pleasant and unpleasant outcomes (i.e. emotional pictures) also appear to modulate heart-rate variability (e.g. Lachnit & Kimmel, 1993; Pollatos & Schandry, 2008). Research into physiological conditioned

responses has relevance to the understanding of how physiological mechanisms underlie the development of conditioned responses (Bindra, 1974), and provide an insight into the relationship between interoceptive processes, the feeling of internal bodily states, and addiction (Gray & Critchley, 2007) discussed later on.

1.5 Implicit learning

1.5.1 General considerations

Implicit learning (Reber, 1989) can be defined as a learning task that results in implicit knowledge, namely knowledge that one is not aware of possessing. In the case of PC, this knowledge consists of stimulus-outcome contingencies subjects integrate without conscious awareness.

The first examples of implicit learning were developed with artificial grammar learning paradigms (Reber, 1967). This kind of paradigm starts with a learning phase in which participants are presented with a series of non-words, the content of which is based on a set of rules determining which letters are contained in each string and their order within it. In a test phase participants are told the strings are built according to a set of rules, but not told which rules those are. They are then presented with a novel set of stimuli of which some strings will respect the original grammar rules and some, that although being similar to the original ones, will violate the rules. Their task is then to classify each novel string as correct or incorrect based on their prior experience. Above chance accuracy on classification indicates the occurrence of learning, and by asking participants about their explicit knowledge about the grammar rules, it is possible to determine whether this learning is implicit or explicit.

Another example of a classic implicit learning experiment is sequence learning. Participants are presented with a sequence of stimuli on screen that follows a structured pattern, for example a dot appearing at different locations within a matrix. After a learning phase, participants are presented with bits of the sequence and are asked to predict the position of the next stimulus, a task they can perform better than at chance levels without being able to explicitly report the structure of the sequence (Cleeremans, Destrebecqz, & Boyer, 1998).

The knowledge being acquired in a conditioning paradigm is that of the contingency between the CS and the reward or outcome. If this can be shown to have been acquired without the participant being aware of possessing that knowledge, then implicit learning can be said to have occurred. The implicit nature of this knowledge is demonstrated if participants develop preferential responses toward CS, in the form of hedonic, attentional or physiological reactivity (see previous section), but are unable to consciously predict the reward or outcome.

1.5.2 Evaluative Conditioning

In the context of conditioning paradigms, the clearest examples of implicit learning have been found within evaluative conditioning.

Evaluative Conditioning (EC) paradigms are based on the principle that value transfers between stimuli presented concurrently (Levey & Martin, 1975). For example, if the picture of an abstract shape is presented at the same time as a very pleasant painting, then the abstract shape may increase its perceived valence, whilst the implicit (without

awareness) knowledge that a certain stimulus is followed by a pleasant (or unpleasant) picture is acquired.

EC has long been seen as different from PC tasks for a variety of reasons (De Houwer, Thomas, & Baeyens, 2001). One of the main distinctions is the alleged ability of EC to occur in the absence of contingency awareness, that is, for participants to be unable to consciously recall the association between pictures as a source of learning (Baeyens, Eelen, & Bergh, 1990; De Houwer et al., 2001; Field, 2000; Hutter, Sweldens, Stahl, Unkelbach, & Klauer, 2012).

The tasks employed to measure EC transfer are, among others, IAT (Mitchell et al., 2003) or pleasantness measurements, which have however been shown to be dependent on CA (Pleyers, Corneille, Luminet, & Yzerbyt, 2007). Indeed, multiple sources of evidence suggest that CA modulates the development of valence transfer (De Houwer, 2014; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010) and that subliminal priming tasks are not effective at generating EC effects (Högden, Hütter, & Unkelbach, 2018; Stahl, Haaf, & Corneille, 2016). It seems therefore that the role of CA in EC is still to be clarified and would benefit from a methodological improvement.

Interestingly, the impact of EC effects on implicit and explicit attitudes (De Houwer, 2014) can have implications for the generation of novel treatments in drug-addiction, targeting attitudes towards alcohol (Zerhouni, Bègue, Comiran, & Wiers, 2018).

1.5.3 Pavlovian Conditioning

1.5.3.1 Against the existence of implicit Pavlovian Conditioning

A similar debate to the one regarding implicit EC has arisen in PC. Here evidence seems to even more strongly support an inability of CS to generate responses in the absence of expectancy awareness (Lovibond & Shanks, 2002); even within my own team (Hogarth, Dickinson, & Duka, 2005; Hogarth, Dickinson, Hutton, Bamborough, et al., 2006; Hogarth, Dickinson, Hutton, Elbers, et al., 2006).

Lovibond and Shanks (2002) reviewed the existing literature at the time on implicit PC and EC, covering subliminal, aversive and autonomous conditioning. They carefully rejected most of the evidence in favour of implicit PC, mostly based on the inadequacy of CA measurements.

For example, CA in PC has been assessed using a variety of techniques, such as post-conditioning assessments, that is, at the end of the conditioning task. Those measurements can either be performed using free reports (i.e. asking participants at the end of the task to recall the rule or structure underlying the task) or with forced choice recognition tasks (e.g. asking them to indicate whether a monetary win or loss did follow a series of stimuli).

Dienes & Berry (1997) already criticized free reports as a measure of explicit knowledge due to their unreliability. For example, a subject unable to freely recall an element at a given time might be able to do so if given another opportunity (Erdelyi & Becker, 1974). Free report may also entail the recollection of large amounts of information of which the subject is not confident, constituting an insensitive measure.

Thus, CA should be assessed with measures sensitive and informative to the task in hand, such as cued reports or forced choice tests (Shanks & St John, 1994). Lovibond and Shanks (2002) also indicated that successful measures of CA must be based on online accounts of US predictability. That is, as post-conditioning or retrospective measures of CA appear to be unreliable, participants should be asked during stimulus presentation if they can predict the occurrence of US. Finally, those measures should be based, when possible, on continuous rating scales rather than forced choice dichotomous questions.

Hogarth et al. (2006) introduced the use of continuous Likert scales to measure CA during an appetitive learning task. Most research in the field of implicit learning considers the absence of significant structural recollection or visual detection as evidence of a lack of conscious awareness. Accordingly, in the above example a participant would be deemed Unaware of contingencies if they fail to assign significantly higher ratings in the Likert scale for stimuli predicting reward than for those not predicting reward.

However, determining the implicit nature of a phenomenon requires accepting the null hypothesis that no knowledge has been developed, and orthodox statistics do not provide a means to determine whether results below common significance thresholds can be informative against the posited theory or are just insensitive. This can have severe implications for the overrepresentation of implicit processes (Shanks, 2016), constituting one of the main weaknesses in implicit PC studies.

For that reason, it is necessary to implement Bayesian approaches to clarify the unconscious nature of a cognitive process (Dienes, 2015). Bayesian statistics provide a tool for examining the sensitivity of non-significant results (Dienes, 2014), it is possible therefore to confidently categorise a subject as Unaware of contingencies if the results from the Bayesian analysis are sensitively null. Otherwise results are non-sensitive, and it is not possible to determine the state of the subject.

1.5.3.2 Evidence for the existence of Implicit Pavlovian Conditioning

One of the earliest evaluations of implicit Pavlovian aversive conditioning effects was reported by Perruchet (1985). In that experiment, a tone was presented on each trial, paired either with an air puff directed to the subject's eye (S+) or with no outcome (S-) on random runs of trials. As a result, participants developed an eye-blink CResp (closing their eyes in the presence of a tone). Interestingly, explicit air puff expectancies were dissociated from eye-blink CResp. During runs in which a puff was delivered after the tone, participants' expectancy ratings decreased, reflecting an anticipation of trial switching towards non-paired tones. These changes in expectancies were however accompanied by an increase in eye-blink CResp. The opposite response pattern occurred for S- trials, demonstrating that CResp can occur in the absence of congruent outcome-contingency expectancies.

Recently, several instances of appetitive PC in the absence of CA have been published. Using a subliminal conditioning task, results supporting implicit PC were reported (Bourgeois, Neveu, & Vuilleumier, 2016). Importantly, this task incorporated both concurrent and post-conditioning measures of prime detection, which resulted

negative, and results show participants developed selective attention towards implicitly rewarded CS.

However, the experiments presented in this thesis are based on paradigms in which stimuli are supraliminally presented, that is, fully accessible to conscious visual observation in contrast to subliminally presented stimuli. The generation of unconscious processes relies therefore on dual-task distractions and on probabilistic manipulations. Anderson and colleagues (2011) used a dual-task paradigm in which they presented an array of stimuli composed of differently coloured circles containing a line. Participants' task was to indicate the direction of the line contained within a specific target circle. Upon correct detection of the line's orientation, a particular coloured target would predict higher chances of obtaining a reward than others. In a test phase, a similar matrix appeared, but this time the task was to detect the orientation of the line within a solitary square surrounded by distracting circles. By including on 50% of the test trials the previously rewarded target as a task-irrelevant distractor, it was found that reward CS can generate value driven attentional responses.

Using a similar paradigm, Anderson found that participants unable to report the contingencies still showed signs of appetitive learning indexed by attentional interference effects (Anderson, 2015b), although the measurement of CA was based on a six-alternative post-conditioning recall, and hence not following the guidelines presented above.

Jeffs and Duka (2017) showed how monetarily rewarded CS in the absence of CA were able to generate subjective hedonic responses, albeit the hedonic value generated by implicit conditioning was not sufficient to trigger behavioural effects, congruent with previous research (Hogarth et al., 2007). However, to our knowledge, neither this experiment nor others have evaluated both Type I and Type II knowledge of outcome expectancies (explained below) using methodologies outlined in the previous sections. The inclusion of such measurements could provide new insights into the understanding of implicit PC.

Finally, using a task irrelevant reward conditioning task, in which pictures of *houses* or *buildings* were associated with different probabilities of monetary outcome, Yokoyama et al. (2015) found that those CS generated attentional preference on an Emotional Attentional Blink task. However, they did not measure CA, preventing an examination of the implicit correlates of their findings.

1.5.4 Signal Detection Theory and metacognition

Importantly, in the context of implicit processing, the differentiation between objective and subjective evaluations is central to the characterisation of learning effects, allowing one to determine the presence of metacognitive knowledge, the awareness of one's own mental states (Metcalfe, 1996). Cheesman and Merikle (1984) studied the detection of subliminal stimulus presentations and found that participants were able to determine whether a stimulus had been presented or not, although they were not confidently able to report their detection accuracy. Therefore, a lack of meta-knowledge can occur if a participant displays above chance scores on an objective

level, but has no confidence or lacks the ability to discriminate the likely accuracy of their different responses. The absence of a correlation between the accuracy of individual responses and confidence in those responses indicates a lack of metacognitive awareness (Dienes, Altmann, Kwan, & Goode, 1995).

In the case of outcome-expectancy measures in a conditioning task, objective measures determine the ability to predict a reward, e.g. using Likert scales ("How likely is X event to occur from 1 to 5") or dichotomous accounts ("Will X event occur: Yes/No").

Out of dichotomous measures of expectancy it is possible to extract Type I d' scores using Signal Detection Theory (SDT) analyses (Green & Swets, 1966; Stanislaw & Todorov, 1999). In traditional terms, Type I d' scores provide a measure of sensitivity in discriminating a stimulus (indicating it is present when it is actually present – a *Hit*– or that it is absent when in truth it is absent – a *Correct Rejection*) accounting for response biases, i.e. the tendency for a participant to indicate they have seen a stimulus or not, regardless of whether it is actually present. The measure is derived from a comparison of Hit rates, the probability of being accurate in the presence of a stimulus, and False Alarm rates, the probability of reporting the presence of a stimulus in its absence. Type I d' is simply the z-score of the Hit rate minus the z-score of the False Alarm rate.

SDT has also been applied to the study of metacognitive knowledge. Using dichotomous confidence responses, Type II d' scores can be computed to determine the consciousness of their ability to make accurate discriminations (Barrett, Dienes, &

Seth, 2013). In the case of Type II d' scores, a *Hit* is considered as a correct response (Type I *Hit* or *Correct Rejection*) accompanied by a high confidence rating, whereas a *False Alarm* is a Type I incorrect response rated as high confident. Higher Type II d' scores signal therefore an appropriate relationship between accuracy and confidence, denoting conscious awareness of their ability to make accurate discriminations.

In instances where a continuous variable, rather than a dichotomous variable, is used to determine confidence in a response, an Area Under the Receiver Operating Characteristic (ROC) curve can provide a measure of metacognitive knowledge. This analysis offers an estimate of the relationship between a state variable (e.g. accuracy, measured via binary response) and a continuous one (confidence, measured via visual analogue or Likert scales). For each detection threshold, Type II *Hit* vs. *False Alarm* rates are computed, generating an area under the curve which shows the extent to which accuracy matches confidence.

Analyses of metacognition based on Type II SDT offer a measure of metacognition independent of confidence biases, that is, the propensity to respond with high or low confidence. However, these analyses can be affected by Type I accuracy scores in the case of extreme decision thresholds for confidence and accuracy (Barrett et al., 2013), and when possible, other measures of metacognition, such as the meta d' (Fleming & Lau, 2014; Maniscalco & Lau, 2014), should be employed.

1.6 Interoception and the generation of Conditioned responses

1.6.1 Interoception and emotion

Interoception is defined as the ability to perceive or detect internal bodily sensations (Cameron, 2001; Sherrington, 1948). Interoception not only corresponds to the sensing or integrating of physiological bodily states, but also has ties to the motivational needs of the subject (Craig, 2009). These processes derive either from basic allostatic and homeostatic information (e.g. thirst), from bodily responses (e.g. skin reactivity), or from the influence of pleasant or unpleasant external stimuli (e.g. skin touch) (Tsakiris & Critchley, 2016).

Visceral, thermoregulatory and inflammatory signals conform the basis of interoceptive signals and are integrated in the insular and orbitofrontal cortices via spinothalomocortical pathways (Craig, 2002), constituting the neural hubs in which interoceptive information is processed (Critchley, Wiens, Rotshtein, Ohman, & Dolan, 2004).

Early theories of emotion posited that emotional experiences are derived from the detection of physiological states induced by stimuli (James, 1884), and as such, arousal states can modulate or participate in the development of feelings (Garfinkel & Critchley, 2013a; Wiens, 2005). The relationship between interoception and emotional processes has been demonstrated using a variety of techniques (Critchley & Garfinkel, 2017). For example, participants with high interoceptive abilities reported increased reactivity towards emotionally laden videos than those with low interoceptive abilities (Wiens, Mezzacappa, & Katkin, 2000). Interoception also seems to modulate the

development of physiological responses (heart-rate reactivity) and subjective hedonic ratings towards emotional stimuli (Pollatos, Herbert, Matthias, & Schandry, 2007), impacting recall accuracy on emotional learning tasks (Pfeifer et al., 2017; Pollatos & Schandry, 2008). Using an emotional attentional blink, Garfinkel et al. (2014) found that the presentation of emotional probes at cardiac systole (marking the ejection of blood into arteries and the signalling of cardiac functioning by baroreflexors) facilitates the detection of fearful faces. This exemplifies how interoceptive signals amplify the emotional salience of stimuli, increasing emotional experience. Finally, from a neuro-anatomical point of view, it has been shown that interoceptive and emotional responses both share common neural substrates in the insula (Zaki, Davis, & Ochsner, 2012). As explained later on, the interoceptive correlates of emotional processing can be linked to the development of learnt responses, providing an explanation for the apparent link between interoception and addiction.

A series of tasks are employed to measure sensitivity to bodily functions, usually focusing on the detection of cardiac responsiveness (Garfinkel, Seth, Barrett, Suzuki, & Critchley, 2015). One of these procedures, known as heartbeat tracking or counting task (Schandry, 1981) consists of requesting participants to report the number of heartbeats they have perceived in different periods of time. The heartbeat discrimination task (Katkin, Blascovich, & Goldband, 1981; Whitehead, Drescher, Heiman, & Blackwell, 1977) on the other hand requires participants to indicate whether a tone is presented in synchrony or not with their own heartbeat.

These two tasks provide measures of interoceptive accuracy, which can be accompanied by trial-by-trial measures of confidence to provide an estimate of

interoceptive sensibility. These two indexes can then be used to compute Type II measures of meta-cognitive interoceptive awareness, also known as interoceptive insight (Khalsa et al., 2017).

1.6.2 Interoception, addiction and reward learning

Much importantly for the field of addiction, interoceptive processes have been found to be related with several aspects of substance use disorders (Paulus & Stewart, 2014; Verdejo-Garcia, Clark, & Dunn, 2012). Besides lesion studies (Naqvi, Rudrauf, Damasio, & Bechara, 2007) showing that insula damage can provoke smoking cessation in tobacco addicts, interoceptive processes can also explain the development or transition to addictive disorders (Stewart, May, Tapert, & Paulus, 2015). Alcoholics on the other hand can show decreases in interoceptive sensibility (Ateş Çöl, Sönmez, & Vardar, 2016).

Particularly linked to reward learning theories of addiction is the evidence for the involvement of interoceptive signalling in the development of reward prediction (Paulus & Stewart, 2014) and hedonic responses towards incentive stimuli through heightened perception of their salience (Paulus, 2007). In that regard, activity of certain insular areas correlates with the detection of reward probabilities (Burke & Tobler, 2011) and reward learning (Cousijn et al., 2013). Interoceptive abilities also seem to correlate with avoidance of loss in a monetary learning task involving risky decision making (Sokol-Hessner, Hartley, Hamilton, & Phelps, 2015). The amplification of aversive physiological responses generated by reward losses (Sokol-Hessner et al., 2009), in combination with high interoceptive abilities, appears to shape behavioural

responses in uncertain conditions. In addition, Kandasamy et al. (2016) showed how interoceptive abilities explain the performance of traders in London floor, further linking reward prediction with the perception of “gut feelings”. Regarding aversive conditioning, insular thickness seems to participate in the development of fear responses (Hartley, Fischl, & Phelps, 2011); and using a subliminal conditioning task it was shown that participants with high interoceptive abilities better learnt aversive stimulus-outcome contingencies (Katkin, Wiens, & Ohman, 2001).

This experimental evidence shows the link between interoception and reward prediction or the development of CA. In conjunction with the amplification of emotional experiences, the representation of physiological reactions generated by rewards (or negative events) in interoceptive hubs could shape the formation of conditioned responses. This could occur through several mechanisms, influenced primarily by interoceptive abilities (Naqvi & Bechara, 2010).

For example, the involvement of interoception in emotional experience supports the alliesthetic perspective of addiction (Paulus, Tapert, & Schulteis, 2009). According to this view, the value of stimuli depends on the subject’s internal states, and interoceptive abilities can mediate and amplify the perception of those states. Higher ability to perceive internal bodily sensations would facilitate the perception of bowel movements, such as emptiness in the stomach. This would promote the generation of increased feelings of hunger, which according to alliesthetic theories would determine the perceived hedonic value of food, following the views outlined earlier on regarding incentive learning (Bindra, 1978; Toates, 1986). Increased perception of the hedonic values of stimuli would intensify their incentive salience and therefore foster conditioned responses.

The generation of craving states on the other hand can be understood as a response to the appraisal of physiological responses generated by reward conditioned stimuli, which activate the mental representations of the stimulus-outcome relationships (Gray & Critchley, 2007).

1.7 Aims of the current thesis

The literature reviewed shows the growing relevance of implicit processes in addictive behaviours. Progressively drawing form early conditioning models to more complex expectancy accounts of reward learning, several theories of addiction have mirrored advances in the field of PC. Therefore, novel research in the field of implicit reward learning could promote innovative perspectives in the field of addiction, particularly in the study of automatic correlates.

On one hand, expectancy theories of addiction suggest that CA is necessary for the development of approach behaviours (Hogarth, Dickinson, Hutton, Elbers, et al., 2006), following the lack of evidence for implicit PC effects (Lovibond & Shanks, 2002; Pleyers et al., 2007). Although not denying a partial role of implicit components in addiction (Hogarth & Duka, 2006), these theories somehow clash with dual process theories that assign a high relevance to unconscious cognitions, particularly in the development of drug-approach behaviours (Belin et al., 2013; Wiers & Stacy, 2006). Wiers and Stacy propose that drug related stimuli can generate approach behaviours in an automatic way. However, the role of outcome expectancies in this model has not yet been clarified. Studies on automatic processes related with addiction usually employ stimuli directly associated with the substance (i.e. pictures of a pack of cigarettes or an alcohol

bottle). In that case, the signalling of drug availability is clear and explicit, and previous consumption experiences can affect responses towards them (Wiers et al., 2002).

In order to tackle that problem, conditioning neutral stimuli with primary or secondary reinforcers can provide a window to the generation of truly implicit processes. The lack of solid evidence towards implicit accounts of PC (Dedonder, Corneille, Bertinchamps, & Yzerbyt, 2013; Le Pelley, Seabrooke, Kennedy, Pearson, & Most, 2017; Lovibond & Shanks, 2002; Pleyers et al., 2007) requires a deeper examination of the techniques used to detect subjective and objective correlates of learning, and where possible their improvement.

As explained before, providing useful and solid evidence towards the existence of implicit processes requires a careful measurement of learning effects (Dienes, 2015; Shanks, 2016). In the experiments presented next an effort is made to follow the guidelines pointed out by previous authors (Lovibond & Shanks, 2002) in order to limit the existence of false negatives (wrongly assuming a participant has no conscious knowledge) leading to the overrepresentation of implicit states.

Previous measurements of learning effects might not have been the most sensitive in order to detect implicit processes (Tibboel et al., 2011). IAT for instance can be affected by the conscious representation of stimuli (Jan De Houwer, 2006; Houben & Wiers, 2006). Moreover, subjective measures of hedonic valence rely on the explicit report of pleasantness to assess the existence of an implicit content which might not be readily accessible in subjective terms. The following experiments will attempt therefore to overcome these pitfalls, utilising procedures and measurements that

allow the study of implicit PC. Accurate measures of metacognition in learning can provide a novel and more complete approach in the study of PC and the role of expectancies in reward learning.

Finally, interoception in its role as a mediator of learning processes, reward prediction and hedonic responses, appears to be an important factor in the study of reward conditioning and in the integration of physiological responses towards CS and US in behavioural outputs. By clarifying the role of interoception in conditioning learning we might partly explain also its involvement in drug addiction (Paulus & Stewart, 2014; Paulus et al., 2009).

The following experiments will focus on the study of PC with the subsequent aims:

1. Providing evidence for the existence of implicit appetitive PC using attentional tasks to evaluate conditioned responses, and cognitive interference paradigms to assess the interplay between explicit and implicit cognitive control systems in reward approach behaviours.
2. Improving the measurements of CA in PC, contemplating previous methodological issues and implementing Bayesian analyses for the determination of unconscious knowledge.
3. Investigating the interoceptive correlates of appetitive PC, both in the development of reward prediction and hedonic responses.
4. Examining the relationship between interoceptive awareness and the perception of acute substance (i.e. alcohol) effects.
5. Integrating the findings within learning and addiction theories.

1.8 References

Anderson, B. A. (2015). Value-driven attentional priority is context specific. *Psychonomic Bulletin & Review*, *22*(3), 750–756. http://doi.org/10.3758/s13423-014-0724-0

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–71. http://doi.org/10.1073/pnas.1104047108

Ateş Çöl, I., Sönmez, M. B., & Vardar, M. E. (2016). Evaluation of Interoceptive Awareness in Alcohol-Addicted Patients. *Noro Psikiyatri Arsivi*, *53*(1), 17–22. http://doi.org/10.5152/npa.2015.9898

Austin, A. J., & Duka, T. (2012). Mechanisms of attention to conditioned stimuli predictive of a cigarette outcome. *Behavioural Brain Research*, *232*(1), 183–189. http://doi.org/10.1016/J.BBR.2012.04.009

Austin, A. J. J., & Duka, T. (2010). Mechanisms of attention for appetitive and aversive outcomes in Pavlovian conditioning. *Behavioural Brain Research*, *213*(1), 19–26. http://doi.org/10.1016/j.bbr.2010.04.019

Baeyens, F., Eelen, P., & Bergh, O. van den. (1990). Contingency awareness in evaluative conditioning: A case for unaware affective-evaluative learning. *Cognition & Emotion*, *4*(1), 3–18. http://doi.org/10.1080/02699939008406760

Baker, T. B., Morse, E., & Sherman, J. E. (1986). The motivation to use drugs: a psychobiological analysis of urges. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, *34*, 257–323.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. http://doi.org/10.1037/a0033268

Bates, M. E., Pawlak, A. P., Tonigan, J. S., & Buckman, J. F. (2006). Cognitive impairment influences drinking outcome by altering therapeutic mechanisms of change. *Psychology of Addictive Behaviors*, *20*(3), 241–253. http://doi.org/10.1037/0893-164X.20.3.241

Bauer, D., & Cox, W. M. (1998). Alcohol-related words are distracting to both alcohol abusers and non-abusers in the Stroop colour-naming task. *Addiction*, *93*(10), 1539–1542. http://doi.org/10.1046/j.1360-0443.1998.9310153910.x

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective. *Nature Neuroscience*, *8*(11), 1458–63. http://doi.org/10.1038/nn1584

Bechara, A., & Noel, X. (2006). Loss of willpower: Abnormal neural mechanisms of impulse control and decision making in addiction. In *Handbook of implicit cognition and addiction* (1st ed., pp. 215–232).

Belin, D., Belin-Rauscent, A., Murray, J. E., & Everitt, B. J. (2013). Addiction: failure of

control over maladaptive incentive habits. *Current Opinion in Neurobiology*, *23*(4), 564–572. http://doi.org/10.1016/J.CONB.2013.01.025

Berridge, K. C. (1991). Modulation of taste affect by hunger, caloric satiety, and sensory-specific satiety in the rat. *Appetite*, *16*(2), 103–120. http://doi.org/10.1016/0195-6663(91)90036-R

Berridge, K. C. (1999). Pleasure, pain, desire, and dread: Hidden core processes of emotion. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations ofhedonic psychology* (pp. 525–557). New York: Russell Sage Foundation.

Berridge, K. C. (2000). Reward learning: Reinforcement, incentives, and expectations. *Psychology of Learning and Motivation*, *40*, 223–278. http://doi.org/10.1016/S0079-7421(00)80022-5

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, *28*(3), 309–369. http://doi.org/10.1016/S0165-0173(98)00019-8

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *The American Psychologist*, *71*(8), 670–679. http://doi.org/10.1037/amp0000059

Berridge, K. C., & Valenstein, E. S. (1991). What psychological process mediates feeding evoked by electrical stimulation of the lateral hypothalamus? *Behavioral Neuroscience*, *105*(1), 3–14.

Berridge, K., & Winkielman, P. (2003). What is an unconscious emotion?(The case for unconscious “liking”). *Cognition & Emotion*, *17*(2), 181–211. http://doi.org/10.1080/02699930302289

Bindra, D. (1968). Neuropsychological interpretation of the effects of drive and incentive-motivation on general activity and instrumental behavior. *Psychological Review*, *75*(1), 1–22. http://doi.org/10.1037/h0025306

Bindra, D. (1974). A motivational view of learning, performance, and behavior modification. *Psychological Review*, *81*(3), 199–213. http://doi.org/10.1037/h0036330

Bindra, D. (1978). How adaptive behavior is produced: a perceptual-motivational alternative to response reinforcements. *Behavioral and Brain Sciences*, *1*(01), 41. http://doi.org/10.1017/S0140525X00059380

Bolles, R. C. (1972). Reinforcement, expectancy, and learning. *Psychological Review*, *79*(5), 394–409. http://doi.org/10.1037/h0033120

Bolles, R. C., & Fanselow, M. S. (1980). PDR - a multi-level model of fear and pain. *Behavioral and Brain Sciences*, *3*(02), 315. http://doi.org/10.1017/S0140525X00005136

Bourgeois, A., Neveu, R., & Vuilleumier, P. (2016). How Does Awareness Modulate Goal-Directed and Stimulus-Driven Shifts of Attention Triggered by Value

Learning? *PLOS ONE*, *11*(8), e0160469. http://doi.org/10.1371/journal.pone.0160469

Brandon, T. H., Herzog, T. A., Irvin, J. E., & Gwaltney, C. J. (2004). Cognitive and social learning models of drug dependence: implications for the assessment of tobacco dependence in adolescents. *Addiction*, *99*, 51–77. http://doi.org/10.1111/j.1360-0443.2004.00737.x

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*(11), 681–684. http://doi.org/10.1037/h0040090

Brown, C. R. H., Duka, T., & Forster, S. (2018). Attentional capture by alcohol-related stimuli may be activated involuntarily by top-down search goals. *Psychopharmacology*. http://doi.org/10.1007/s00213-018-4906-8

Burke, C. J., & Tobler, P. N. (2011). Reward skewness coding in the insula independent of probability and loss. *Journal of Neurophysiology*, *106*(5), 2415–22. http://doi.org/10.1152/jn.00471.2011

Cabanac, M. (1971). Physiological role of pleasure. *Science (New York, N.Y.)*, *173*(4002), 1103–7.

Cabanac, M. (1979). Sensory Pleasure. *The Quarterly Review of Biology*, *54*(1), 1–29. http://doi.org/10.1086/410981.

Cameron, O. G. (2001). Interoception: The Inside Story—a Model for Psychosomatic Processes. *Psychosomatic Medicine*, *63*(5), 697–710.

Carter, B. L., & Tiffany, S. T. (1999). Meta-analysis of cue-reactivity in addiction research. *Addiction (Abingdon, England)*, *94*(3), 327–40.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, *36*(4), 387–395. http://doi.org/10.3758/BF03202793

Clark, R. E., Manns, J. R., & Squire, L. R. (2001). Trace and Delay Eyeblink Conditioning: Contrasting Phenomena of Declarative and Nondeclarative Memory. *Psychological Science*, *12*(4), 304–308. http://doi.org/10.1111/1467-9280.00356

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: news from the front. *Trends in Cognitive Sciences*, *2*(10), 406–416. http://doi.org/10.1016/S1364-6613(98)01232-7

Cornell, C. E., Rodin, J., & Weingarten, H. (1989). Stimulus-induced eating when satiated. *Physiology & Behavior*, *45*(4), 695–704. http://doi.org/10.1016/0031-9384(89)90281-3

Cousijn, J., Wiers, R. W., Ridderinkhof, K. R., van den Brink, W., Veltman, D. J., Porrino, L. J., & Goudriaan, A. E. (2013). Individual differences in decision making and reward processing predict changes in cannabis use: a prospective functional magnetic resonance imaging study. *Addiction Biology*, *18*(6), 1013–1023. http://doi.org/10.1111/j.1369-1600.2012.00498.x

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews. Neuroscience*, *3*(8), 655–66.

http://doi.org/10.1038/nrn894

Craig, A. D. (2009). How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1), 59–70. http://doi.org/10.1038/nrn2555

Cristea, I. A., Kok, R. N., & Cuijpers, P. (2016). The Effectiveness of Cognitive Bias Modification Interventions for Substance Addictions: A Meta-Analysis. *PLOS ONE*, *11*(9), e0162226. http://doi.org/10.1371/journal.pone.0162226

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*(2), 189–95. http://doi.org/10.1038/nn1176

Critchley, H., & Garfinkel, S. (2017). Interoception and emotion. *Current Opinion in Psychology*, *17*, 7–14. http://doi.org/10.1016/J.COPSYC.2017.04.020

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. http://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J. (2014). A propositional perspective on context effects in human associative learning. *Behavioural Processes*, *104*, 20–25. http://doi.org/10.1016/J.BEPROC.2014.02.002

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: a review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*(6), 853–69.

Dedonder, J., Corneille, O., Bertinchamps, D., & Yzerbyt, V. (2013). Overcoming Correlational Pitfalls: Experimental Evidence Suggests That Evaluative Conditioning Occurs for Explicit But Not Implicit Encoding of CS-US Pairings. *Social Psychological and Personality Science*, *5*(2), 250–257. http://doi.org/10.1177/1948550613490969

Dickinson, A. (1989). *Expectancy theory in animal conditioning.*

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18. http://doi.org/10.3758/BF03199951

Dickinson, A., & Balleine, B. (1995). Motivational Control of Instrumental Action. *Current Directions in Psychological Science*, *4*(5), 162–167. http://doi.org/10.1111/1467-8721.ep11512272

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. http://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. *Behavioural Methods in Consciousness Research*, 199–220.

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1322–1338.

http://doi.org/10.1037/0278-7393.21.5.1322

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, *4*(1), 3–23.

Duka, T., & Townshend, J. M. (2004). The priming effect of alcohol pre-load on attentional bias to alcohol-related stimuli. *Psychopharmacology*, *176*(3–4), 353–61. http://doi.org/10.1007/s00213-004-1906-7

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology*, *6*(1), 159–171. http://doi.org/10.1016/0010-0285(74)90008-5

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, *2*(2–3), 101–118. http://doi.org/10.1080/13506289508401726

Evans, A. H., Pavese, N., Lawrence, A. D., Tai, Y. F., Appel, S., Doder, M., … Piccini, P. (2006). Compulsive drug use linked to sensitized ventral striatal dopamine transmission. *Annals of Neurology*, *59*(5), 852–8. http://doi.org/10.1002/ana.20822

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. http://doi.org/10.1038/nn1579

Field, a P. (2000). I like it, but I'm not sure why: can evaluative conditioning occur without conscious awareness? *Consciousness and Cognition*, *9*(1), 13–36. http://doi.org/10.1006/ccog.1999.0402

Field, M., & Cox, W. M. (2008). Attentional bias in addictive behaviors: a review of its development, causes, and consequences. *Drug and Alcohol Dependence*, *97*(1–2), 1–20. http://doi.org/10.1016/j.drugalcdep.2008.03.030

Field, M., Mogg, K., & Bradley, B. P. (2004). Eye movements to smoking-related cues: effects of nicotine deprivation. *Psychopharmacology*, *173*(1–2), 116–23. http://doi.org/10.1007/s00213-003-1689-2

Field, M., Munafò, M. R., & Franken, I. H. A. (2009). A meta-analytic investigation of the relationship between attentional bias and subjective craving in substance abuse. *Psychological Bulletin*. Field, Matt: School of Psychology, University of Liverpool, Liverpool, United Kingdom, Liverpool, mfield@liverpool.ac.uk: American Psychological Association. http://doi.org/10.1037/a0015843

Field, M., Wiers, R. W., Christiansen, P., Fillmore, M. T., & Verster, J. C. (2010). Acute alcohol effects on inhibitory control and implicit cognition: implications for loss of control over drinking. *Alcoholism, Clinical and Experimental Research*, *34*(8), 1346–52. http://doi.org/10.1111/j.1530-0277.2010.01218.x

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. http://doi.org/10.3389/fnhum.2014.00443

Garfinkel, S. N., & Critchley, H. D. (2013). Interoception, emotion and brain: new

insights link internal physiology to social behaviour. Commentary on:: “Anterior insular cortex mediates bodily sensibility and social anxiety” by Terasawa et al. (2012). *Social Cognitive and Affective Neuroscience*, *8*(3), 231–4. http://doi.org/10.1093/scan/nss140

Garfinkel, S. N., Minati, L., Gray, M. A., Seth, A. K., Dolan, R. J., & Critchley, H. D. (2014). Fear from the heart: sensitivity to fear stimuli depends on individual heartbeats. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(19), 6573–82. http://doi.org/10.1523/JNEUROSCI.3507-13.2014

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. http://doi.org/10.1016/j.biopsycho.2014.11.004

Garland, E. L., Franken, I. H. A., & Howard, M. O. (2012). Cue-elicited heart rate variability and attentional bias predict alcohol relapse following treatment. *Psychopharmacology*, *222*(1), 17–26. http://doi.org/10.1007/s00213-011-2618-4

Goldstein, R. Z., Craig, A. D. (Bud), Bechara, A., Garavan, H., Childress, A. R., Paulus, M. P., & Volkow, N. D. (2009). The Neurocircuitry of Impaired Insight in Drug Addiction. *Trends in Cognitive Sciences*, *13*(9), 372–380. http://doi.org/10.1016/J.TICS.2009.06.004

Gray, M. A., & Critchley, H. D. (2007). Interoceptive Basis to Craving. *Neuron*, *54*(2), 183–186.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *New York: Wiley*.

Greenwald, A. G., Mcghee, D. E., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Hardman, C. A., Herbert, V. M. B., Brunstrom, J. M., Munafò, M. R., & Rogers, P. J. (2012). Dopamine and food reward: effects of acute tyrosine/phenylalanine depletion on appetite. *Physiology & Behavior*, *105*(5), 1202–7. http://doi.org/10.1016/j.physbeh.2011.12.022

Hartley, C. A., Fischl, B., & Phelps, E. A. (2011). Brain structure correlates of individual differences in the acquisition and inhibition of conditioned fear. *Cerebral Cortex (New York, N.Y. : 1991)*, *21*(9), 1954–62. http://doi.org/10.1093/cercor/bhq253

Hendershot, C. S., Witkiewitz, K., George, W. H., & Marlatt, G. A. (2011). Relapse prevention for addictive behaviors. *Substance Abuse Treatment, Prevention, and Policy*, *6*(1), 17. http://doi.org/10.1186/1747-597X-6-17

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

Hogarth, L. C., Mogg, K., Bradley, B. P., Duka, T., & Dickinson, A. (2003). Attentional orienting towards smoking-related stimuli. *Behavioural Pharmacology*, *14*(2),

153–60. http://doi.org/10.1097/01.fbp.0000063527.83818.9e

Hogarth, L., & Chase, H. W. (2011). Parallel goal-directed and habitual control of human drug-seeking: implications for dependence vulnerability. *Journal of Experimental Psychology. Animal Behavior Processes*, *37*(3), 261–76. http://doi.org/10.1037/a0022913

Hogarth, L., Dickinson, A., Austin, A., Brown, C., & Duka, T. (2008). Attention and expectation in human predictive learning: The role of uncertainty. *Quarterly Journal of Experimental Psychology*, *61*(11), 1658–1668. http://doi.org/10.1080/17470210701643439

Hogarth, L., Dickinson, A., & Duka, T. (2005). Explicit knowledge of stimulus?outcome contingencies and stimulus control of selective attention and instrumental action in human smoking behaviour. *Psychopharmacology*, *177*(4), 428–437. http://doi.org/10.1007/s00213-004-1973-9

Hogarth, L., Dickinson, A., & Duka, T. (2010). The associative basis of cue-elicited drug taking in humans. *Psychopharmacology*, *208*(3), 337–51. http://doi.org/10.1007/s00213-009-1735-9

Hogarth, L., Dickinson, A., Hutton, S. B., Bamborough, H., & Duka, T. (2006). Contingency knowledge is necessary for learned motivated behaviour in humans: relevance for addictive behaviour. *Addiction (Abingdon, England)*, *101*(8), 1153–66. http://doi.org/10.1111/j.1360-0443.2006.01459.x

Hogarth, L., Dickinson, A., Hutton, S. B., Elbers, N., & Duka, T. (2006). Drug expectancy is necessary for stimulus control of human attention, instrumental drug-seeking behaviour and subjective pleasure. *Psychopharmacology*, *185*(4), 495–504. http://doi.org/10.1007/s00213-005-0287-x

Hogarth, L., Dickinson, A., Wright, A., Kouvaraki, M., & Duka, T. (2007). The role of drug expectancy in the control of human drug seeking. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(4), 484–496. http://doi.org/10.1037/0097-7403.33.4.484

Hogarth, L., & Duka, T. (2006). Human nicotine conditioning requires explicit contingency knowledge: is addictive behaviour cognitively mediated? *Psychopharmacology*, *184*(3–4), 553–66. http://doi.org/10.1007/s00213-005-0150-0

Högden, F., Hütter, M., & Unkelbach, C. (2018). Does evaluative conditioning depend on awareness? Evidence from a continuous flash suppression paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://doi.org/10.1037/xlm0000533

Houben, K., & Wiers, R. W. (2006). Assessing implicit alcohol associations with the Implicit Association Test: Fact or artifact? *Addictive Behaviors*, *31*(8), 1346–1362. http://doi.org/10.1016/J.ADDBEH.2005.10.009

Hull, C. L. (1943). Principles of Behavior. An Introduction to Behavior Theory. *The Journal of Philosophy*, *40*(20), 558. http://doi.org/10.2307/2019960

Hutter, M., Sweldens, S., Stahl, C., Unkelbach, C., & Klauer, K. C. (2012). Dissociating Contingency Awareness and Conditioned Attitudes: Evidence of Contingency-Unaware Evaluative Conditioning. *Journal of Experimental Psychology: General*, *141*(3), 539–557.

James, W. (1884). II.—What is an emotion? *Mind*.

Janes, A. C., Pizzagalli, D. A., Richardt, S., Frederick, B. de B., Holmes, A. J., Sousa, J., … Kaufman, M. J. (2010). Neural substrates of attentional bias for smoking-related cues: an FMRI study. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *35*(12), 2339–45. http://doi.org/10.1038/npp.2010.103

Jeffs, S., & Duka, T. (2017). Predictive but not emotional value of Pavlovian stimuli leads to pavlovian-to-instrumental transfer. *Behavioural Brain Research*, *321*, 214–222. http://doi.org/10.1016/j.bbr.2016.12.022

Jellinek, E. M. (1955). The craving for alcohol. *Quarterly Journal of Studies on Alcohol*, *16*(1), 35–8.

Kandasamy, N., Garfinkel, S. N., Page, L., Hardy, B., Critchley, H. D., Gurnell, M., & Coates, J. M. (2016). Interoceptive Ability Predicts Survival on a London Trading Floor. http://doi.org/10.1038/srep32986

Kang, O.-S., Chang, D.-S., Jahng, G.-H., Kim, S.-Y., Kim, H., Kim, J.-W., … Chae, Y. (2012). Individual differences in smoking-related cue reactivity in smokers: An eye-tracking and fMRI study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *38*(2), 285–293. http://doi.org/10.1016/J.PNPBP.2012.04.013

Katkin, E. S., Blascovich, J., & Goldband, S. (1981). Empirical assessment of visceral self-perception: Individual and sex differences in the acquisition of heartbeat discrimination. *Journal of Personality and Social Psychology*, *40*(6), 1095–1101. http://doi.org/10.1037/0022-3514.40.6.1095

Katkin, E. S., Wiens, S., & Ohman, A. (2001). Nonconscious Fear Conditioning, Visceral Perception, and the Development of Gut Feelings. *Psychological Science*, *12*(5), 366–370. http://doi.org/10.1111/1467-9280.00368

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., … Zucker, N. (2017). Interoception and Mental Health: A Roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. http://doi.org/10.1016/J.BPSC.2017.12.004

Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of addiction. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 217–38. http://doi.org/10.1038/npp.2009.110

Koob, G. F., & Volkow, N. D. (2016). Neurobiology of addiction: a neurocircuitry analysis. *The Lancet Psychiatry*, *3*(8), 760–773. http://doi.org/10.1016/S2215-0366(16)00104-8

Krebs, R. M., Boehler, C. N., & Woldorff, M. G. (2010). The influence of reward associations on conflict processing in the Stroop task. *Cognition*, *117*(3), 341–347.

http://doi.org/10.1016/j.cognition.2010.08.018

Lachnit, H., & Kimmel, H. D. (1993). Positive and negative patterning in human classical skin conductance response conditioning. *Animal Learning & Behavior*, *21*(4), 314–326. http://doi.org/10.3758/BF03197997

Lê, A. D., Quan, B., Juzytch, W., Fletcher, P. J., Joharchi, N., & Shaham, Y. (1998). Reinstatement of alcohol-seeking by priming injections of alcohol and exposure to stress in rats. *Psychopharmacology*, *135*(2), 169–174. http://doi.org/10.1007/s002130050498

Le Pelley, M. E., Seabrooke, T., Kennedy, B. L., Pearson, D., & Most, S. B. (2017). Miss it and miss out: Counterproductive nonspatial attentional capture by task-irrelevant, value-related stimuli. *Attention, Perception, & Psychophysics*, *79*(6), 1628–1642. http://doi.org/10.3758/s13414-017-1346-1

Levey, A. B., & Martin, I. (1975). Classical conditioning of human 'evaluative' responses. *Behaviour Research and Therapy*, *13*(4), 221–226. http://doi.org/10.1016/0005-7967(75)90026-1

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26.

Lubman, D. I., Peters, L. A., Mogg, K., Bradley, B. P., & Deakin, J. F. (2000). Attentional bias for drug cues in opiate dependence. *Psychological Medicine*, *30*(1), 169–75.

MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology*. US: American Psychological Association. http://doi.org/10.1037/0021-843X.95.1.15

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d', Response-Specific Meta-d', and the Unequal Variance SDT Model. In *The Cognitive Neuroscience of Metacognition* (pp. 25–66). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-45190-4_3

Mathew, R. J., Claghorn, J. L., & Largen, J. (1979). Craving for alcohol in sober alcoholics. *The American Journal of Psychiatry*, *136*(4B), 603–6.

McHugo, M., Olatunji, B. O., & Zald, D. H. (2013). The emotional attentional blink: what we know so far. *Frontiers in Human Neuroscience*, *7*, 151. http://doi.org/10.3389/fnhum.2013.00151

Metcalfe, J. (1996). Metacognitive Processes. *Memory*, 381–407. http://doi.org/10.1016/B978-012102570-0/50013-6

Miles, F. J., Everitt, B. J., & Dickinson, A. (2003). Oral cocaine seeking by rats: action or habit? *Behavioral Neuroscience*, *117*(5), 927–38. http://doi.org/10.1037/0735-7044.117.5.927

Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the Implicit Association Test. *Learning and Motivation*, *34*(2), 203–217. http://doi.org/10.1016/S0023-9690(03)00003-1

Mogg, K., Bradley, B. P., Field, M., & De Houwer, J. (2003). Eye movements to smoking-related pictures in smokers: relationship between attentional biases and implicit and explicit measures of stimulus valence. *Addiction*, *98*(6), 825–836. http://doi.org/10.1046/j.1360-0443.2003.00392.x

Naqvi, N. H., & Bechara, A. (2010). The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. *Brain Structure & Function*, *214*(5–6), 435–50. http://doi.org/10.1007/s00429-010-0268-7

Naqvi, N. H., Rudrauf, D., Damasio, H., & Bechara, A. (2007). Damage to the insula disrupts addiction to cigarette smoking. *Science (New York, N.Y.)*, *315*(5811), 531–4. http://doi.org/10.1126/science.1135926

Nikolaou, K., Field, M., & Duka, T. (2013). Alcohol-related cues reduce cognitive control in social drinkers. *Behavioural Pharmacology*, *24*(1), 29–36. http://doi.org/10.1097/FBP.0b013e32835cf458

Paulus, M. P. (2007). Neural basis of reward and craving--a homeostatic point of view. *Dialogues in Clinical Neuroscience*, *9*(4), 379–87.

Paulus, M. P., & Stewart, J. L. (2014). Interoception and drug addiction. *Neuropharmacology*, *76 Pt B*, 342–50. http://doi.org/10.1016/j.neuropharm.2013.07.002

Paulus, M. P., Tapert, S. F., & Schulteis, G. (2009). The role of interoception and alliesthesia in addiction. *Pharmacology, Biochemistry, and Behavior*, *94*(1), 1–7. http://doi.org/10.1016/j.pbb.2009.08.005

Peeters, M., Wiers, R. W., Monshouwer, K., van de Schoot, R., Janssen, T., & Vollebergh, W. A. M. (2012). Automatic processes in at-risk adolescents: the role of alcohol-approach tendencies and response inhibition in drinking behavior. *Addiction*, *107*(11), 1939–1946. http://doi.org/10.1111/j.1360-0443.2012.03948.x

Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. *The Pavlovian Journal of Biological Science : Official Journal of the Pavlovian*, *20*(4), 163–170. http://doi.org/10.1007/bf03003653

Pfeifer, G., Garfinkel, S. N., Gould van Praag, C. D., Sahota, K., Betka, S., & Critchley, H. D. (2017). Feedback from the heart: Emotional learning and memory is controlled by cardiac cycle, interoceptive accuracy and personality. *Biological Psychology*, *126*, 19–29. http://doi.org/10.1016/j.biopsycho.2017.04.001

Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis)liking: item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(1), 130–44. http://doi.org/10.1037/0278-7393.33.1.130

Pollatos, O., Herbert, B. M., Matthias, E., & Schandry, R. (2007). Heart rate response after emotional picture presentation is modulated by interoceptive awareness. *International Journal of Psychophysiology*, *63*(1), 117–124. http://doi.org/10.1016/J.IJPSYCHO.2006.09.003

Pollatos, O., & Schandry, R. (2008). Emotional processing and emotional memory are modulated by interoceptive awareness. *Cognition & Emotion*, *22*(2), 272–287. http://doi.org/10.1080/02699930701357535

Pool, E., Sennwald, V., Delplanque, S., Brosch, T., & Sander, D. (2016). Measuring wanting and liking from animals to humans: A systematic review. *Neuroscience and Biobehavioral Reviews*. http://doi.org/10.1016/j.neubiorev.2016.01.006

Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., … Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychological Assessment*, *27*(2), 365–76. http://doi.org/10.1037/pas0000036

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849–860. http://doi.org/10.1037/0096-1523.18.3.849

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863. http://doi.org/10.1016/S0022-5371(67)80149-X

Reber, A. S., & S., A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. http://doi.org/10.1037/0096-3445.118.3.219

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, *18*(3), 247–291. http://doi.org/10.1016/0165-0173(93)90013-P

Robinson, T. E., & Berridge, K. C. (2001). Incentive-sensitization and addiction. *Addiction (Abingdon, England)*, *96*(1), 103–14. http://doi.org/10.1080/09652140020016996

Robinson, T. E., & Berridge, K. C. (2003). Addiction. *Annual Review of Psychology*, *54*, 25–53. http://doi.org/10.1146/annurev.psych.54.101601.145237

Schandry, R. (1981). Heart Beat Perception and Emotional Experience. *Psychophysiology*, *18*(4), 483–488. http://doi.org/10.1111/j.1469-8986.1981.tb02486.x

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*, *19*(7), 595–605. http://doi.org/10.1002/per.554

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology*, *60*(3), 291–309. http://doi.org/10.1080/17470210601000581

Shanks, D. R. (2016). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. http://doi.org/10.3758/s13423-016-1170-y

Shanks, D. R., & Dickinson, A. (1988). Associative Accounts of Causality Judgment.

Psychology of Learning and Motivation, *21*, 229–261. http://doi.org/10.1016/S0079-7421(08)60030-4

Shanks, D. R., & St John, M. F. (1994). Characteristics of dissociable human learning systems. *BEHAVIORAL AND BRAIN SCIENCES*, *17*, 367–447. http://doi.org/10.1017/S0140525X00035032

Sherrington, C. S. (1948). The integrative action of the nervous system, (Cambridge Univ. Pess, Cambridge, UK).

Siegel, S. (1975). Evidence from rats that morphine tolerance is a learned response. *Journal of Comparative and Physiological Psychology*, *89*(5), 498–506.

Sienkiewicz-Jarosz, H., Scinska, A., Swiecicki, L., Lipczynska-Lojkowska, W., Kuran, W., Ryglewicz, D., … Bienkowski, P. (2013). Sweet liking in patients with Parkinson's disease. *Journal of the Neurological Sciences*, *329*(1–2), 17–22. http://doi.org/10.1016/j.jns.2013.03.005

Skinner, B. (2011). *About behaviorism*. Vintage.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*. US: American Psychological Association. http://doi.org/10.1037/h0054367

Smith, S. D., Most, S. B., Newsome, L. A., & Zald, D. H. (2006). An emotion-induced attentional blink elicited by aversively conditioned stimuli. *Emotion (Washington, D.C.)*, *6*(3), 523–7. http://doi.org/10.1037/1528-3542.6.3.523

Sokol-Hessner, P., Hartley, C. A., Hamilton, J. R., & Phelps, E. A. (2015). Interoceptive ability predicts aversion to losses. *Cognition and Emotion*, *29*(4), 695–701. http://doi.org/10.1080/02699931.2014.925426

Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., & Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(13), 5035–40. http://doi.org/10.1073/pnas.0806761106

Stacy, A. W., & Wiers, R. W. (2010). Implicit cognition and addiction: a tool for explaining paradoxical behavior. *Annual Review of Clinical Psychology*, *6*(1), 551–75. http://doi.org/10.1146/annurev.clinpsy.121208.131444

Stahl, C., Haaf, J., & Corneille, O. (2016). Subliminal evaluative conditioning? Above-chance CS identification may be necessary and insufficient for attitude learning. *Journal of Experimental Psychology: General*, *145*(9), 1107–1131. http://doi.org/10.1037/xge0000191

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, &*.

Stewart, J., de Wit, H., & Eikelboom, R. (1984). Role of unconditioned and conditioned drug effects in the self-administration of opiates and stimulants. *Psychological Review*, *91*(2), 251–268. http://doi.org/10.1037/0033-295X.91.2.251

Stewart, J. L., May, A. C., Tapert, S. F., & Paulus, M. P. (2015). Hyperactivation to pleasant interoceptive stimuli characterizes the transition to stimulant addiction.

Drug and Alcohol Dependence, *154*, 264–270. http://doi.org/10.1016/J.DRUGALCDEP.2015.07.009

Stroop. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6).

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i-109. http://doi.org/10.1037/h0092987

Thorndike, E. L. (1911). Individuality. *Houghton, Mifflin*.

Thush, C., Wiers, R. W., Ames, S. L., Grenard, J. L., Sussman, S., & Stacy, A. W. (2008). Interactions between implicit and explicit cognition and working memory capacity in the prediction of alcohol use in at-risk adolescents. *Drug and Alcohol Dependence*, *94*(1–3), 116–24. http://doi.org/10.1016/j.drugalcdep.2007.10.019

Tibboel, H., De Houwer, J., Spruyt, A., Field, M., Kemps, E., & Crombez, G. (2011). Testing the validity of implicit measures of wanting and liking. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(3), 284–92. http://doi.org/10.1016/j.jbtep.2011.01.002

Tibboel, H., De Houwer, J., & Van Bockstaele, B. (2015). Implicit measures of "wanting" and "liking" in humans. *Neuroscience and Biobehavioral Reviews*. http://doi.org/10.1016/j.neubiorev.2015.09.015

Tiffany, S. T. (1990). A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes. *Psychological Review*, *97*(2), 147–68. http://doi.org/10.1037/0033-295X.97.2.147

Tiffany, S. T., Warthen, M. W., & Goedeker, K. C. (2009). The functional significance of craving in nicotine dependence. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, *55*, 171–97.

Toates, F. M. (Frederick M. . (1986). *Motivational systems*. Cambridge University Press.

Townshend, J. M., & Duka, T. (2001). Attentional bias associated with alcohol cues: differences between heavy and occasional social drinkers. *Psychopharmacology*, *157*(1), 67–74. http://doi.org/10.1007/S002130100764

Tsakiris, M., & Critchley, H. (2016). Interoception beyond homeostasis: affect, cognition and mental health. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708), 20160002. http://doi.org/10.1098/rstb.2016.0002

Verdejo-García, A., Bechara, A., Recknor, E. C., & Pérez-García, M. (2006). Executive dysfunction in substance dependent individuals during drug use and abstinence: An examination of the behavioral, cognitive and emotional correlates of addiction. *Journal of the International Neuropsychological Society*, *12*(03), 405–415. http://doi.org/10.1017/S1355617706060486

Verdejo-Garcia, A., Clark, L., & Dunn, B. D. (2012). The role of interoception in addiction: A critical review. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1857–

1869. http://doi.org/10.1016/j.neubiorev.2012.05.007

Volkow, N. D., Wang, G.-J., Fowler, J. S., Logan, J., Jayne, M., Franceschi, D., … Pappas, N. (2002). "Nonhedonic" food motivation in humans involves dopamine in the dorsal striatum and methylphenidate amplifies this effect. *Synapse*, *44*(3), 175–180. http://doi.org/10.1002/syn.10075

Waters, A. J., Heishman, S. J., Lerman, C., & Pickworth, W. (2007). Enhanced identification of smoking-related words during the attentional blink in smokers. *Addictive Behaviors*, *32*(12), 3077–82. http://doi.org/10.1016/j.addbeh.2007.05.016

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158–177. http://doi.org/10.1037/h0074428

Weingarten, H. P. (1983). Conditioned cues elicit feeding in sated rats: A role for learning in meal initiation. *Science*, *220*(4595), 431–433. http://doi.org/10.1126/science.6836286

Whitehead, W. E., Drescher, V. M., Heiman, P., & Blackwell, B. (1977). Relation of heart rate control to heartbeat perception. *Biofeedback and Self-Regulation*, *2*(4), 371–392. http://doi.org/10.1007/BF00998623

Wiens, S. (2005). Interoception in emotional experience. *Current Opinion in Neurology*, *18*(4), 442–7.

Wiens, S., Mezzacappa, E. S., & Katkin, E. S. (2000). Heartbeat detection and the experience of emotions. *Cognition & Emotion*, *14*(3), 417–427. http://doi.org/10.1080/026999300378905

Wiers, R. W., Gladwin, T. E., Hofmann, W., Salemink, E., & Ridderinkhof, K. R. (2013). Cognitive Bias Modification and Cognitive Control Training in Addiction and Related Psychopathology. *Clinical Psychological Science*, *1*(2), 192–212. http://doi.org/10.1177/2167702612466547

Wiers, R. W., & Stacy, A. W. (2006). Implicit Cognition and Addiction. *Current Directions in Psychological Science*, *15*(6), 292–296. http://doi.org/10.1111/j.1467-8721.2006.00455.x

Wiers, R. W., Stacy, A. W., Ames, S. L., Noll, J. A., Sayette, M. A., Zack, M., & Krank, M. (2002). Implicit and Explicit Alcohol-Related Cognitions. *Alcoholism: Clinical and Experimental Research*, *26*(1), 129–137. http://doi.org/10.1111/j.1530-0277.2002.tb02441.x

Wikler, A. (1948). Recent progress in research on the neurophysiologic basis of morphine addiction. *American Journal of Psychiatry*, *105*(5), 329–338. http://doi.org/10.1176/ajp.105.5.329

Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality & Social Psychology Bulletin*, *31*(1), 121–35. http://doi.org/10.1177/0146167204271309

Winkler, M. H., Weyers, P., Mucha, R. F., Stippekohl, B., Stark, R., & Pauli, P. (2011). Conditioned cues for smoking elicit preparatory responses in healthy smokers. *Psychopharmacology*, *213*(4), 781–789. http://doi.org/10.1007/s00213-010-2033-2

Wise, R. A. (1988). The neurobiology of craving: implications for the understanding and treatment of addiction. *Journal of Abnormal Psychology*, *97*(2), 118–32.

Wise, R. A., & Koob, G. F. (2014). The development and maintenance of drug addiction. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *39*(2), 254–62. http://doi.org/10.1038/npp.2013.261

Wyvell, C. L., & Berridge, K. C. (2000). Intra-Accumbens Amphetamine Increases the Conditioned Incentive Salience of Sucrose Reward: Enhancement of Reward "Wanting" without Enhanced "Liking" or Response Reinforcement. *J. Neurosci.*, *20*(21), 8122–8130.

Yokoyama, T., Padmala, S., & Pessoa, L. (2015). Reward learning and negative emotion during rapid attentional competition. *Frontiers in Psychology*, *6*, 269. http://doi.org/10.3389/fpsyg.2015.00269

Zaki, J., Davis, J. I., & Ochsner, K. N. (2012). Overlapping activity in anterior insula during interoception and emotional experience. *NeuroImage*, *62*(1), 493–499. http://doi.org/10.1016/j.neuroimage.2012.05.012

Zerhouni, O., Bègue, L., Comiran, F., & Wiers, R. W. (2018). Controlled and implicit processes in evaluative conditioning on implicit and explicit attitudes toward alcohol and intentions to drink. *Addictive Behaviors*, *76*, 335–342. http://doi.org/10.1016/J.ADDBEH.2017.08.026

2 Attentional responses to stimuli associated with a reward can occur in the absence of knowledge of their predictive values

2.1 Abstract

Classical conditioning theories of addiction suggest that stimuli associated with rewards acquire incentive salience, inducing emotional and attentional conditioned responses. It is not clear whether those responses occur without contingency awareness (CA), i.e. are based on explicit or implicit learning processes. Examining implicit aspects of stimulus-reward associations can improve our understanding of addictive behaviours, supporting treatment and prevention strategies. However, the acquisition of conditioned responses without CA has yet to be rigorously demonstrated, as the existing literature shows a lack of methodological agreement regarding the measurement of implicit and explicit processes.

The purpose of two experiments presented here was to study the emotional value acquired by CS through implicit emotional and attentional processes, trying to overcome critical methodological issues.

Experiment 1 (n = 48) paired two stimuli categories (houses/buildings) with high (HR) or low (LR) probabilities of monetary reward. An Emotional Attentional Blink revealed preferential attention for HR over LR regardless of CA; while pleasantness ratings were unaffected, probably due to the intrinsic nature of CS.

Experiment 2 (n = 60) replicated the effect of conditioning on the Emotional Attentional Blink utilising abstract CS (octagons/squares). In addition, increased pleasantness for HR over LR was found significant overall, and marginally significant for Aware but not for Unaware participants. Here CA was rigorously determined using a signal-detection analysis and metacognitive-awareness measurements. Bayesian analyses verified the unconscious nature of the learning.

These findings demonstrate that attentional conditioned responses can occur without CA and advance our understanding of the mechanisms by which implicit conditioning can occur and becomes observable. Furthermore, these results can highlight how addictive behaviours might develop.

2.2 Introduction

Processes related to Classical conditioning have been proven to determine the development of addictive behaviours. Stimuli conditioned (CS) with addictive drugs acquire hedonic and reinforcing properties associated with the substance [1,2], ultimately driving and maintaining drug-seeking behaviours. As part of this process, drug-related cues generate preferential attention and emotional ratings in heavy drinkers of alcohol, cocaine addicts [3] and smokers [4,5]. Most of the conditioned responses occur with subjects' awareness [6], as substance expectancy generated by drug-paired stimuli is responsible for attentional, instrumental and hedonic conditioned responses [7].

However, conditioned responses may also occur without subjects' awareness, and could be studied from an implicit point of view.

The study of implicit processes within addiction has gained increasing relevance, as pointed by Wiers and Stacy [8], leading to dual process theories of addiction. Dual process theories of addiction conceptualize addictive behaviours as the combination of automatic appetitive appraisals generated by associations [9] in opposition to regulatory executive signals based on propositional knowledge [10]. However, the model presented by Wiers and Stacy does not specify the nature (implicit and/or explicit) of the associations generating automatic responses, obviating the role implicit learning may have in the development of automatic responses towards drug related stimuli.

Determining the extent to which implicitly generated associations can induce conditioned responses without awareness could provide a better understanding of addictive behaviours.

Implicit processes in addiction are most commonly studied via attentional bias measurements, memory associations or action tendencies [8] using naturalistic stimuli, materials conceptually related to a substance. The use of naturalistic stimuli we think is a limitation, as the explicit conceptual representation of the substance is necessarily bounded to drug-related stimuli, compromising the dissociation between implicit and explicit processes even in automatic detection tasks.

Neutral cues paired with alcoholic drinks or tobacco can also generate attentional and autonomous reactions through classical conditioning [11,12]. Furthermore, stimuli associated with non-drug rewards can also be conditioned to generate incentive responses equivalent to those elicited by drug-related stimuli [13]. Given that CSs in

this case are originally devoid of any motivational attributes, conditioning paradigms can provide an opportunity to investigate the development of implicit (as well as explicit) processes through learned associations between such stimuli and rewards in the laboratory.

Within the Evaluative Conditioning (EC) paradigm, the modification or generation of emotional responses towards cues paired with positively or negatively valenced stimuli [14], has led to confronting viewpoints about the necessity of learning to be conscious in order to elicit measurable responses. A meta-analysis [15] showed that pleasantness can occur without Contingency Awareness (CA), that is, without conscious knowledge, i.e. knowledge that the neutral CS had been associated with a highly emotional stimulus. However, opposing views are still prevalent [16,17]. Whether conditioning can occur without CA has generated a discussion regarding the methods most appropriate to measure implicit [18], or explicit knowledge about contingencies [19].

Furthermore, implicit learning in Pavlovian conditioning tasks is most commonly demonstrated using direct self-report measurements of liking [20,21]. This type of assessment is based on Likert or Visual Analog scales in which participants evaluate the pleasantness of a stimulus. De Houwer [22] however, has advocated for the need to study emotional reactions in the context of Implicit learning through indirect measures of automatic behaviour.

Attentional processes are strongly affected by the emotional salience of stimuli (see Ref. [23] for a review), and attentional correlates of conditioned stimuli have been employed in Pavlovian conditioning procedures. Hogarth et al., has demonstrated

attentional orientation to stimuli associated with a monetary reward versus non-reward using eye tracking [7], whereas Anderson, Laurent and Yantis [24] showed that stimuli paired with high-reward (HR) versus low reward (LR) probabilities were more distracting on a visual search task. Automatic attention allocation towards CS has also been demonstrated using rapid serial visual presentation tasks (RSVP). In this kind of task, also known as Attentional Blink [25], a stream of pictures is presented, and participants have to detect a target embedded within the stream. Before the target, a distractor is also presented, affecting the accuracy on detection of the target. Emotional Attentional Blink (EAB) tasks [26] employ emotional stimuli as distractors during the RSVP.

Conditioned stimuli have been used during RSVP tasks both as distractors during an aversive conditioning task [27], and as targets during a task irrelevant Pavlovian conditioning procedure [28], providing evidence towards the ability of CS to capture attention. However, no measures of CA (see below) were included in this task. Thus, it cannot be excluded that the ability of the CS to guide attention is based on explicit knowledge about stimulus-outcome contingencies, a matter that will be addressed in this paper.

Measurements of CA in PC are often based on post-hoc ratings, i.e. asking participants how stimuli and outcomes were related to each other during learning. Lovibond and Shanks [19], proposed as the most valid measures of CA to ask participants on a trial by trial basis to anticipate the outcome using Likert or Visual Analogue Scales in the presence of CS. Several studies have considered such criteria for measures of Expectancy Awareness (e.g. [21,29]) showing that pleasantness ratings for the CS can

occur in the absence of CA. Others have appealed to retrospective measurements when measuring CA [30]. These differences in terms of CA measurement could in turn explain the inconsistencies found in experimental literature related to Implicit Conditioning [16].

Prior literature in the field of Implicit learning has primarily focused on artificial grammar-learning tasks [31,32] and sequence learning [33], but few studies have used self-reports in conditioning experiments [34,35]. In these latter studies, however, no implicit conditioned responses were found.

Measures of CA and Implicit learning as previously described respond to the criteria set out by classical implicit learning theories [32] in which objective measurements (i.e. accuracy on a detection task) are combined with subjective evaluations (i.e. ability to report a rule). However, a third layer of measurement can be implemented, testing whether participants have developed Metacognitive awareness about their knowledge of the set of rules underlying the procedure, that is, whether they can explicitly report those rules [36,37].

To our knowledge, there are no studies which have successfully incorporated the criteria set out by Dienes and Perner [36] for a distinction of explicit from implicit knowledge using Metacognitive measures (e.g. [38,39]) within a Pavlovian conditioning task.

The purpose of this paper was to examine if stimuli conditioned to a reward outcome would implicitly generate attentional and/or emotional conditioned responses. A task irrelevant conditioning task was used to limit the extent to which participants reached

CA. An EAB studied the ability of CS to overcome the effect of emotional distractors, therefore assessing their ability to generate preferential attention. In this kind of task, aversive distractors decrease the accuracy on detection of targets compared to neutral distractors. We hypothesize that detection of LR stimuli will decrease when distractors are aversive compared to when they are neutral, an effect that should be weaker for HR stimuli. This would show HR stimuli develop preferential attention as they are able to overcome the effect of negative distractors.

A novel approach for conditioning learning was employed in Experiment 2 following methods originating from the Implicit Learning literature [40], to classify participants in different groups according to their awareness level. Three levels of CA were hypothesized: complete Unawareness of Contingencies; Partial Awareness, being able to predict the nature of an outcome without explicit knowledge about it; and Metacognitive Awareness, in which explicit knowledge about the contingencies is developed.

2.3 Experiment 1

2.3.1 Aims

The aim of this first experiment was to examine the occurrence of implicitly conditioned responses. Particularly, we assessed whether awareness about contingencies between CS and its outcome is necessary to develop preferential attention towards stimuli predicting higher chances of reward. Another focus of the experiment was to investigate whether this preferential attention occurs together with a development of Pleasantness towards the aforementioned stimuli.

2.3.2 Methods

2.3.2.1 Participants

Forty-Eight University of Sussex students (28 females), mean age 22.7 years, were recruited via an online participant database and compensated for their time financially or with course credit. Participants gave written consent before beginning the study, with ethical approval being granted by the University of Sussex Life Sciences ethics committee. Inclusion criteria were that they were in a state of

good health, whereas exclusion criteria were that they were currently taking prescription medication (excluding the contraceptive pill) or reported having been diagnosed with a mental illness.

2.3.2.2 Apparatus and stimuli

During the conditioning phase, pictures depicting Buildings or Houses (36 of each category) were used as CSs throughout the experiment. Pictures were selected to be neutral with regard to pleasantness ratings of a pool of 50 House and 50 Building pictures before the experiment. An independent sample (n = 16) rated the pleasantness of the pictures via Likert scales (from 1 to 9). From each category, 14 pictures with the highest deviation in pleasantness ratings from the mean were excluded to generate a definitive final list of 72 stimuli as neutral as possible. No significant differences in Pleasantness between Houses (5.14, SD = .86) and Buildings (5.14, SD = .56), were found $p > .538$. Pleasantness towards the selected Houses and Buildings was measured also by a larger independent sample of participants (n = 40) to

confirm stimuli similarity. No significant differences were found between Houses (5.01, SD = .89) and Buildings (4.83, SD = .97) regarding pleasantness, $p > .3$.

During the Emotional Attentional Blink stimuli were used as targets among a series of picture fillers, composed by 4 × 5 matrices of Houses and Buildings (see Fig. 2 for an example). Distractors, aversive and neutral pictures, were obtained from the IAPS picture data base [41] with additional matched aversive pictures from the internet.

Stimuli were presented on a Dell ACPI 64-bit PC, screen refresh rate = 16.6 ms. Procedures were performed using E-prime 2. Data analysis was performed using SPSS and Matlab.

During the conditioning phase 10, 20 and 50 p coins were used as tangible reinforcers at the end of each block.

2.3.2.3 Procedure

Participants following informed consent completed questionnaires regarding their demographics and drug use, the AUQ [42] and AUDIT [43] questionnaires measuring alcohol consumption, the BIS-11 impulsivity questionnaire [44], the BIS BAS questionnaire on approach-avoidance behaviours [45] and the PANAS mood questionnaire [46]. Participants also took an Interoception Assessment involving Heartbeat measurements [47], not described in this paper. Afterwards, they completed Conditioning, Emotional Attentional Blink and Pleasantness measurement tasks.

The experiment lasted approximately 80 min.

2.3.2.4 Conditioning task

A task irrelevant Conditioning procedure was used to pair stimuli belonging to one of the categories (Houses vs. Buildings) with high (80%) or low (20%) probabilities of obtaining 10p [28]. For half of the participants, High Reward probability (HR) stimuli consisted of Houses and Low Reward probability (LR) stimuli consisted of Buildings, and vice versa.

CSs appeared on screen with an overlaid green or yellow square for 2000 ms or until a response was recorded (max recorded time 1499 ms). Participants were instructed to press a green or yellow key on the keyboard depending on the colour of the square. Participants were also informed that from time to time they would obtain money but were kept naïve about the nature of stimuli predicting reward or about the contingencies between stimuli and reward. As this conditioning procedure was task-irrelevant, the stimulus category (House or Building) was the only factor predictive of reward. Feedback was on screen for 1500 ms indicating whether the participant had obtained 10 p or nothing on that particular trial.

Following correct responses to the yellow or green key participants were asked to indicate on a Likert scale from 1 to 9 how likely they were to win 10 p whilst the stimulus remained on the screen (measurement of expectancy Awareness (EA)). Immediately after the response, they received feedback about the outcome of the trial (Fig. 1).

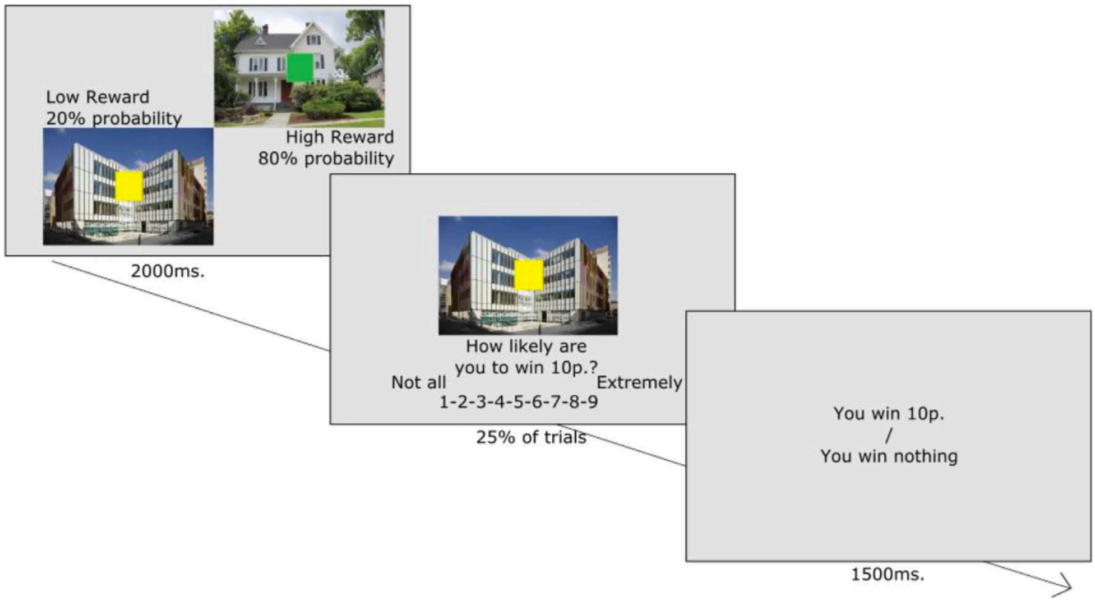


Figure 1: Conditioning task.

Each trial consisted of the presentation of a House or a Building with an overlaid coloured square. For half of the participants, Buildings were associated with High reward (80% probability of wining 10p) and Houses with Low reward (20% probability). For the other half of the sample, probabilities were inverted across stimuli categories. Participants had to press a key depending on the colour of the square (Green/Yellow) in the middle of the picture. Stimuli appeared on screen for 2000 ms or until a response was given. After their response participants were informed whether they had won 10p or not on that trial. On 25% of the trials, right after their response, participants were asked to indicate from 1 = "not at all" to 9 "extremely" how likely they were to win 10p (measurement of Expectancy Awareness). Immediately after Expectancy Awareness measurement, feedback about earnings on that trial appeared on screen.

Only in 25% of the Conditioning trials EA was measured as a means to avoid excessive priming towards contingency elaboration [19]; EA evaluation was pseudorandomized to occur every 3, 4 or 5 trials to prevent participants from establishing rules regarding its measurement. Trial order was pseudorandomized so the same kind of CS (HR/LR) or the same coloured square could not appear more than 4 times in a row.

In total 5 blocks of 72 trials were presented. At the end of each block the total amount earnt appeared on screen and participants had to grab the equivalent amount in coins and transfer it from a Bank box to their earnings box.

2.3.2.5 Emotional Attentional Blink

After the Conditioning task, participants took the Emotional Attentional Blink. The purpose of this task was to evaluate the ability of CS to overcome the influence of aversive distractors.

Each trial consisted of a RSVP of 17 stimuli. Fillers composed of jittered matrices of Houses and Buildings appeared at the beginning of the trial. Aversive or neutral distractors appeared on the 4th, 6th or 8th position of the series, followed by another filler and the presentation of the target, a HR or LR CS. Such a short lag between distractor and target was used as a means to increase the interference of distractors [48]. Finally, more fillers were presented to complete the stream of 17 images.

Participants were notified they would not be able to obtain money any longer and instructed to detect the presence in the stream of a House or a Building. At the end of each trial they had to press one of two keys depending on the category of the target detected.

The task started with a practice block of 12 trials in which each stimulus appeared on screen for 100 ms and feedback about accuracy was presented after each response.

The main task consisted of 3 blocks of 48 trials. Presentation time of each stimulus was 83 ms for participants with accuracy on detection above 75% on the practice block, and 100 ms for those less accurate (Fig. 2).

The amount of trials displaying aversive or neutral distractors was equally distributed among target type and no feedback was displayed during the task.

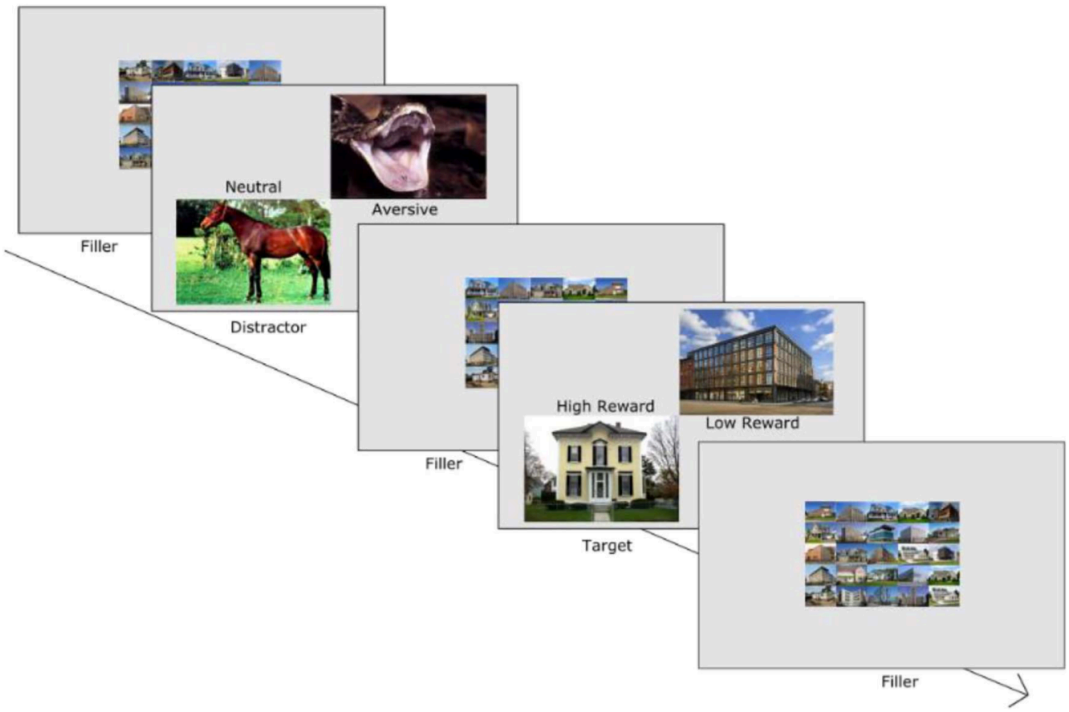


Figure 2: Emotional Attentional Blink.

Seventeen stimuli were presented in each trial on a RSVP stream for 83–100 ms each. The stimuli presented were a series of Fillers composed by jittered pictures of Houses and Buildings and an Aversive or Neutral distractor appeared on position 4, 6 or 8 of

the series. After the distractor, a Filler was presented followed by a target consisting of a House or a Building. Finally, more fillers appeared to complete the 17 stimuli sequence. Participants' task was to indicate the detection of a House or a Building by pressing one of two keys at the end of the trial.

2.3.2.6 Pleasantness measurement

At the end of the experiment CS pleasantness was measured via Likert scale. Eighteen House pictures and 18 Building pictures from the stimuli presented during conditioning appeared in random sequence on screen and participants had to indicate from 1 to 9 how pleasant each of them was. Each of the pictures remained on screen until response.

2.3.3 Data analysis

2.3.3.1 Expectancy evaluation.

First, participants were categorized as Aware or Unaware of the contingencies associated with CS depending on their responses to the EA assessment. One sample t-test comparisons were performed both for the 4th and 5th block of trials (9 ratings per stimulus per block). Participants were deemed to be Aware if their expectancy ratings were significantly above 5 for the HR and below 5 for the LR stimuli on one of these blocks of trials; 5 was the rating denoting "I don't know". On the basis of this approach only 4 participants were classified as aware. Therefore data only from the Unaware participants are presented. Data on pleasantness for Aware participants are presented in Appendix A.

2.3.3.2 Emotional Attentional Blink(EAB).

Accuracy on detection of targets was the dependent variable for the EAB analysis; a 2-way Repeated measures ANOVA was conducted with CS target (HR vs LR) and distractor type (aversive vs. neutral) as within subject factors for Unaware participants. Planned post-hoc t-tests will examine the hypothesized detrimental effect of aversive distractors on detection of LR and HR targets. Differences between HR and LR targets under baseline condition (neutral distractors) were also explored using paired t-test. Descriptives for Aware participants are included in Appendix A.

2.3.3.3 Pleasantness.

Paired samples t-tests compared pleasantness ratings towards HR and LR stimuli for Unaware participants. Descriptives for Aware participants are included in Appendix A.

2.3.3.4 Supplementary analyses.

RT towards HR and LR stimuli during conditioning were log transformed in order to compare them. Paired samples t-tests were performed for Unaware participants, see Appendix A.

2.3.4 Results

2.3.4.1 Outcome expectancy measurements

Out of 48 participants, only five met the two awareness criteria (Blocks 4&5, HR=5.19, SD=.78; LR=4.4, SD=1.24) and were classified as Aware. The rest, 43 participants, were considered Unaware of the contingencies associated with CS (Blocks 4&5, HR = 5.42, SD = 1.72; LR = 5.4, SD = 1.72).

2.3.4.2 Emotional Attentional Blink

There was a significant interaction between target and distractor type for Unaware participants, F(1,43) = 6.762, p = .013. Detection of LR stimuli was significantly lower when distractors were aversive compared with neutral, t(43) = 2.796, p = .008; this was not the case for HR stimuli. Only a marginal effect was found with higher detection for HR when distractors were negative compared to neutral, t (43) = 1.821, p = .076. There were no significant differences between HR and LR targets under neutral distractors, t(43) = 1.085 p = .284, Fig. 3.

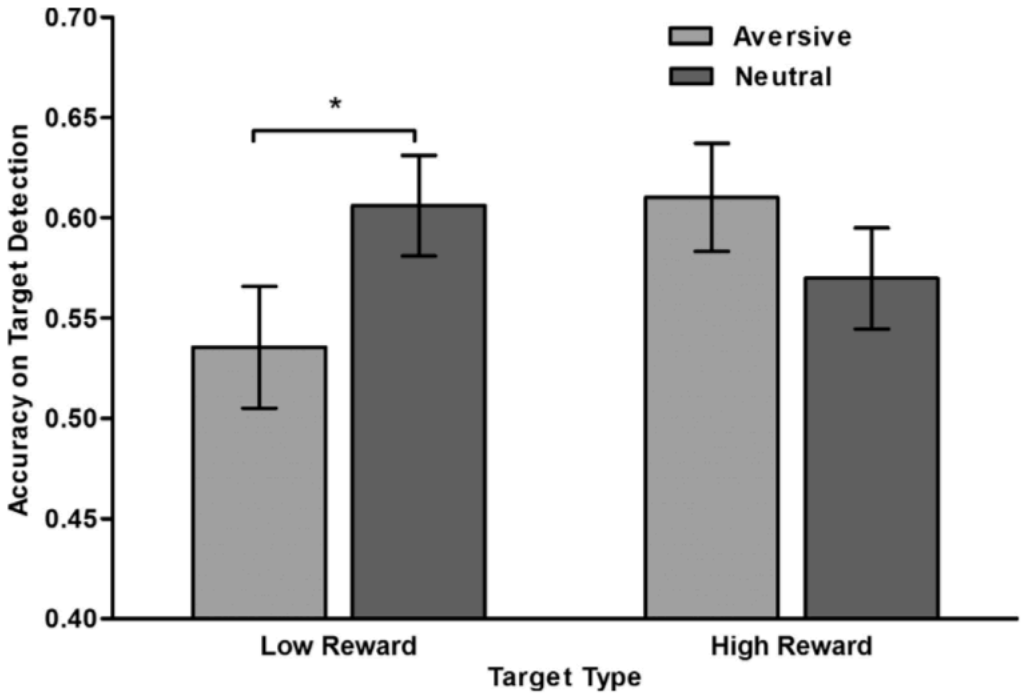


Figure 3: Accuracy of target detection (High reward or Low Reward) depending on Distractor type (Aversive or neutral), Experiment 1.

*Accuracy on target detection depending on reward associated and distractor type for Unaware participants. *Low reward targets under aversive distractors compared to neutral distractors, p < 0.01.*

2.3.4.3 Pleasantness measurement

No significant differences between HR and LR stimuli in terms of pleasantness t(43) = .273, p = .786, were found for Unaware participants, see Fig. 4.

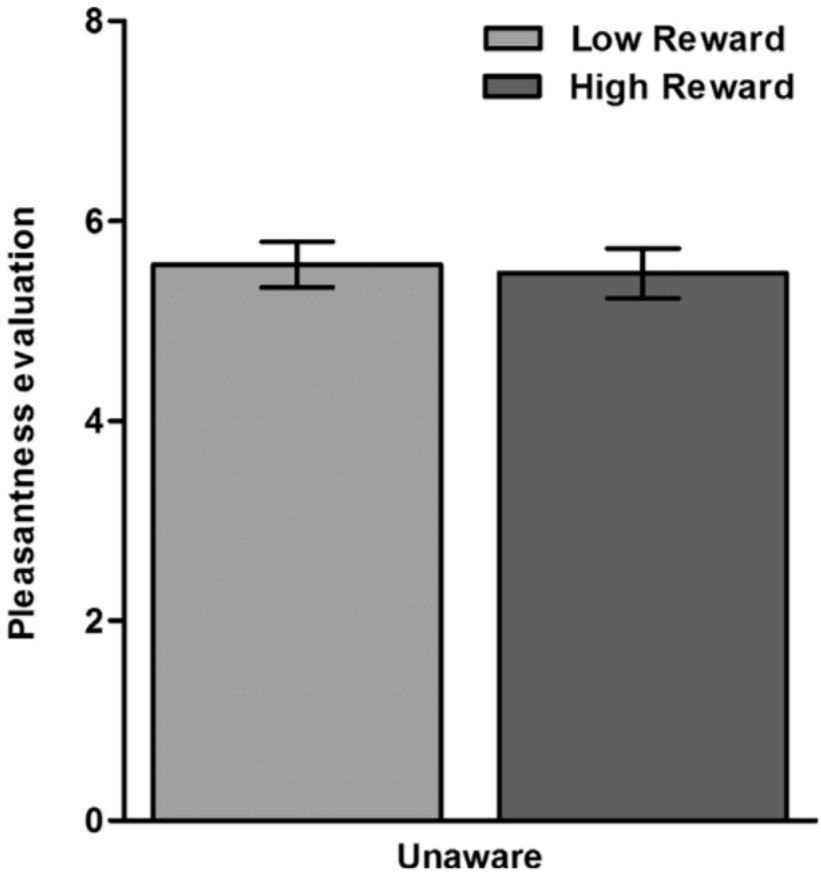


Figure 4: Pleasantness ratings depending on stimulus type for contingency Unaware participants, Experiment 1.

Pleasantness ratings towards High reward and Low reward stimuli for Unaware participants; there were no significant differences in either of the groups, ps > 0.7.

In order to understand the inability of HR stimuli to generate preferential pleasantness, an exploratory analysis examined pleasantness development depending on stimulus category (Houses vs. Buildings) for Unaware participants. A paired samples t-test found that Houses were evaluated as more pleasant than Buildings when they were HR stimuli, $t(20) = 2.687$, $p = .014$, but also when they were associated with LR, $t(22) = 2.915$, $p = .008$, see Fig. 5. This might be due to an effect of the intrinsic value of CS, which may be higher for stimuli more related to comfort (Houses), than work and business (Buildings). No differences in pleasantness ratings were found between the two types of stimuli outside the conditioning procedure (see Section 2.2.2 Apparatus and stimuli).

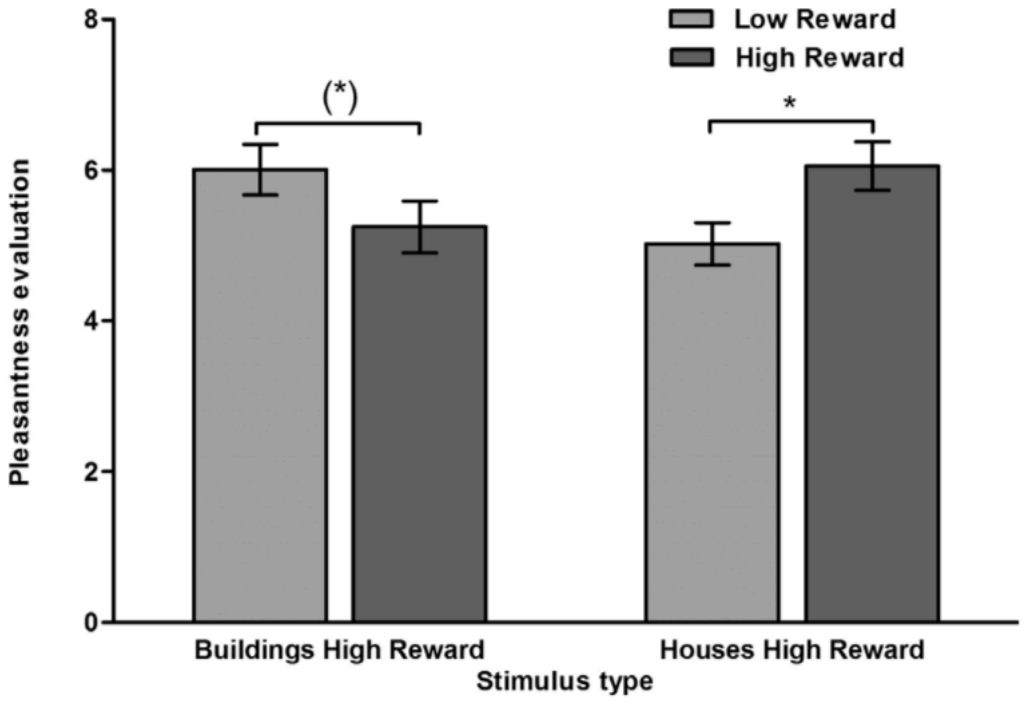


Figure 5: Pleasantness ratings per stimulus category for Unaware participants, Experiment 1.

*Pleasantness towards High Reward and Low reward CS depending on the stimulus (houses or buildings) associated with High reward. (*) Low reward compared to High reward when Buildings were High reward, p < .02; * High reward compared to Low reward when Houses were High reward, p < .01. Houses were overall more pleasant than Buildings (main effect of stimulus type).*

2.3.4.4 Questionnaires

There were no significant differences between the Contingency Aware and Unaware groups and questionnaire scores (see Table A.1 in Appendix A).

2.3.5 Supplementary analyses

No significant differences were found, see Appendix A.

2.3.6 Discussion Experiment 1

The present task irrelevant conditioning procedure induced expectancy awareness in only 5 out of 48 participants using probabilities of reward of 80% for HR and 20% for LR stimuli. As predicted, when measuring attentional preference towards HR CS compared to LR CS, HR stimuli were more resistant to the interference of aversive distractors than LR stimuli regardless of expectancy awareness. These results replicate those obtained by Yokoyama et al. [28], who did not, however, measure CA. These findings accurately show that CS acquire implicit incentive salience, which attracts

attention, thus providing evidence towards the ability of implicit processes to govern the development of conditioned responses.

Concerning pleasantness, neither Aware nor Unaware participants developed preferential subjective pleasantness towards HR stimuli. One reason that Unaware participants failed to develop heightened pleasantness towards HR stimuli compared to LR might be the nature of the stimuli used during the conditioning task. Despite independent measures of pleasantness revealing no differences between Houses and Buildings, after the conditioning task Houses were evaluated as more pleasant than Buildings overall. It is possible that the intrinsic preference for Houses as a safe and comfortable space opposed to Buildings as workspace, in conjunction with the monetary conditioning procedure, overrode the development of contingency congruent hedonic responses in Unaware participants.

Importantly the classification of Aware and Unaware participants did not take into account metacognitive awareness measures and the criteria used to separate Aware from Unaware participants may have been therefore not rigorous.

In order to tackle these issues, Experiment 2 was designed to replicate Experiment 1 using abstract geometric shapes devoid of any intrinsic positive meaning and confidence ratings of awareness were introduced to classify participants more rigorously as Aware and Unaware.

2.4 Experiment 2

2.4.1 Aims

The second Experiment addressed some of the limitations of Experiment 1, aiming at strengthening evidence towards the existence of implicit emotional and attentional conditioned responses.

By implementing abstract stimuli instead of Houses and Buildings, we tried to come up with a Conditioning procedure able to generate Pleasantness towards HR CS together with preferential attentional salience on participants Unaware of the contingencies.

In addition, we aimed to improve our classification of participants to different degrees of awareness by incorporating confidence ratings on each EA measurement. With this addition we hoped participants could be classified in different levels of CA, from unawareness to Metacognitive Contingency Awareness, fulfilling criteria established in the Implicit learning literature [49]; Bayesian factors were introduced to determine the presence of unconscious processes [50].

2.4.2 Methods

2.4.2.1 Participants

Sixty Sussex University Students (52 females, mean age=20.51, SD = 3.41) took part in the experiment. Participation conditions were identical to those of Experiment 1.

2.4.2.2 Apparatus and stimuli

Two types of abstract stimuli were used as CS in this experiment, Squares and Octagons. A set of 72 stimuli was developed using InkScape software. Stimuli consisted

of Octagons or Squares filled with parallel stripes. In order to generate different stimuli belonging to the same category (Squares or Octagons), 5 filling patterns were developed, differentiated in terms of stripe thickness. Then, each of the patterns was rotated multiple times, avoiding vertical and horizontal orientations as well as alignment with the edges of the figure contour [51]. This way, 36 Squares unique in terms of filling orientation and pattern and 36 matched Octagons were obtained. As this conditioning procedure was task-irrelevant, the stimulus category (Squares or Octagons) was the only factor predictive of reward.

The EAB fillers consisted of geometrical figures combining the contour of a Square and an Octagon, filled with the same patterns as CS. During all the procedures, geometrical shapes were presented superimposed on neutral landscape pictures to match the visual characteristics of aversive and neutral distractors. For that purpose, 15 neutral pictures were selected from the internet to compose the background on each presentation. For examples of fillers and conditioning stimuli see Fig. 6. Distractors consisted of the aversive and neutral stimuli as used in Experiment 1. The rest of the apparatus was identical to Experiment 1.

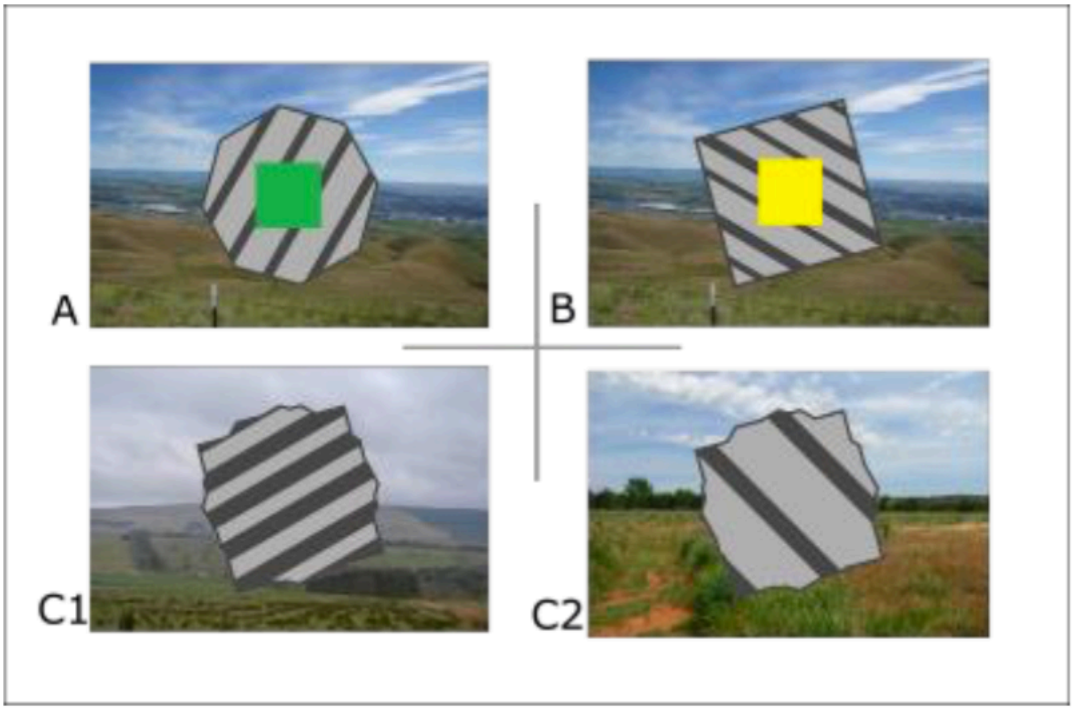


Figure 6: Stimuli used in Experiment 2

36 geometrical shapes representing an Octagon (A) and 36 representing a Square (B) were designed for this task, including different patterns as filler. 16 Geometrical shapes (C1 and C2) were designed as fillers for the Emotional Attentional Blink.

2.4.2.3 Procedure

The procedure was similar to Experiment 1, except for the fact that Pleasantness measurements took place after the Conditioning task, and an extra measurement of Post-hoc EA was included at the end of the experiment.

2.4.2.4 Conditioning task

A task irrelevant conditioning was employed as in Experiment 1, with Squares and Octagons as HR or LR CS. In order to increase the proportion of Aware participants, the

probability of obtaining 10p after a HR stimulus was raised to 90% and decreased to 10% for LR.

EA was measured via a dichotomous question ("Will you get money?" – Yes/No) on 25% of the trials. After their response participants had to indicate how confident they were in their judgment using a 1–5 Likert scale, (1. "completely guessing", 2. "more or less guessing", 3. "fairly sure", 4. "almost certain", 5. "completely certain"). The two different types of measurement, dichotomous and Likert, were employed to reduce interference between the two responses. The rest of the procedure was the same as in Experiment 1.

2.4.2.5 Pleasantness measurement and Emotional Attentional Blink

The procedure was equivalent to the one used in Experiment 1, this time using the set of stimuli described above. Pleasantness was measured before the EAB.

2.4.2.6 Post-hoc expectancy measurement

Expectancy was measured again at the end of the experiment using a 1–9 Likert scale to compare dichotomous online expectancy measurements with post-hoc assessments. Eighteen CS for each category were presented and participants asked to indicate how likely they thought they were to earn money after each of them. With this confirmatory analysis we aimed at reducing regression to the mean effects due to post-hoc categorizations [52].

2.4.3 Data analysis

2.4.3.1 Bayesian analysis.

A Bayesian analysis [53,54] allows determining the sensitivity of results obtained and extracting scientific conclusions out of non-significant results. A Bayes factor (B) below 1/3 provides substantial evidence for the null hypothesis (i.e. there is no difference between two means) and a B above 3 shows substantial evidence for the alternative hypothesis. Results between 1/3 and 3 indicate data are insensitive. These factors will be used throughout these analyses as the main source of CA categorization.

2.4.3.2 Contingency awareness.

Claiming that learning occurs implicitly requires accepting the null hypothesis that participants have not been able to perform above chance level on the task. Orthodox statistics based on p-values do not permit the validation of such claims [50]. Therefore, a Bayesian approach will be used to establish the existence of unconscious states [55].

In this experiment CA categorization was performed using Signal Detection Theory (SDT) methods [37,56]. In order to compute participants' accuracy taking into account response bias, log Type I d' (d1') scores for each participant were computed, using the number of individual Hits (H, answering Yes on a HR trial), Correct Rejections (CR, answering No on a LR trial), False Alarms (FA, answering Yes on a LR

trial) and Misses (M, answering No on a HR trial) [57]. Only results from blocks 4 and 5 were considered to account for the progressive development of learning.

In order to run a Bayesian analysis at an individual level for each participant, logistic d1' using Odds ratio (OR) Eq. (1) and Standard Errors (SE) Eq. (2) [58] were computed for each participant:

$$log_{d1'} = In(OR) * \frac{\sqrt{3}}{\pi} \quad (1)$$

$$SE_{d1'} = \sqrt{\frac{1}{H} + \frac{1}{CR} + \frac{1}{FA} + \frac{1}{M}} * \frac{\sqrt{3}}{\pi} \quad (2)$$

Type II d' (d2') scores [59] allow determining metacognitive knowledge using accuracy and confidence responses for each participant. Each of the confidence ratings was converted from Likert scales (1–5), to a dichotomous variable (confident/not confident). Responses equal to or below 2 ("more or less guessing"), were considered to be "low confidence", the rest of the responses were considered as "confident".

From a SDT point of view, for log d2' scores [37,54] H, FA, CR and M are computed as follows: accurate responses on expectancy discrimination (Type I Hits or Correct rejections) accompanied by a confident response are considered as Hits. Incorrect responses (Type I False Alarms or Misses) with high confidence as False Alarms. Correct Rejections are incorrect responses rated with low confidence, and Misses are accurate responses rated with low confidence.

Logistic d2' and SE d2' were obtained using the same method as previously described.

A Bayes factor was then computed for each participant on their log d2', modelling H1 with a Uniform going from 0 to their own log d1' as a maximum, given that d2' rarely exceeds d1' [50]. Participants with a B > 3 were categorized as metacognitively Aware,

those with B < 1/ 3 as metacognitively Unaware, and the rest had an undetermined metacognitive state.

The mean log d1's of metacognitively Aware participants was then used as the maximum for a Uniform to model H1 for testing each individuals d1's to determine their CA, [50]. The interpretation of individual Bs was then used to categorize them as contingency Aware, Unaware, or undetermined.

2.4.3.3 Post-hoc contingency measurement

Data extracted from Post-hoc contingency measurements was analysed, as in Experiment 1, performing independent samples t-tests on HR and LR stimuli compared to 5 (rating indicating “I don’t know”).

2.4.3.4 Emotional Attentional Blink

For Aware and Unaware participants, a 2-way Repeated measures ANOVA was conducted on accuracy to detect the targets, with CS targets (HR vs LR) and distractor type (aversive vs. neutral) as within factors. Sample sizes and variances differed between groups and therefore no between group comparisons were performed (Levene’s for HR neutral with distractors, $F(1,41) = 7.013$, $p = .011$, LR with negative distractors, $F(1,41)=7.181$, $p=.011$). We hypothesized the same results as in Experiment 1, the detection of LR targets would be affected by aversive distractors compared to neutral, and detection of HR would not be affected by distractor type. We also explored differences between HR and LR targets under baseline condition (neutral distractors).

2.4.3.5 Pleasantness

A paired samples t-test compared pleasantness towards HR and LR stimuli for Unaware and Aware participants separately. No between group comparisons were performed as sample sizes and variances were different between groups, Box M F (3,21531) = 7.071, $p < .001$.

2.4.3.6 Supplementary analyses

RT towards HR and LR stimuli during conditioning were log transformed in order to compare them. Paired samples t-tests were performed for Unaware and Aware participants. Accuracy towards HR and LR stimuli was compared within Aware and Unaware participants as well as Type 2 d' scores. Results are reported on Appendix B.

2.4.4 Results

2.4.4.1 Contingency awareness

Using the Bayesian approach for metacognitive CA using log d2' scores, 27 participants were deemed sensitively meta-Unaware, 30 didn't show any sensitive results ($3 < B > 1/3$), and 3 participants were categorized as metacognitively Aware. Using the mean log d1' score of metacognitively Aware participants as prior (2.72) to establish CA, 28 participants had a sensitive null on log d1' and were effectively contingency Unaware, 6 of them being metacognitively insensitive. Fifteen participants were deemed as contingency Aware, 3 of them belonging to the metacognitive Aware group, 1 of them to the metacognitively Unaware group, and 11 having insensitive log d2' scores.

Another 17 participants had an insensitive log d1', their CA could not be established and will be excluded from further analyses, see Table 1. For further analysis, results

will be reported for contingency Aware and Unaware participants, avoiding differentiating them in terms of metacognitive knowledge due to the small sample size of meta-aware participants and the high number of insensitive ones.

		Type I: Contingency Awareness			
		Unaware	Insensitive	Aware	
Type II: Metacognitive Knowledge	Unaware	22	4	1	27
	Insensitive	6	13 **(1)**	11 **(9)**	30
	Aware	0	0	3 **(2)**	3
		28	17	15	60

Table 1:

Contingency table presenting the categorization of participants according to the results on their individual Bayes Factors for Type I outcome-Contingency Awareness and on Type II tests of Metacognitive Contingency Awareness. (x) participants deemed Contingency Aware via post-hoc categorization.

2.4.4.2 Post-hoc contingency awareness

Out of 60 participants, 12 were deemed Aware following the procedure on Experiment 1, eleven of them being Contingency Aware according to Bayesian analyses and one having originally an insensitive B. Four participants did not pass the post-hoc categorization, implying that some forgetfulness might have occurred over time, but generally confirming that the two measurement methods are congruent (see Table 1).

2.4.4.3 Pleasantness

Results show that HR stimuli (mean=.55, SD=.17) were more pleasant than LR (mean = .49, SD = .17), $t(42) = 2.276$, $p = .028$ collapsing both groups. Analysing separately

Aware and Unaware participants, a marginal increase in pleasantness towards HR stimuli compared to LR for Aware participants, t(14) = 1.830, p = .089 was found. The difference was even weaker for Unaware participants, t (27) = 1.545, p = .134, Fig. 7.

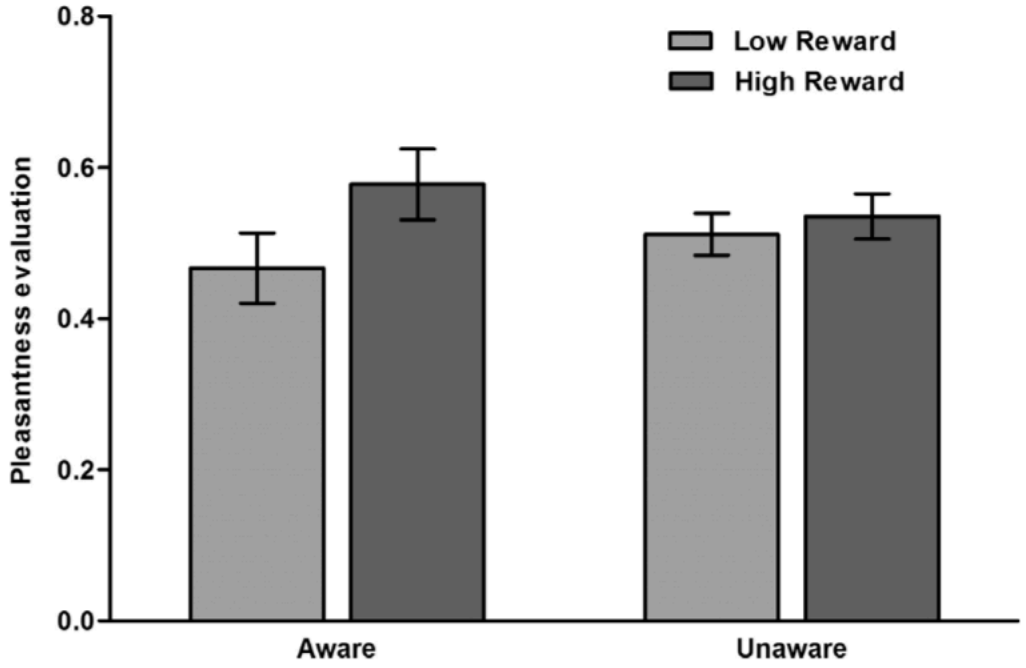


Figure 7: Pleasantness ratings depending on stimulus type and contingency awareness, Experiment 2.

There were no significant differences between High reward and Low reward stimuli in either Aware or Unaware participants, ps > .08. However, a significant stimulus effect (p = 0.28) was Found when the 2 groups were collapsed.

2.4.4.4 Emotional Attentional Blink

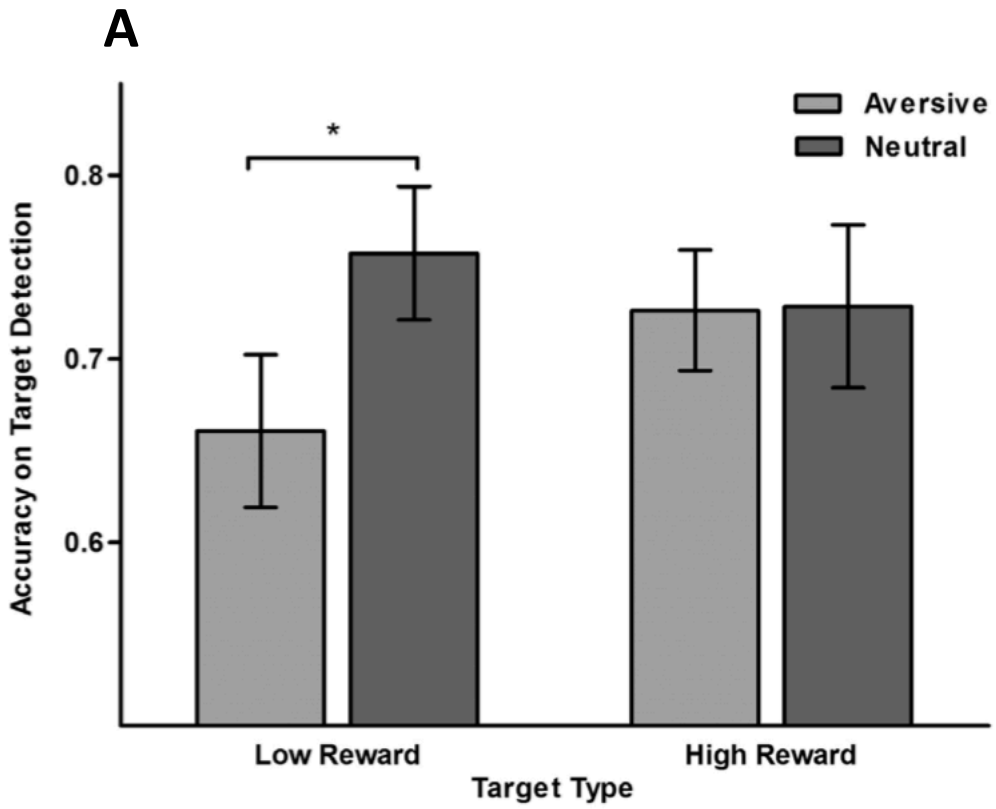
For the Unaware group, there was a main effect of distractor type on accuracy, F(1,27) = 8.064, p = .008. No significant interaction between target and distractor type, F(1,27) = 1.87, p = .183, was found. However, due to the hypothesized effects and taking into account the results obtained on Experiment 1, we performed planned paired samples t-tests.

In the Unaware group aversive distractors decreased detection of LR stimuli, t(27) = 2.668, p = .013 compared to neutral distractors.

Distractor type did not have any significant effect on detection of HR stimuli, t(27) = .−052, p > .9, see Fig. 8. There were no significant differences between HR and LR targets under neutral distractors, t (27) = −.729 p = .472.

These results show again that stimuli conditioned with HR are less affected by the interference of aversive distractors than LR stimuli for Contingency Unaware participants.

For Contingency Aware participants, there was again a main effect of distractor type, F(1,14) = 11.760, p = .004 but no significant interaction between stimulus and distractors, F(1,14)=2.484, p=.137. Here aversive distractors decreased detection of HR compared to neutral distractors, t(14) = 4.185, p = .001. There was no significant effect of distractor on LR target detection, t(14)=−.574, p=.575, see Fig. 8. There was finally a marginally significant difference between HR and LR targets under neutral distractors, t(14) = 1.966, p = .069.


A
Aversive
Neutral
*
0.8
0.7
0.6
Accuracy on Target Detection
Low Reward
High Reward
Target Type

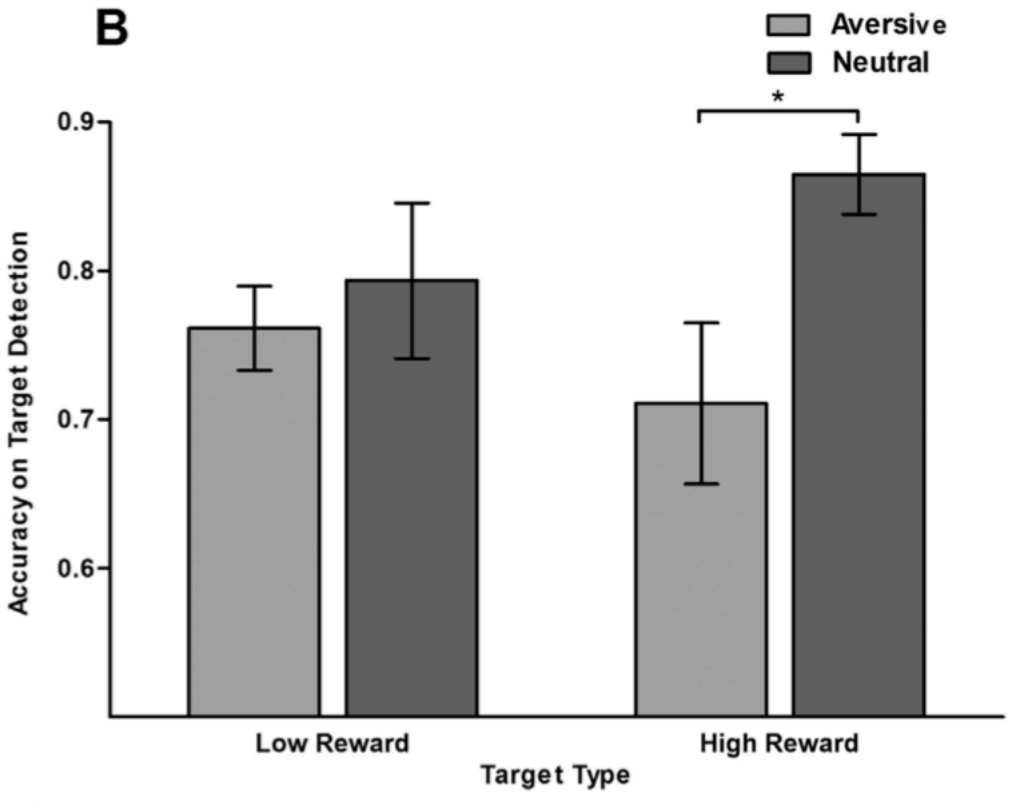

B
Aversive
Neutral
*
0.9
0.8
0.7
0.6
Accuracy on Target Detection
Low Reward
High Reward
Target Type

Figure 8: Emotional Attentional Blink results for Unaware (A) and Aware (B) participants, Experiment 2.

*Accuracy on target detection depending on reward associated and distractor type for Unaware participants. A: * Low reward targets under aversive distractors compared to neutral distractors, p < 0.025. B: * High reward targets under aversive distractors compared to neutral distractors, p < .001. There was a marginally significant difference between HR and LR targets under neutral distractors, p = .069.*

2.4.4.5 Questionnaires

There were no corrected significant differences between Unaware and CA groups in the questionnaire scores (see Table B.1 in Appendix B).

2.4.4.6 Supplementary analyses

No significant differences were found, see Appendix B.

2.4.5 Discussion Experiment 2

Measures of metacognitive awareness incorporated in this experiment gave rise to 3 distinct groups. As expected, participants were categorized as Unaware of the contingencies, as partially Aware, being able to predict the outcomes associated with CSs; and metacognitively Aware, having developed metacognitive knowledge about contingencies. Bayes factors were used to perform this classification, a vital step when determining the existence of unconscious processes. One of the limitations of this analysis is the fact that many participants showed insensitive results, and couldn't therefore be correspondingly classified.

Importantly, we replicated the results obtained on the EAB in Experiment 1. We failed at obtaining heightened emotional responses towards HR stimuli for Unaware and Aware participants separately. However, when considering all participants, HR stimuli were more pleasant than LR. This finding shows that abstract stimuli can acquire emotional salience with conditioning, and partly helped in overcoming the limitations of Experiment 1 in relation to the intrinsic nature of stimuli used.

2.5 General Discussion

Both experiments showed that expectancy awareness is not necessary to generate preferential attention towards CSs. During an EAB task, detection of stimuli associated with LR but not HR probabilities decreased in the presence of aversive emotional stimuli as distractors. That occurred both in Experiment 1 (all participants were unaware) and in Experiment 2 only in participants who were Unaware, and not in those who were Aware of the contingencies. Importantly in Experiment 2 awareness criteria were more rigorous. The ability of HR stimuli to overcome the interference of aversive distractors as opposed to LR stimuli is a proof that attention was preferentially allocated towards HR stimuli. These findings taken together provide a first strong account of an implicitly conditioned attentional response using an EAB task [28].

In experiment 2 the number of participants with awareness of stimulus-outcome contingencies increased allowing us to examine how attention developed in Aware participants. During the EAB, Aware participants seemed to allocate more attention to LR stimuli, as attention to LR targets was not affected by aversive distractors; however

attention to HR stimuli was affected by aversive distractors, implying a decrease in attention allocation to HR stimuli for those participants. This might be explained by differences in the predictive power explicitly obtained by CS.

According to the Pearce-Hall theory of attention [60], it is possible that the increased predictability of HR stimuli in Aware participants leads to a decreased necessity to focus attention on HR stimuli [61], in order to perform accurately, leading to a higher effect of interference by the salient negative distractors (but also to a marginal higher accuracy in the presence of neutral distractors (compared to LR)). On the other hand, for Unaware participants, for which knowledge about stimuli paired with reward is obviously not sufficient to generate correct predictions, the valence obtained by CS through non explicit associations drives their attention preferentially towards HR CS, in accordance with Mackintosh's theory [62] of associative learning. Such an explanation could be supported by the fact that Aware and Unaware participants combined showed pleasantness towards HR over LR stimuli (main effect of stimulus, $p = 0.028$), a measurement of the emotional value of HR stimuli. However when the two groups were separately analysed, both demonstrated a marginal effect, although stronger for Aware participants. Of course this explanation should be taken into consideration with caution as our data did not demonstrate any effect of contingency awareness between HR and LR stimuli; and our data on pleasantness were not as clear for Unaware participants. More research on how metacognition about stimuli-reward associations and emotional and attentional responses to stimuli associated with reward develops may help to integrate both theories [63] and support an understanding on attentional and emotional conditioned responses.

During our first experiment, HR CSs were not evaluated as more pleasant than LR stimuli, inconsistent with previous research [21]. However, a series of factors could have affected the pleasantness ratings. Firstly, stimuli depicting Houses and Buildings were used in Experiment 1. Even though independent measures carried out in a pilot study had discarded a preferential hedonic appraisal of any of the two categories, after conditioning, Houses were evaluated as more pleasant than Buildings, even when the latter stimulus type was associated with higher probabilities of obtaining money. In previous studies when houses and buildings were used as CSs [28], pleasantness ratings were not taken at the end of the conditioning task. During the second experiment, abstract geometric stimuli were used as CSs. Those stimuli were specifically crafted to prevent them from generating any intrinsic emotional reactions [64–66]. This time, stimuli associated with HR were consistently more pleasant than those associated with LR if considering the whole sample. These differences highlight one of the limitations of stimuli used in experiments evaluating preference towards drug cues, as the graphic nature of those stimuli can hinder their ability to generate automatic implicitly learned reactions. These findings also point out the importance of using neutral stimuli in conditioning paradigms (i.e. stimuli with no possible previous value).

Another reason why Experiment 1 may have failed at generating preferential hedonic reactions towards HR stimuli is the fact that pleasantness was measured after the EAB task. This means that CSs had been extensively presented under extinction, as during EAB trials there was no reward following CS presentation. It is possible that the effects of conditioning in Unaware participants were not strong enough to withstand that kind

of extinction procedure, explaining why the intrinsic value of the images took over during the pleasantness evaluation task. On Experiment 2, pleasantness was measured between the conditioning and EAB tasks, and thus that may contribute to the task generating the expected results.

Importantly, in Experiment 2 higher pleasantness for HR over LR was found, but not for Unaware participants in isolation, failing to support previous findings [21].This is in line with previous data showing that CA was necessary for the generation of emotional responses [7], but still cannot rule out the ability of Implicitly conditioning to generate hedonic responses, as a very marginal effect was seen also in unaware participants and as mentioned above a main stimulus effect was highly significant.

Regarding, the EAB for Unaware participants, we find in both experiments that LR stimuli are less detectable in the presence of aversive distractors than neutral, whereas HR stimuli were not. However these effects were more pronounced in Experiment 1. It is possible that these effects were weaker in Experiment 2 due to differences in sample size (n = 43 in experiment 1 and n = 28 in experiment 2) or due the fact that EAB was measured later on in the procedure (after the measurements of pleasantness) pointing again towards a possible effect of extinction.

A recent experiment by Le Pelley et al. [67] showed using a RSVP task that distractors associated with reward only affect target detection under conditions of CA, results that somehow clash with our findings. However, in their task, the conditioning procedure was embedded within the RSVP instead of occurring previously and separately. Moreover, CS in their case acted as distractors and not as targets. Their findings

suggest that CA is necessary for conditioned stimuli during a learning task to affect target detection, whereas our findings suggest that CS paired with high reward probabilities can resist the interference of aversive distractors after a conditioning task.

Our procedure used money as reward, which may be considered not as high in value as primary reinforcers (e.g. food, drugs etc.). However, as Hogarth et al. [35] already posited, conditioning procedures using tobacco or other substance administration as reward can lead to reduced reward value by the occurrence of satiety effects. Satiety decreases pleasantness attributed towards the substance itself [68]. It is possible that this decrease in pleasantness blocks the development of positive attentional responses towards CSs under conditions of Contingency Unawareness. Conditioning paradigms targeting the generation of implicit conditioned responses should therefore use rewards as outcomes for which satiation is difficult to achieve (i.e. money) instead of drug substances or food.

An important aspect of Experiment 2 is that a parsimonious analysis of CA using a statistical approach originating in Implicit learning theories [40] and recurring to Bayes factors [55] allowed us to classify participants in three different groups: those Unaware of the contingencies governing the conditioning task; those able to predict the outcomes associated with each CS; and those able to explicitly describe those contingencies. Most importantly, the rigorous classification obtained using Bayes factors allows determining the true nature of conscious or unconscious processes [50]. Post-hoc expectancy measurements using Likert scales were also compared to online measurements of awareness using d1’ score categorizations. Interestingly, a high

congruency between both kinds of methodologies was found. Arguments against the existence of Implicit conditioning are based often on the types of EA measurements used to classify awareness [19]. Our results show that a sensible approach towards EA measurements suffices in order to obtain reliable implicit measures.

This fact suggests that the problem underlying inconsistent results in the implicit learning literature [16,19,35] lays more within the kind of conditioning procedure or the type of stimuli used, or the measurement of learning by-products (i.e. conditioned emotional responses), rather than EA measurements.

We think that the separation of participants in three groups depending on their CA and metacognition is a useful tool to help us understand learning processes, and hence this rigorous methodology should be prioritised. In our conditioning the number of Aware participants was relatively small to obtain a better differentiation between these subgroups. It is possible that the small number of Aware participants is due to the use of a task-irrelevant conditioning task that may have impeded explicit learning.

A high proportion of participants could not definitely be classified as Aware or Unaware of contingencies on Experiment 2. Their meta-cognitive state was also undetermined due to their Bayes factors for both measurements being insensitive. As learning is a progressive phenomenon, initial trials are uninformative of contingency knowledge compared to later blocks. In this experiment EA was measured every 4 trials so as to prevent excessive priming of awareness development [19]. Those two factors combined lead to a small amount of trials being used for awareness categorization, and therefore generated a higher rate of insensitive results.

In summary, this paper shows convincing evidence of the occurrence of Implicit Pavlovian conditioning whilst presenting a novel approach of CA measurement based on Bayes factors. It suggests that appetitive CSs can elicit increased attention in conditions of contingency unawareness. The attentional correlates of implicit learning appears to match those generated by explicitly learned appetitive CSs as reported in the literature. Our data also indicated a development of implicit emotional responses albeit not as clearly. These findings therefore highlight a possible role of implicit learning in the development of addictive behaviours and support dual process theories of addiction. More research needs to address the development of emotional responses in implicit conditioning, as results have proven to be inconclusive.

The utility of the emotional and attentional responses to stimuli associated with reward for seeking that reward (i.e. the behavioural response) in the absence of awareness remains to be shown.

2.6 Conflicts of interest

Participants provided informed consent to take part in this experiment. Ethical approval was granted by the University of Sussex Life Sciences ethics committee. We declare no conflicts of interest in the development of this research.

2.7 Funding

This research was funded by the University of Sussex.

2.8 Acknowledgements

We thank Professor Zoltan Dienes from the University of Sussex for his help developing the methodology for participant categorization based on Bayes factors.

2.9 Appendix A. Supplementary data

Table A.1

Data for demographic and questionnaire information for Experiment 1 depending on Awareness group and statistics for ANOVA comparing Contingency Aware and Unaware groups.

	Unaware n= 44		**Aware n=4**			
	Mean	SD	Mean	SD	*F*	*p*
Age	23.43	7.32	28.00	13.59	1.23	0.27
Panas positive	2.90	0.79	2.65	1.49	0.32	0.58
Panas negative	1.52	0.48	1.30	0.42	0.76	0.39
BIS	2.82	0.32	3.00	0.42	1.09	0.30
BAS	2.96	0.41	2.90	0.53	0.07	0.80
BAS Drive	2.74	0.64	2.75	0.74	0.00	0.99
BAS Reward	3.35	0.44	3.65	0.25	1.80	0.19
BAS Fun seeking	2.79	0.68	2.31	0.75	1.78	0.19
Barrat Total	2.15	0.41	1.71	0.26	4.34	0.04
Barrat Attentional	2.24	0.48	1.81	0.33	2.99	0.09
Barrat Motor	2.04	0.44	1.66	0.33	2.83	0.10
Barrat Nonplanning	2.16	0.52	1.66	0.30	3.54	0.07
Alcohol Use Total	29.92	24.67	8.45	10.46	2.93	0.09
Binge score	20.20	17.36	3.38	4.03	3.67	0.06
Alcohol units/week	9.71	9.94	5.08	6.47	0.83	0.37
AUDIT	7.98	6.16	2.75	3.40	2.76	0.10

Experiment 1

Pleasantness for Aware participants

For the four Aware participants, pleasantness towards HR (mean=6.97, SD=.89) and LR (mean= 5.74, SD=1.43) stimuli was computed.

Emotional Attentional Blink for Aware participants

For the four Aware participants, accuracy on the EAB was computed. Under neutral distractors, HR stimuli had an accuracy of .67 (SD=.08) and under aversive distractors of .73 (SD=.09). Under neutral distractors, LR stimuli had an accuracy of .55 (SD=.20) and under aversive distractors of .39 (SD=.20).

RT during the conditioning task

Results show no significant differences in RT towards HR and LR stimuli for Unaware participants, there was only a marginal difference, with increased RT towards HR stimuli (mean=2.67, SD=.08) than LR (mean=2.66, SD=.08), $t(43)=1.705$, $p=.095$. For Aware participants, RT towards HR stimuli were 2.74 (SD=.16) and LR stimuli were 2.74 (SD=.16).

2.10 Appendix B. Supplementary data

Table B.1

Data for demographic and questionnaire information for Experiment 2 depending on Awareness group and statistics for ANOVA comparing Contingency Aware and Unaware groups.

	Unaware n= 28		**Aware n=15**		F	p
	Mean	SD	Mean	SD		
Age	20.50	4.10	20.60	1.99	.008	.930
Panas positive	2.76	0.75	3.01	0.65	1.248	.270
Panas negative	1.44	0.41	1.57	0.52	.770	.385
BIS	1.60	0.35	1.89	0.44	5.837	.020
BAS	2.03	0.28	2.05	0.22	.056	.814
BAS Drive	2.35	0.42	2.32	0.55	.044	.834
BAS Reward	1.64	0.33	1.60	0.33	.114	.737
BAS Fun seeking	2.21	0.50	2.35	0.47	.861	.359
Barrat Total	2.02	0.23	2.11	0.27	1.401	.243
Barrat Attentional	2.12	0.43	2.09	0.44	.071	.791
Barrat Motor	1.90	0.30	2.00	0.42	.780	.382
Barrat Nonplanning	2.06	0.27	2.15	0.42	.773	.385
Alcohol Use Total	28.40	19.64	32.57	27.28	.335	.566
Binge score	15.38	11.20	18.73	15.62	.663	.420
Alcohol units/week	13.02	13.63	13.84	13.39	.036	.851
AUDIT	6.36	4.35	7.93	5.39	1.085	.304

RT during the conditioning task

Results show no significant differences in RT towards HR (mean=2.66, SD=.10) and LR (mean=2.66, SD=.10) stimuli for Unaware participants, $t(27)=.480$, $p=.635$. For Aware

participants, there was no significant difference between HR (mean=2.70, SD=.09) and LR (mean=2.69, SD=.09) stimuli, $t(14)=.560$, $p=.584$.

Accuracy towards HR and LR stimuli

There were no significant differences in accuracy towards HR and LR stimuli neither for Unaware, $t(27)=.055$, $p=.956$ (HR=.46, SD=.20; LR=.46, SD=.19), nor for Aware participants, $t(14)=.760$, $p=.460$ (HR=.89, SD=.15; LR=.85, SD=.18).

Metacognition towards HR and LR stimuli

There were no significant differences in d2' scores towards HR and LR stimuli neither for Unaware, $t(27)=1.415$, $p=.169$ (HR=-.19, SD=.54; LR=.12, SD=.76), nor for Aware participants, $t(14)=.704$, $p=.493$ (HR=.1.08, SD=.77; LR=.89, SD=.87).

2.11 References

[1] J. Stewart, H. de Wit, R. Eikelboom, Role of unconditioned and conditioned drug effects in the self-administration of opiates and stimulants, Psychol. Rev. 91 (1984) 251–268, http://dx.doi.org/10.1037/0033-295X.91.2.251.

[2] S. Glautier, Measures and models of nicotine dependence: positive reinforcement, Addiction 99 (2004) 30–50, http://dx.doi.org/10.1111/j.1360-0443.2004.00736.x.

[3] S.J. Moeller, T. Maloney, M.A. Parvaz, et al., Enhanced choice for viewing cocaine pictures in cocaine addiction, Biol. Psychiatry 66 (2009) 169–176, http://dx.doi.org/10.1016/j.biopsych.2009.02.015.

[4] M. Field, K. Mogg, B.P. Bradley, Eye movements to smoking-related cues: effects of nicotine deprivation, Psychopharmacology (Berl.) 173 (2004) 116–123, http://dx.doi.org/10.1007/s00213-003-1689-2.

[5] L.C. Hogarth, K. Mogg, B.P. Bradley, et al., Attentional orienting towards smoking-related stimuli, Behav. Pharmacol. 14 (2003) 153–160, http://dx.doi.org/10.1097/ 01.fbp.0000063527.83818.9e.

[6] L. Hogarth, T. Duka, Human nicotine conditioning requires explicit contingency knowledge: is addictive behaviour cognitively mediated? Psychopharmacology (Berl.) 184 (2006) 553–566, http://dx.doi.org/10.1007/s00213-005-0150-0.

[7] L. Hogarth, A. Dickinson, S.B. Hutton, et al., Drug expectancy is necessary for stimulus control of human attention, instrumental drug-seeking behaviour and subjective pleasure, Psychopharmacology (Berl.) 185 (2006) 495–504, http://dx.doi. org/10.1007/s00213-005-0287-x.

[8] A.W. Stacy, R.W. Wiers, Implicit cognition and addiction: a tool for explaining paradoxical behavior, Annu. Rev. Clin. Psychol 6 (2010) 551–575, http://dx.doi.org/10.1146/annurev.clinpsy.121208.131444.

[9] B. Gawronski, G.V. Bodenhausen, Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change, Psychol.

Bull. 132 (2006) 692–731, http://dx.doi.org/10.1037/0033-2909.132.5.692.

[10] R.W. Wiers, B.D. Bartholow, E. van den Wildenberg, et al., Automatic and con trolled processes and the development of addictive behaviors in adolescents: a review and a model, Pharmacol. Biochem. Behav. 86 (2007) 263–283, http://dx.doi. org/10.1016/j.pbb.2006.09.021.

[11] M. Field, T. Duka, Cues paired with a low dose of alcohol acquire conditioned incentive properties in social drinkers, Psychopharmacology (Berl.) 159 (2002) 325–334, http://dx.doi.org/10.1007/s00213-001-0923-z.

[12] L. Hogarth, A. Dickinson, T. Duka, Explicit knowledge of stimulus? Outcome contingencies and stimulus control of selective attention and instrumental action

in human smoking behaviour, Psychopharmacology (Berl.) 177 (2005) 428–437, http://dx.doi.org/10.1007/s00213-004-1973-9.

[13] A.J. Austin, T. Duka, Mechanisms of attention for appetitive and aversive outcomes in pavlovian conditioning, Behav. Brain Res. 213 (2010) 19–26, http://dx.doi.org/ 10.1016/j.bbr.2010.04.019.

[14] J. De Houwer, S. Thomas, F. Baeyens, Associative learning of likes and dislikes: a review of 25 years of research on human evaluative conditioning, Psychol. Bull. 127 (6) (2001) 853.

[15] W. Hofmann, J. De Houwer, M. Perugini, et al., Evaluative conditioning in humans: a meta-analysis, Psychol. Bull. 136 (2010) 390–421, http://dx.doi.org/10.1037/ a0018916.

[16] S. Sweldens, O. Corneille, V. Yzerbyt, The role of awareness in attitude formation through evaluative conditioning, Pers. Soc. Psychol. Rev. 18 (2014) 187–209, http://dx.doi.org/10.1177/1088868314527832.

[17] G. Weidemann, M. Satkunarajah, P.F. Lovibond, I think, therefore eyeblink: the importance of contingency awareness in conditioning, Psychol. Sci. 27 (2016) 467–475, http://dx.doi.org/10.1177/0956797615625973.

[18] J. De Houwer, Using the implicit association test does not rule out an impact of conscious propositional knowledge on evaluative conditioning, Learn. Motive 37 (2006) 176–187, http://dx.doi.org/10.1016/j.lmot.2005.12.002.

[19] P. Lovibond, D. Shanks, The role of awareness in pavlovian conditioning: empirical evidence and theoretical implications, J. Exp. Psychol. Anim. Behav. Process. 28 (2002) 3–26.

[20] E. Pool, V. Sennwald, S. Delplanque, et al., Measuring wanting and liking from animals to humans: a systematic review, Neurosci. Biobehav. Rev. 63 (2016) 124–142, http://dx.doi.org/10.1016/j.neubiorev.2016.01.006.

[21] S. Jeffs, T. Duka, Predictive but not emotional value of pavlovian stimuli leads to pavlovian-to-instrumental transfer, Behav. Brain Res. 321 (2017) 214–222, http:// dx.doi.org/10.1016/j.bbr.2016.12.022.

[22] J. De Houwer, What are implicit measures and why are we using them, Handb. Implicit Cogn. Addict. (2006).

[23] J. Yiend, The effects of emotion on attention: a review of attentional processing of emotional information, Cognit. Emot. 24 (2010) 3–47, http://dx.doi.org/10.1080/ 02699930903205698.

[24] B.A. Anderson, P.A. Laurent, S. Yantis, Value-driven attentional capture, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 10367–10371, http://dx.doi.org/10.1073/pnas. 1104047108.

[25] J.E. Raymond, K.L. Shapiro, K.M. Arnell, Temporary suppression of visual processing in an RSVP task: an attentional blink? J. Exp. Psychol. Hum. Percept. Perform. 18 (1992) 849–860, http://dx.doi.org/10.1037/0096-1523.18.3.849.

[26] M. McHugo, B.O. Olatunji, D.H. Zald, The Emotional Attentional Blink: what we know so far, Front. Hum. Neurosci. 7 (2013) 151, http://dx.doi.org/10.3389/fnhum.2013.00151.

[27] S.D. Smith, S.B. Most, L.A. Newsome, D.H. Zald, An emotion-induced attentional blink elicited by aversively conditioned stimuli, Emotion 6 (2006) 523–527, http://dx.doi.org/10.1037/1528-3542.6.3.523.

[28] T. Yokoyama, S. Padmala, L. Pessoa, Reward learning and negative emotion during rapid attentional competition, Front. Psychol. 6 (2015) 269, http://dx.doi.org/10. 3389/fpsyg.2015.00269.

[29] L. Hogarth, A. Dickinson, A. Wright, et al., The role of drug expectancy in the control of human drug seeking, J. Exp. Psychol. Anim. Behav. Process. 33 (2007) 484–496, http://dx.doi.org/10.1037/0097-7403.33.4.484.

[30] J. Hur, A.D. Iordan, H. Berenbaum, F. Dolcos, Emotion–attention interactions in fear conditioning: moderation by executive load, neuroticism, and awareness, Biol. Psychol. 121 (2016) 213–220, http://dx.doi.org/10.1016/j.biopsycho.2015.10.007.

[31] C.M. Chubala, B.T. Johns, R.K. Jamieson, D.J.K. Mewhort, Applying an exemplar model to an implicit rule-learning task: implicit learning of semantic structure, Q. J. Exp. Psychol 69 (2016) 1049–1055, http://dx.doi.org/10.1080/17470218.2015.1130068.

[32] Z. Dienes, D. Broadbent, D.C. Berry, Implicit and explicit knowledge bases in artificial grammar learning, J. Exp. Psychol. Learn. Mem. Cognit. 17 (1991) 875–887, http://dx.doi.org/10.1037/0278-7393.17.5.875.

[33] M.A. Stadler, H.L. Roediger III, The Question of Awareness in Research on Implicit Learning, Sage Publications, Inc., 1998.

[34] P.F. Lovibond, D.R. Shanks, The role of awareness in pavlovian conditioning: empirical evidence and theoretical implications, J. Exp. Psychol. Anim. Behav. Process. 28 (2002) 3–26.

[35] L. Hogarth, A. Dickinson, S.B. Hutton, et al., Contingency knowledge is necessary for learned motivated behaviour in humans: relevance for addictive behaviour, Addiction 101 (2006) 1153–1166, http://dx.doi.org/10.1111/j.1360-0443.2006.01459.x.

[36] Z. Dienes, J. Perner, Implicit Knowledge in People and Connectionist Networks, Oxford University Press, 1996.

[37] A.B. Barrett, Z. Dienes, A.K. Seth, Measures of metacognition on signal-detection theoretic models, Psychol. Methods 18 (2013) 535–552, http://dx.doi.org/10.1037/a0033268.

[38] E. Konstantinidis, D.R. Shanks, Don't bet on it! Wagering as a measure of awareness in decision making under uncertainty, J. Exp. Psychol. Gen. 143 (2014) 2111–2134, http://dx.doi.org/10.1037/a0037977.

[39] G. Pleyers, O. Corneille, O. Luminet, V. Yzerbyt, Aware and (dis)liking: item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness, J. Exp. Psychol. Learn. Mem. Cognit. 33 (2007) 130–144, http://dx.doi.org/10.1037/0278-7393.33.1.130.

[40] S.M. Fleming, H.C. Lau, How to measure metacognition, Front. Hum. Neurosci. 8 (2014) 443, http://dx.doi.org/10.3389/fnhum.2014.00443.

[41] P. Lang, M. Bradley, B. Cuthbert, International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. Tech. Rep. A-8, (2008).

[42] A. Mehrabian, J.A. Russell, A questionnaire measure of habitual alcohol use, Psychol. Rep. 43 (1978) 803–806, http://dx.doi.org/10.2466/pr0.1978.43.3.803.

[43] J.B. Saunders, O.G. Aasland, T.F. Babor, et al., Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption—II, Addiction 88 (1993) 791–804.

[44] J.H. Patton, M.S. Stanford, E.S. Barratt, Factor structure of the barratt impulsiveness scale, J. Clin. Psychol. 51 (1995) 768–774, http://dx.doi.org/10.1002/1097 4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1 .

[45] C.S. Carver, T.L. White, Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales, J. Pers. Soc. Psychol. 67 (1994) 319–333, http://dx.doi.org/10.1037/0022-3514.67.2.319.

[46] D. Watson, L.A. Clark, A. Tellegen, Development and validation of brief measures of positive and negative affect: the PANAS scales, J. Pers. Soc. Psychol. 54 (1988) 1063–1070, http://dx.doi.org/10.1037/0022-3514.54.6.1063.

[47] S.N. Garfinkel, A.K. Seth, A.B. Barrett, et al., Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness, Biol. Psychol. 104 (2014) 65–74, http://dx.doi.org/10.1016/j.biopsycho.2014.11.004.

[48] S.B. Most, M.M. Chun, D.M. Widders, D.H. Zald, Attentional rubbernecking: cognitive control and personality in emotion-induced blindness, Psychon. Bull. Rev. 12 (2005) 654–661.

[49] Z. Dienes, D. Berry, Implicit learning: below the subjective threshold, Psychon. Bull. Rev. 4 (1997) 3–23.

[50] Z. Dienes, How bayesian statistics are needed to determine whether mental states are unconscious, Behav. Methods Conscious. Res. (2015) 199–220.

[51] C. Velasco, A. Salgado-Montejo, A.J. Elliot, et al., The shapes associated with approach/avoidance words, Motive Emot. 40 (2016) 689–702, http://dx.doi.org/10. 1007/s11031-016-9559-5.

[52] D.R. Shanks, Regressive Research: The Pitfalls of Post Hoc Data Selection in the Study of Unconscious Mental Processes, (2016), http://dx.doi.org/10.3758/ s13423-016-1170-y.

[53] H. Jeffreys, The Theory of Probability, 1st ed., Oxford Oxford Univ. Press, 1939.

[54] Z. Dienes, Using bayes to get the most out of non-significant results, Front. Psychol. 5 (2014) 781, http://dx.doi.org/10.3389/fpsyg.2014.00781.

[55] A. Sand, M.E. Nilsson, Subliminal or not? Comparing null-hypothesis and bayesian methods for testing subliminal priming, Conscious. Cognit. 44 (2016) 29–40, http://dx.doi.org/10.1016/j.concog.2016.06.012.

[56] H.C. Lau, R.E. Passingham, Relative blindsight in normal observers and the neural correlate of visual consciousness, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 18763–18768, http://dx.doi.org/10.1073/pnas.0607716103.

[57] H. Stanislaw, N. Todorov, Calculation of signal detection theory measures, Behav. Res. Methods Instrum. 31 (1) (1999) 137–149.

[58] S. Chinn, A simple method for converting an odds ratio to effect size for use in meta- analysis, Stat. Med. 19 (22) (2000) 3127–3131.

[59] C. Kunimoto, J. Miller, H. Pashler, Confidence and accuracy of near-threshold discrimination responses, Conscious. Cognit. 10 (2001) 294–340, http://dx.doi. org/10.1006/ccog.2000.0494.

[60] J.M. Pearce, G. Hall, The influence of context-reinforcer associations on instru mental performance, Anim. Learn. Behav. 7 (1979) 504–508, http://dx.doi.org/10. 3758/BF03209710.

[61] L. Hogarth, A. Dickinson, M. Janowski, et al., The role of attentional bias in mediating human drug-seeking behaviour, Psychopharmacology (Berl.) 201 (2008) 29–41, http://dx.doi.org/10.1007/s00213-008-1244-2.

[62] N. Mackintosh, The Psychology of Animal Learning, (1974).

[63] J. Pearce, N. Mackintosh, Two theories of attention: a review and a possible integration, Atten. Assoc. Learn. (2010) 11–39.

[64] G. Gómez-Puerto, E. Munar, M. Nadal, Preference for curvature: a historical and conceptual framework, Front. Hum. Neurosci. 9 (2015) 712, http://dx.doi.org/10. 3389/fnhum.2015.00712.

[65] N. Chen, K. Tanaka, D. Matsuyoshi, K. Watanabe, Cross preferences for colors and shapes, Color Res. Appl. 41 (2016) 188–195, http://dx.doi.org/10.1002/col.21958.

[66] S.J. Westerman, P.H. Gardner, E.J. Sutherland, et al., Product design: preference for rounded versus angular design elements, Psychol. Mark. 29 (2012) 595–605, http://dx.doi.org/10.1002/mar.20546.

[67] M.E. LePelley, T. Seabrooke, B.L. Kennedy, et al., Miss it and miss out: counter productive nonspatial attentional capture by task-irrelevant, value-related stimuli, Atten. Percept. Psychophys. 79 (2017) 1628–1642, http://dx.doi.org/10.3758/ s13414-017-1346-1.

[68] K.C. Berridge, Modulation of taste affect by hunger, caloric satiety, and sensoryspecific satiety in the rat, Appetite 16 (1991) 103–120, http://dx.doi.org/10.1016/ 0195-6663(91)90036-R.

3 Knowledge about the Predictive Value of Reward Conditioned Stimuli Modulates their Interference with Cognitive Processes

3.1 Abstract

Stimuli conditioned with a substance can generate drug approach behaviours due to their acquired motivational properties. According to implicit theories of addiction these stimuli can decrease cognitive control automatically. The present study (n=49) examined whether reward-associated stimuli can interfere with cognitive processes in the absence of knowledge about stimulus-outcome contingencies. Abstract conditioned stimuli (CS) were paired with High (HR) or Low (LR) probabilities of monetary reward using a Pavlovian learning task. Participants were categorised as Aware or Unaware of contingencies using a Bayesian analysis. CS were then used as task irrelevant distractors in modified Flanker and N-back tasks. Results show HR CS can generate increased interference in the Flanker task for participants Unaware of contingencies, contributing further evidence for the existence of implicit Pavlovian conditioning. For the N-back task, working memory performance was affected by HR CS, albeit only for Aware participants. These results suggest that CS can interfere implicitly with cognitive processes in a similar way to drug-related stimuli. Such an effect could occur in a stimulus-driven fashion, devoid of top-down goal directedness. These findings have implications for the conceptualisation and study of implicit

processes in addiction and highlights the necessity to reconsider the measurement of such phenomena.

3.2 Introduction

Motivational properties of stimuli associated with substances are known to play a crucial role in the development of addictive behaviours. Through repeated associations with drug effects, drug related stimuli acquire incentive salience (Berridge & Robinson, 2003; Goldstein & Volkow, 2002), a quality that ultimately drives and directs motivational responses. The instatement of those responses is posited to occur via mechanisms similar to that of Pavlovian appetitive conditioning (Stewart et al., 1984).

In humans, responsiveness to drug-related stimuli has been evaluated via attentional processes (Field & Cox, 2008), and emotional (Pool et al., 2016) or autonomous (Carter & Tiffany, 1999) reactivity. Attentional biases towards drug-related stimuli have been consistently observed for different substances (Bonson et al., 2002; Bradley et al., 2008; Field et al., 2013; Garland et al., 2012; Moeller et al., 2009), commonly evaluated using dot probe tasks (i.e. Townshend and Duka, 2001) targeting overt attention allocation.

Other techniques assessing biases towards drug-related stimuli consist of eye-gaze measurements (Hogarth et al., 2006b) or interference tasks such as the Addiction-Stroop test (Cox & Fadardi, 2006). Interference by task-irrelevant drug-related stimuli has also been evaluated using working memory (WM) tasks (Hester & Garavan, 2009) particularly with cocaine addicts.

Nikolaou et al. (2013) further examined the mechanisms underlying cue interference in cognitive control using a modified Flanker task (Eriksen & Schultz, 1979); they found that task-irrelevant alcohol related stimuli increased RT under high cognitive load, thus proving the effect of drug-related stimuli in the attenuation of cognitive control resources necessary for correct task performance.

Others have also examined the ability of emotionally salient stimuli to interfere with WM using an N-back task (Ladouceur et al., 2009). In this procedure, participants have to respond to targets occurring in a sequence of stimuli. In the case of a 0-back condition this simply means responding when a given stimulus appears. In the case of an n-back condition a response is required when a stimulus is the same as one appearing n stimuli earlier in the sequence. As the n increases the cognitive load on working memory also increases. Interference on performance resulting from the inclusion of drug-related non-target stimuli in the sequence may result from a combination of explicit and implicit processes in drug addiction.

According to the dual process theory of addiction (Wiers & Stacy, 2006), the cascade of events leading to drug approach behaviours is supposed to occur implicitly, under the influence of stimuli associated with the substance (see also Tiffany, 1990). Explicit motives activated in parallel support cognitive control mechanisms and are meant to impede or limit such tendencies. This trade-off between an implicit appetitive system triggered by drug-related stimuli and an explicit cognitive control system based on regulatory executive signals can explain results such as those obtained in the alcohol Flanker task (Nikolaou et al., 2013b).

Most experiments have investigated this matter using stimuli explicitly associated with a substance (e.g. alcohol bottles or cigarettes) for which the stimulus outcome contingency is clearly defined due to their own nature, thus impeding an accurate exploration of implicit components of drug-addiction. Even though the procedures themselves, based on task-irrelevant distractor effects, are thought to be implicit, the explicit attributes of drug related stimuli generated through prior consumption experiences (Wiers et al., 2002) may affect task outcomes (Leganes-Fonteneau et al., 2018). Stimuli conditioned with non-drug rewards can generate value-driven responses equivalent to those of drug-related stimuli (Anderson et al., 2011), both attentional (Hogarth et al., 2006a; Jeffs and Duka, 2017) and emotional (i.e. Austin and Duka, 2010).

However, in seeking to investigate the mechanisms underlying responsiveness towards reward-related stimuli, a key question is whether those responses can occur without conscious awareness of outcome-contingencies (CA), that is, in the absence of predictive knowledge about associations between Conditioned Stimuli (CS) and rewards. The ability of implicit reward-CS associations to produce hedonic and attentional responses has generated an extensive discussion (Lovibond & Shanks, 2002), with research showing inconsistent results. Recent findings (Le Pelley et al., 2017), as well as previous research (i.e. Hogarth et al., 2005, 2006a), appear to show that CA is necessary for the development of responses in Pavlovian appetitive conditioning. These results may stem from the inadequacy of procedures employed to assess learning, both in the measurement of conditioned responses (De Houwer, 2006) and CA (Lovibond & Shanks, 2002).

In a recent series of experiments, in which CA was carefully measured using a novel Bayesian approach (Dienes, 2015; Sand & Nilsson, 2016), we found that targets, paired with increased probabilities of monetary reward, gathered preferential attention in an Emotional Attentional Blink task (Leganes-Fonteneau et al., 2018). Importantly, this was observed in participants Unaware of stimulus-reward contingencies. These results are in line with the postulate that reward predictive stimuli can modulate attentional processes (Failing & Theeuwes, 2017) even in the absence of CA.

As discussed above, research in the field of drug-addiction has demonstrated preferential attentional responses towards task-irrelevant reward-related stimuli; made apparent by their influence on cognitive processes (Hester and Garavan, 2009; Nikolaou et al., 2013b). Similar results have been obtained using secondary reinforcers. For example, using a modified colour-naming Stroop task in which certain task-relevant colours (e.g. a blue font) were associated with monetary outcomes (Krebs et al., 2010), it was found that those colours facilitated task performance. Interestingly, the incentive value of rewarded colours subsequently transferred to task-irrelevant words associated with rewarded stimuli (e.g. the word “blue”), this time generating increased interference. This was considered an example of implicit appetitive learning. However, in this case, as in others (i.e. Anderson, 2015), the implicit nature of the effect generated by CS can only be postulated as no stringent measures of CA were implemented.

Dual process theories of addiction (Wiers & Stacy, 2006) indicate that the chain of events leading to appetitive behaviours is based on implicit processes triggered by conditioned stimuli. It is not clear however if stimuli implicitly associated with a drug,

without conscious knowledge of outcome contingencies, can set off drug-approach behaviours (Hogarth et al., 2006a).

To help clarify this matter we investigated whether CS in the absence of CA can generate task-irrelevant interferences in cognitive control, allowing us to draw a parallel between cognitive processes associated with drug cues and implicitly CS.

To make sure that Pavlovian associations would occur in the absence of awareness we utilised a task-irrelevant reward learning procedure (Yokoyama et al., 2015), pairing stimuli belonging to two different categories with high (HR) or low (LR) probabilities of monetary reward. Using task-irrelevant procedures, it is possible to direct the focus of attention away from the stimulus-reward outcome, and in this way, delay explicit learning of stimulus-reward associations. We measured CA and meta-cognitive knowledge about contingencies on a trial-by-trial basis (Leganes-Fonteneau et al., 2018) to determine the explicit knowledge about outcome-contingencies gained by participants. Using a Bayesian analysis for this purpose we were able to gather sensitive evidence for the existence of non-conscious learning. Finally, emotional responses towards CS were measured, and the interference of CS on cognitive processes was assessed using modified Flanker and N-back tasks with different degrees of cognitive load.

We hypothesized that CS would have an effect on performance for both the N-back and Flanker tasks depending on their value (HR vs. LR) and that the extent to which participants were Aware or Unaware of the contingencies would modulate that interference.

3.3 Results

3.3.1 Questionnaires

Groups were matched on all baseline indices (i.e. Barrat Impulsiveness scale - BIS, Alcohol use disorder identification test - AUDIT, Alcohol use questionnaire - AUQ, Positive and negative affect schedule - PANAS and Bodily perception questionnaire - BPQ; p>.2, in all cases) except for Reversed Digit Span, see Table 1. Groups also did not differ in the distribution of male and female participants, $\chi(1, N = 36) = 1.446$, p=.229 (ratio of female/male: 8/20 for Unaware and 10/16 for Aware).

	Contingency Aware	**n=16**	**Contingency Unaware**	**n=20**		
Between groups comparison						
	Mean	SD	Mean	SD	$t(34)$	p
Age	20.13	5.39	19.55	1.23	0.165	.870
BIS-11	63.81	8.20	64.2	5.87	1.136	.264
AUDIT	7.69	4.70	7.8	5.52	0.886	.382
Binge Score	25.69	16.45	19.29	17.06	0.486	.630
AUQ score	41.43	25.53	33.5	27.51	0.983	.333
PANAS Positive	0.47	0.15	0.44	0.20	0.165	.870
Porges	2.52	0.50	2.31	0.75	1.136	.264
	Mean	SD	Mean	SD	Z	p
Reverse Digit	3.21	0.97	4.13	1.26	2.903	.043
PANAS Negative	0.10	0.07	0.12	0.11	0.161	.888

Table 1: Results and descriptives comparing questionnaires and demographic scores between Contingency Aware and Unaware participants.

3.3.2 Pleasantness

There was a main effect of stimulus-type, $F(1,34)= 10.015$, $p= .003$, reflecting, irrespective of CA, increased pleasantness ratings towards High Reward (HR) (mean= 55.29, SD=.16) compared to Low Reward (LR) CS (mean= 43.97, SD= .17).

This main effect was quantified by a significant CA by stimulus-type interaction, $F(1,34)= 7.899$, $p=.008$. Thus Aware participants rated HR CS as being more pleasant than LR CS, $t(15)= 3.182$, $p= .006$, $B_{U(0,0.06)}= 3.3998$. By contrast, there was no sensitive difference in pleasantness ratings between HR and LR CS in the Unaware group, $t(19)= 0.362$, $p= .721$, $B_{U(0,0.06)}= 0.9054$, see Figure 1.

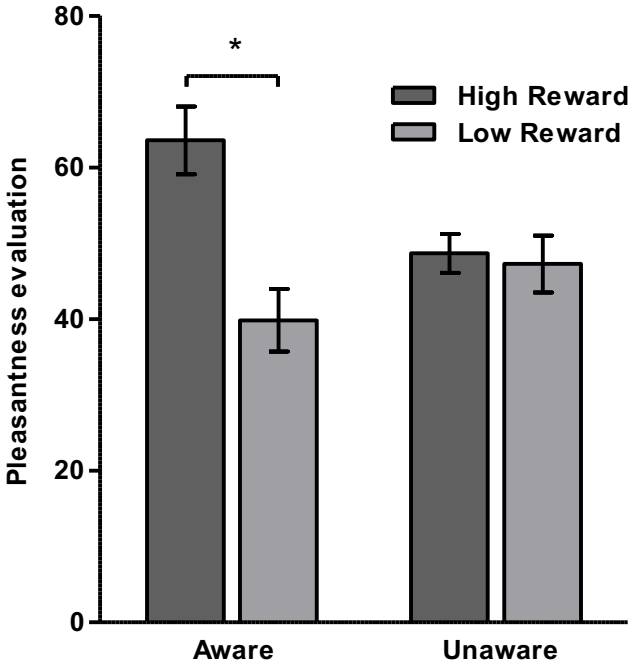


Figure 1: Pleasantness towards Conditioned Stimuli.

Pleasantness ratings towards High Reward and Low Reward stimuli depending on Contingency Awareness.

**Only Aware participants developed preferential emotional responses towards High Reward stimuli, t(15)= 3.182, p= .006.*

3.3.3 N-back

3.3.3.1 Accuracy N-back

For Aware participants, the analysis of the percentage of correct responses showed a main effect of load, $F(1,45)= 29.307$, $p< .001$, with participants performing less accurately in 2-back blocks (mean= 79.43, SD= 8.22) than in 0-back blocks (mean= 90.10, SD= 7.89). Similarly, Unaware participants also had lower accuracy scores in the 2-back (mean= 84.65, SD= 7.95) than the 0-back condition (mean=88.49, SD=10.83); main effect of load, $F(1,54)= 8.006$, $p= 0.007$. There were no other significant main effects or interactions neither for Aware nor for Unaware participants when looking at the percentage of correct responses, $Fs< 0.427$, $p> .516$.

Analyses of net differences, resulted in no significant main effects or interactions neither for Aware, nor for Unaware participants, $Fs< 0.935$, $ps> .338$.

For proportion of correct Hits there were no significant interactions or main effects, neither for Aware, nor for Unaware participants, $Fs< 1.638$, $ps> .207$.

3.3.3.2 Latencies N-back

With respect to latencies taking into account all trials, Aware participants were overall slower on 2-back trials (mean= 2.76, SD=.13) compared to 0-back trials (mean= 2.74, SD= .10); main effect of load, $F(1,15)= 5.624$, $p=.032$. There were no other significant main effects or interactions in the Aware group, $Fs < 2.597$, ps > .128. The analysis performed in the Unaware group did not result in any statistically significant main effect or interaction, $Fs< 1.015$, $ps> .327$.

With respect to latencies on target trials only: The analysis in the Aware group showed a marginally significant main effect of stimulus-type, $F(1,15)= 3.367$, $p= .086$, indicating that latencies to LR targets were faster (mean=2.70, SD=0.07) than to HR targets (mean= 2.72, SD= 0.07). In addition, there was a marginally significant stimulus-type by load interaction, $F(1,15)= 3.875$, $p=.068$. Post-hoc paired samples t-tests showed that responses on HR target trials were marginally slower than responses on LR target trials in the 2-back block, $t(15)= 2.122$, $p= .051$ (see Figure 2). No other simple effect was statistically significant, $p> .149$. Finally, the main effect of load was not statistically significant, $F(1,15)=0.587$, $p= .456$.

For Unaware participants, there were no statistically significant main effects or interactions, $F<1.733$, $p>.205$, in all cases.

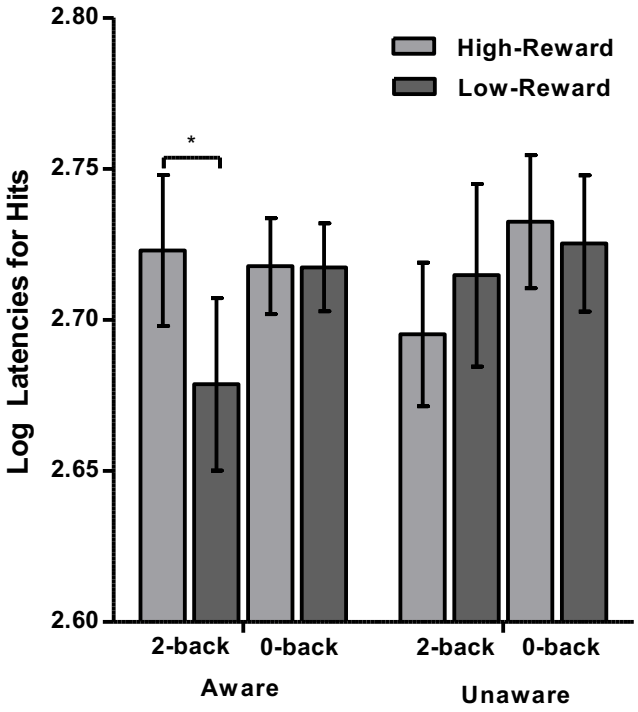


Figure 2: Reaction times during the N-back task between Aware and Unaware participants for Hits.

For Aware participants Latencies in the 2-back condition towards High Reward were higher than those towards Low Reward, p=.051.

3.3.4 Flanker Task

3.3.4.1 Accuracy Flanker

For Unaware participants there was a main effect of stimulus-type on accuracy, $F(2,85)= 6.834$, $p= .002$, explained by significantly lower accuracy overall on HR (mean= 94.90, SD= 3.84) compared to Control trials (mean= 97.07, SD= 2.28), $Z= 3.675$, $p< .001$, and marginally lower accuracy on LR (mean= 95.68, SD= 3.84) than Control trials, $Z= 1.945$, $p=.052$. There was no significant difference in accuracy between HR and LR trials, $Z= 0.857$, $p= .391$. The main effect of congruency was also significant, $F(1,85)=$

152.970, $p < .001$, and reflected increased accuracy in general on congruent (mean= 99.38, SD= 1.05) than incongruent trials (mean= 92.39, SD= 5.03). Both of these main effects were quantified by an interaction between stimulus type and congruency in the Unaware group, $F(2,85) = 6.420$, $p = .002$, due to differences between Control and HR stimuli across levels of congruency, $\chi(1) = 12.807$, $p = .001$. Wilcoxon post-hoc tests showed a significant difference between Control and HR trials in the incongruent condition, $Z = 2.255$, $p = .024$, that did not occur in the congruent condition, $Z = .447$, $p = .655$, see Table 2.

Similarly, for Aware participants, there was also a main effect of stimulus-type on accuracy, $F(2,70) = 7.820$, $p < .001$, explained by overall lower accuracy for HR (mean= 93.43, SD= 5.39) compared to Control trials (mean= 95.37, SD= 5.98), $Z = 3.819$, $p < .001$, and lower accuracy on LR (mean=93.06, SD=7.48) compared to Control trials, $Z = 2.111$, $p = .035$. There was no difference in accuracy between HR and LR trials, $Z = 0.175$, $p = .861$. The main effect of congruency was also significant, $F(1,70) = 146.051$, $p < .001$, and reflected increased accuracy in general on congruent (mean= 99.14, SD= 1.13) than incongruent trials (mean= 88.77, SD= 11.08). There was however no interaction between stimulus-type and congruency, $F(2, 70) = 2.360$, $p = .102$.

	Congruent						**Incongruent**					
	High-reward		Low-reward		Control		High-reward		Low-reward		Control	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Unaware	99.54	1.07	98.92	2.88	99.69	0.90	90.28	7.70	92.44	5.94	94.44	4.57
Aware	99.07	1.36	98.52	2.54	99.81	0.72	87.78	10.85	87.59	13.36	90.93	11.38

Table 2: Accuracy during the Flanker task.

Accuracy during the Flanker task depending on contingency awareness, cognitive load and stimulus type.

3.3.4.2 Latencies Flanker

For Unaware participants there was a main effect of stimulus-type, $F(2, 34)= 29.828$, $p< .001$, with faster latencies on Control trials (mean= 568.88, SD= 56.62) compared to both HR (mean= 594.74, SD= 53.89), $p< .001$, and LR trials (mean= 596.79, SD= 55.76), $p< .001$. The main effect of congruency was also significant, $F(1,17)= 425.936$, $p< .001$, with Unaware participants responding faster on congruent (mean= 537.21, SD= 52.71) than on incongruent trials (mean= 636.40, SD= 58.15). These main effects were quantified by a significant stimulus-type by congruency interaction, $F(2,34)= 6.506$, $p= .004$. Follow-up tests showed a significant main effect of stimulus-type in the congruent condition, $F(2,34)= 10.153$, $p< .001$. Latencies were slower on LR trials than on both HR, $p= .036$, and Control trials, $p< .001$. No significant difference was found between HR and Control trials, $p= .346$. In the incongruent condition, the main effect of stimulus type, $F(2,34)= 19.017$, $p< .001$, reflected slower latencies on both HR, $p< .001$, and LR trials, $p= .002$, compared to Control trials. No significant difference was found between HR and LR trials, $p= .552$, see Table 3 for descriptive statistics.

Aware participants also showed a main effect of stimulus-type, $F(2,28)= 11.011$, $p< .001$. This was explained by faster latencies on Control trials (mean= 575.03, SD= 57.84) compared to both HR (mean= 592.39, SD=54.72), $p= .005$, and LR trials (mean= 591.87, SD= 59.35), $p= .006$. The main effect of congruency was also significant,

F(1,14)= 268.607, *p*< .001, with Aware participants also responding slower on incongruent (mean= 643.027, SD= 64.47) than on congruent trials (mean= 529.59, SD= 51.02). However, in this group, the stimulus-type by congruency interaction was not statistically significant, *F*(2,28)= 0.606, *p*= .553.

	Congruent						**Incongruent**					
	High-reward		Low-reward		Control		High-reward		Low-reward		Control	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Unaware	536.16	54.26	548.37	53.03	527.09	54.63	653.32	57.04	645.21	0.04	610.68	70.05
Aware	534.41	49.31	533.31	57.33	521.05	49.17	650.37	64.03	650.42	65.64	624.09	70.47

Table 3: Flanker task descriptive statistics for Latencies.

Latencies for the Flanker task depending on cognitive load, stimulus type and Contingency Awareness.

3.3.4.3 Flanker effect

Unaware participants showed a main effect of stimulus-type, *F*(2,34)= 6.506, *p*= .004. HR CS generated more interference than both LR, *p*= .042, or Control stimuli, *p*= .006, see Figure 3.

This was not observed, in Aware participants, *F*(2,28)= 0.606, *p*= .553.

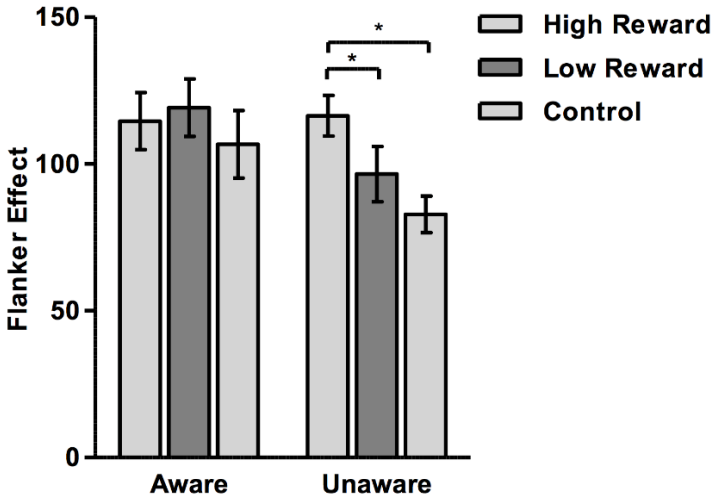


Figure 3: Flanker scores.

Latency difference-scores for congruent versus incongruent trials in the Flanker task by contingency awareness and stimulus type. For Unaware participants high cognitive load generates more interference in high-reward trials compared to low-reward, t(17)= 2.740, p= .014, and control stimuli, t(17)= 3.615, p= .002.

3.4 Discussion

The present study set out to examine the extent to which reward CS can interfere with cognitive processing and the degree to which this remains true in the absence of CA. By using a task-irrelevant Pavlovian conditioning paradigm and setting stringent criteria in the determination of CA it was possible to reliably identify implicit influences. The use of a Bayesian approach provides a sensitive measurement of the unconscious nature of learning, countering otherwise well-grounded methodological

criticisms regarding measurements of implicit learning (Lovibond & Shanks, 2002), due to lax determinations of unconscious processes (Shanks, 2016).

A significant Flanker effect was found in response latencies generated by HR CS, albeit only in participants Unaware of the contingencies. On the other hand, in the N-back task, under high cognitive load, a higher interference in the presence of HR CS compared to LR CS was only found in Aware participants. Subjective hedonic responses to HR CS, as measured by pleasantness ratings, were also only seen in participants Aware of contingencies; Unaware participants displayed an insensitive response pattern.

The development of hedonic responses congruent with reward signalling only in Aware participants clarifies recent findings (Leganes-Fonteneau et al., 2018) suggesting that CA is necessary for the development of subjective emotional responses. Previous research from our own group had found instances of hedonic responses in the absence of CA (Jeffs & Duka, 2017). These incongruent results may again be due to the use of less stringent CA categorizations. However, insensitive results, with regard to pleasantness for Unaware participants (as shown with Bayesian analyses), indicate that we cannot fully discard the development of such responses also in this group. Other factors contributing to the development of hedonic conditioned responses, such as interoceptive abilities (Pollatos and Schandry, 2008), may ultimately help resolve such disparities. However, it seems clear that the assessment of subjective hedonic responses is by no means an adequate tool for the examination of implicit learning (De Houwer, 2006).

As expected, decreases in performance (accuracy and latencies) were found for the Flanker task, in the incongruent compared to the congruent condition; also, as expected in the N-back task, accuracy decreased in the 2-back compared to the 0-back blocks, showing the suitability of the procedures to generate increased cognitive load.

Results of the Flanker task are in line with data recently obtained by our group (Leganes-Fonteneau et al., 2018) and by others (i.e. Bourgeois et al., 2016) demonstrating the ability of reward CS to grab preferential attention implicitly. HR CS, paired with increased probabilities of reward, generated more cognitive interference compared to LR CS. Importantly that happened solely for participants Unaware of contingencies. These results are relevant enough by themselves as they provide further evidence for the existence of implicit Pavlovian conditioning effects.

Research findings relating to the nature of the attentional processes underlying task interference by irrelevant distractors have proven to be inconsistent. On one hand, attentional resources are posited to be necessary for task interference (Pessoa & Ungerleider, 2004) countering the argument of automaticity of salience effects and converging with the interpretation of attentional biases towards drug cues acting as a top-down albeit involuntary mechanism (Brown et al., 2018). However, value driven attentional capture (Anderson et al., 2011) is also posited to occur involuntarily. In our case, as the interference generated by HR CS occurs without conscious awareness of their reward predictive value, we can posit that their influence is implicit, devoid of goal-directedness.

The ability of CS to interfere with cognitive control tasks in a stimulus-driven fashion has two-fold implications for implicit theories of addiction. On the one hand, this mere effect shows that stimuli associated with a reward can generate approach behaviours in a non-declarative or explicit way and can interfere with cognitive control mechanisms, providing further evidence for the existence of implicit processes in drug-addiction (Wiers & Stacy, 2006). This implies a need to reconsider how cognitive control interference induced by drug-related stimuli implicitly can be tested. Explicit drug cues typically utilised in relevant experiments (e.g. pictures of alcohol bottles for alcohol related cognitive bias) might not allow implicit mechanisms involved in cognitive biases to be revealed. On the other hand, if we consider that the effects of reward associated cues can occur on cognitive interference implicitly, as shown here, we can extrapolate and add to our understanding of mechanisms underlying drug-cue interference in cognitive control (i.e. on an equivalent Flanker task Nikolaou et al., 2013b), that drug-cue interference could be affected, at least partially, via implicit and involuntary processes detached from goal-directedness (Hester & Garavan, 2009).

It is puzzling that Aware participants did not show a stronger Flanker effect of HR CS compared to LR CS compared to Unaware participants. This might be explained by an overall decrease in baseline performance for Aware participants (illustrated by higher Flanker effect for Control trials in Aware compared to Unaware participants, ($t(31)=2.302$, $p=.028$). Such an effect could impede an appropriate interference by HR CS. Another possibility is that due to the conscious knowledge about contingencies, participants were consciously trying to identify the HR CS (exploiting attentional

resources to achieve this), in order to minimize their interference. Such a cognitive process could explain the high flanker effect seen also in Control trials for that group.

On the N-back task stimulus interference was observed only for Aware participants. Under high cognitive load (2-back condition), latencies for Hits in the presence of HR CS were longer than towards LR CS. We can assume that this effect is due to the increased salience of HR CS.

It seems therefore that cognitive mechanisms underlying approach to reward related stimuli tested by different tasks are influenced differently by CA. CS effects in the Flanker task can be explained according to bias competition models (Desimone & Duncan, 1995) by which under limited attentional resources (such as high cognitive load) salient stimuli will grab attention and interfere with the task (Vuilleumier, 2005), in our case implicitly. We need however to explain why no effects were found in the N-back task for Unaware participants. The N-back task differs from the Flanker task as it requires less attentional demand (detecting a colour vs. a target arrow embedded in distractors) but involves more complex cognitive processes (recalling the colour of a previous image to respond, whilst at the same time ignoring target irrelevant information vs. overcoming distractors). Previous research found that masked presentations of emotional distractors altered N-back performance, albeit not for high load trials (Uher et al., 2014). It is possible that interference on the N-back task does not directly target attentional mechanisms but rather working memory (WM) and inhibitory processes (Kensinger & Corkin, 2003), and that a more conscious presentation of distractors, or of their incentive value, is necessary for interference to occur in WM (LeDoux, 2002).

The results obtained in this study are subject to a number of limitations. The effects of HR CS on response latencies in the N-back task are observed solely on accurate target trials, that is, on relatively few instances, maybe because HR CS only affect trials in which the coloured frame matches the 2-back stimulus, interfering with target recognition. The observation that interference in the Flanker task was not found for Aware participants remains to be fully explained. While it may be possible to provide an account of these results based on the effect of conscious knowledge about contingencies, it would be informative to compare the effect of emotional stimuli or masked distractors with the interference generated by CS on both tasks (Nikolaou et al., 2013a).

In conclusion, Implicit processes play a crucial role in the development of drug addiction, particularly in drug-approach behaviours. We observed a clear interference in cognitive control by stimuli conditioned with reward in the absence of CA. This effect provides further evidence for the existence of implicit Pavlovian conditioning and has implications for the understanding of dual-process theories of addiction. Uncovering implicit mechanisms of drug-approach behaviours may prove essential for the development of novel treatments in substance use disorders.

3.5 Methods and materials

3.5.1 Participants

Forty-nine Psychology students from the University of Sussex completed the experiment (mean age= 20.04, SD= 3, 34; 25 females). Exclusion criteria were a history of mental disease and undergoing heavy medical treatment at the time of the study.

All participants were given course credits and £2 for taking part in the study, and the study was approved by the University of Sussex Life Sciences ethics committee.

3.5.2 Measures

3.5.2.1 Questionnaires

Reversed Digit Span measurements were used to index working memory capacity (Redick & Lindsey, 2013; Wechsler, 2008).

The Alcohol Use Questionnaire (AUQ; Mehrabian and Russell, 1978) was used to assess average weekly alcohol use over the past six months. The questionnaire also provides a binge drinking score based on the speed of alcohol consumption, and the number as well as the proportion of times that participants were drunk in the last six months (Townshend and Duka, 2002).

Severity of alcohol use was measured with the 10-item Alcohol Use Disorder Identification Test (AUDIT; Saunders et al., 1993).

The Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) was used to measure positive and negative mood at the start of the study session. Participants rate how they feel at that moment using a 5-point Likert scale (1 = “very slightly” – 5 = “Extremely”). The questionnaire consists of 10 items per construct.

The Body Perception Questionnaire (BPQ; Porges, 1993) is a 45-item questionnaire evaluating the subjective ability to detect internal bodily sensations. Participants had to indicate on a 5-point Likert scale (1 = “never” – 5 = “always”) the frequency with which they felt different sensations, for example facial twitches or bowel movements.

Finally, the Barratt Impulsiveness Scale (BIS; Patton et al., 1995) is a 30 item questionnaire evaluating different factors contributing to overall impulsiveness, namely Attentional, Motor and Non-planning impulsivity.

3.5.2.2 Conditioning task

A task-irrelevant conditioning procedure (Leganes-Fonteneau et al., 2018; Yokoyama et al., 2015) was implemented in order to train participants to associate high and low probabilities of monetary reward with two different categories of CS. Thirty-six geometrical stimuli belonging to each CS category (squares vs. octagons) were produced with Ink-Scape vector design software[2].

On each trial, a stimulus from one CS category (i.e. a square or an octagon) was presented on the computer screen with an overlaid green or yellow coloured square. Participants were asked to press a green or yellow key depending on the colour of the square. Stimuli remained on screen for 2000ms or until a response was made. If the response was correct, participants could win 10p. For HR CS, the probability of winning was 90%. For LR CS, the probability was 10%. After a response was recorded feedback about the outcome of the trial was provided ("You win 10p" or "You win nothing") for 1500ms. Trial outcomes depended solely on the CS presented during the trial and the associated probabilities of reward. The stimulus category was counterbalanced across participants.

---

[2] see Leganes-Fonteneau et al. 2018 for a detailed description of stimulus development, Figure 4-5 for an example, and supplementary materials for the complete collection of stimuli: https://osf.io/t9qu6/?view_only=242170271cc1418aae7bb65f6c744f85

On 50% of the HR trials and on 50% of the LR trials, participants had to indicate if they thought they would win money (Yes/No expectancy responses). Following the expectancy response, they were also asked to rate how confident they were about their response on a 5-point Likert Scale (1. “completely guessing”, 2. “more or less guessing”, 3. “fairly sure”, 4. “almost certain”, 5. “completely certain”). The measurement of accuracy and confidence on reward prediction allowed us to determine participants’ metacognitive knowledge about contingencies (Barrett et al., 2013).

The conditioning procedure comprised a total of 5 blocks, with 72 trials in each block. At the end of each block participants transferred the amount of earned coins from a bank box to their “earnings” box. At the end of the final block, they were told how much money they had won in total.

Participants were kept naïve about the contingencies between the CS categories and the reward-outcome probabilities, although they were told that they could win money at the end of each trial. They were also told that they would be asked a series of questions about their expectancy and confidence and that their responses to those questions would not affect the outcome of the trial in terms of reward probability.

3.5.2.3 Pleasantness measurement

Immediately after the conditioning procedure participants rated the pleasantness associated with each category of CS. Eighteen squares and eighteen octagons (randomly selected from the original 72 stimuli) were presented in a random order,

one at a time, and participants were asked to rate how pleasant they found each stimulus on a 5-point Likert scale (1- Not pleasant at all / 5-Extremely pleasant).

3.5.2.4 CS N-back task

Each trial of the CS N-back task began with a fixation cross in the centre of screen (jittered 1-3secs; average 2 secs). This was followed by the Stimulus display for 500ms, followed by a response interval for 1000ms.

The Stimulus display consisted of a CS surrounded by a coloured frame. On 50% of the trials, the CS was a HR CS, while on the remainder 50% of the trials it was a LR CS. The colour of the frame was either a primary colour (i.e. red, blue, and yellow) or a non-primary colour (pink, orange, and green).

The task consisted of two conditions. In the 0-back condition participants were instructed to press one button if the colour of the frame was a primary colour (target trial) and another button if it was a non-primary colour (control trial) as quickly and as accurately as possible. In the 2-back condition, participants had to remember the colour of the frame and press one button if the colour of the frame matched the one shown 2 trials before (target trial) and another button if the colours did not match (see Figure 4).

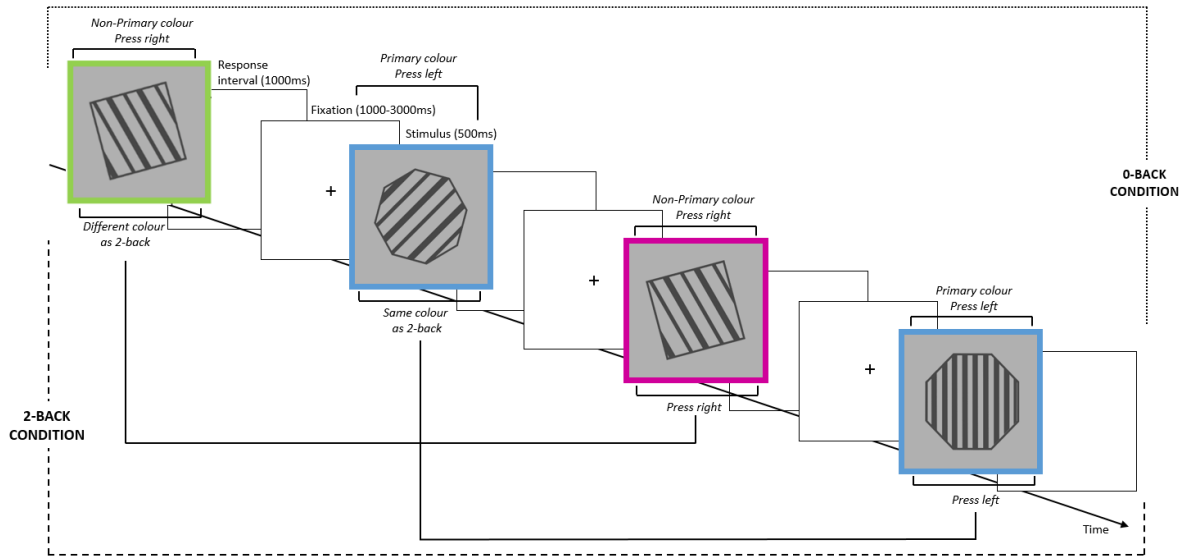


Figure 4: N-back task.

The task consisted of two different conditions. In the 0-back condition participants indicated whether the colour of the frame was a primary or a non-primary colour. In the 2-back condition they had to indicate whether the colour of the frame matched the one presented two trials before. High and Low Reward conditioned stimuli were task irrelevant.

A total of eight blocks (four 0-back and four 2-back) were presented in an ABAB order. A 0-back block was always presented first. At the start of each block, a short instructions screen (jittered duration 4-8secs, average 6secs) reminded participants what they should do in that block.

Each block consisted of eight control and four target trials. In half of the control trials and in half of the target trials the image presented was a HR CS, with the other half being LR CS. In order to match the visual properties of each block, target trials always

consisted of primary colours and control trials of non-primary colours regardless of the N-back condition. Within blocks, the colours surrounding HR and LR CS were also matched.

In each 2-back block, at least one (maximum two) lure trials were introduced. These trials were control trials in which the colour of the frame matched the one presented 1-back or 1-forward. Lure trials were allocated equally often between HR and LR image conditions.

The percentage of correct responses was calculated for each N-back and stimulus-type condition separately. We also computed the proportion of Hits (proportion of accurate Target trials), and net differences subtracting False Alarms rates (proportion of "target-like" responses on Control trials) from the proportion of Hits for each of the conditions. Reaction times to correct responses (i.e. latencies) were calculated for each N-back and stimulus type condition. Reaction times to correct target trials (i.e. latencies for Hits) were also computed separately for each N-back and stimulus type condition.

Finally, participants completed two practice blocks, one for each N-back condition, in which a plain grey background was presented as part of the stimulus display. A minimum accuracy of 65% in each practice block was required to proceed to the real task. Participants not reaching this threshold were given the task instructions again and repeated the practice blocks.

3.5.2.5 CS Flanker task

The task was adapted from Nikolaou and colleagues (2013).

Each trial began with the presentation of a fixation cross for a jittered duration (850-1150 ms), followed by the stimulus display for 800ms, and a response interval for 700ms (see Figure 5).

The stimulus display consisted of a horizontal row of five arrows superimposed on either a plain grey background, or on task-unrelated background images that belonged to either the high or the low reward CS categories. The central arrow was the target, and was surrounded by two distracting arrows (flankers) on either side. Participants were instructed to ignore the flankers and press one key if the central arrow was pointing to the left, and another key if it was pointing to the right, as quickly and accurately as possible. In the congruent condition flankers pointed in the same direction as the target (e.g. < < **<** < <). In the incongruent condition flankers pointed in the opposite direction to the target (e.g. > > **<** > >, bold font added only for illustration). There were 4 different flanker and target combinations (target pointing left or right and flankers pointing left or right).

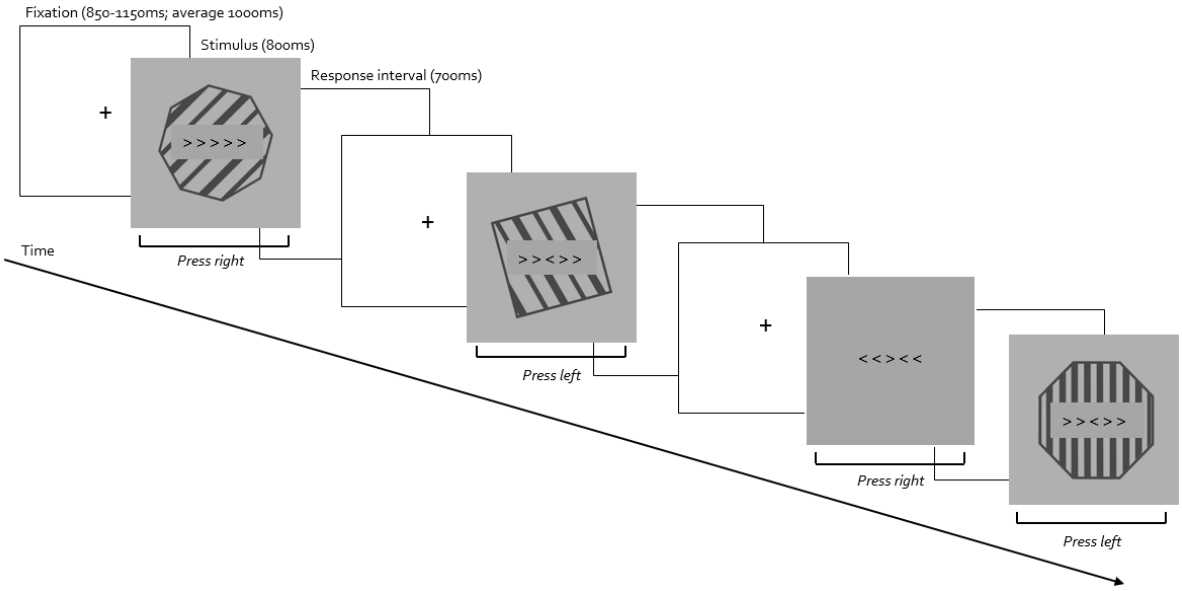


Figure 5: Flanker task.

Participants responded depending on the direction of the central arrow, which could be surrounded by congruent or incongruent arrows. High and Low reward conditioned stimuli as well as Grey control backgrounds were task irrelevant.

Each of the flanker combinations was superimposed on each of 20 selected HR and LR CS images, to generate 80 HR and 80 LR trials (40 in the congruent and 40 in the incongruent condition). An additional 40 congruent and 40 incongruent trials with a plain grey background (i.e. control trials) were also included to generate a single task block of 240 trials in total. Trial order was pseudorandomized to avoid presentation of the same background and same congruency condition for more than 3 consecutive trials.

Mean latencies and accuracy scores (% of correct responses) were computed for each background image (i.e. HR, LR and Control) under both the congruent and incongruent

conditions. The flanker effect was also computed for each image condition separately as the difference in average latency in the incongruent condition minus the average latency in the congruent condition.

Participants completed 40 practice trials, 20 congruent and 20 incongruent, with only the grey plain background before the main block. A minimum accuracy of 70% was required to proceed to the experimental block (achieved by all participants).

3.5.3 Procedure

Each participant completed a single testing session. They first gave written informed consent that they agreed to take part in the study. This was followed by the completion of the AUQ, AUDIT, PANAS, BIS-11, and BPQ. They then completed two tasks designed to measure interoceptive awareness (detailed in Garfinkel et al. (2015), see supplemental materials for data), followed by the reward conditioning procedure and pleasantness evaluation. Finally, they completed the CS N-back and Flanker tasks in a counterbalanced order. At the end of the experiment they were debriefed and compensated for their participation.

3.5.4 Data analysis

3.5.4.1 Bayesian analysis

Bayesian analyses provide a statistical tool with which the sensitivity of results can be determined. This way it is possible to extract conclusions out of non-significant findings generated by frequentist statistical approaches (Zoltan Dienes, 2014). A Bayes factor (*B*) of above 3 shows compelling evidence towards the alternative hypothesis (i.e. two means are different) whereas a *B* below 1/3 provides substantial evidence

towards the null hypothesis (i.e. there is no difference between two means). A *B* between 3 and 1/3 implies there is not enough evidence in either direction.

Bayes factors were used (a) as a tool for sensitive categorization of participants into those who were aware of the CS-HR/LR contingencies and those who were not (see Leganes-Fonteneau et al., 2018 for details and below for a summary); and (b) to examine the sensitivity of within group comparisons of the pleasantness ratings (see "Pleasantness" below).

3.5.4.2 Contingency Awareness Categorization

Determining the unconscious nature of a mental process requires evidence of an inability or failure to consciously perceive that process. This typically involves asserting the null hypothesis that performance on some knowledge related task is no different to chance, which cannot be evaluated using traditional frequentist statistical analyses (Zoltan Dienes, 2015). Therefore a Bayesian approach (Sand & Nilsson, 2016) was used to categorize participants as Aware or Unaware of contingencies using the data gathered from the conditioning procedure (Leganes-Fonteneau et al., 2018).

Using Signal Detection Theory (SDT) methods (Barrett et al., 2013; Stanislaw and Todorov, 1999), we computed the number of Type I: Hits (responding "Yes" on a HR trial), Correct Rejections (responding "No" on a LR trial), False Alarms (responding "Yes" on a LR trial) and Misses (responding "No" on a HR trial) from blocks 4 and 5 of the conditioning task. Type I scores reflect performance accuracy.

Type II scores, provide an account of the metacognitive knowledge generated during the task using the relationship between accuracy and confidence responses (Kunimoto

et al., 2001). Confidence responses were transformed to a dichotomous variable: responses of two or below (i.e. “more or less guessing” or “completely guessing”) were classified as “not confident”, and those above two as “highly confident”. Thus, a Type II Hit was defined as a Type I Hit or Correct Rejection with high confidence; a Correct Rejection as a Type I Miss or False Alarm with low confidence; a False Alarm as a Type I Miss or False Alarm with high confidence; and a Miss as a Type I Hit or Correct Rejection with low confidence.

Logistic d1’(Logd1’) and Standard Error d1’ (SEd1’) scores were calculated for each participant as well as Logd2’ and Sed2’ scores. As d2’ scores rarely exceed d1’ scores (Zoltan Dienes, 2015), a Bayes factor was calculated for each participant on their Logd2’ modelling H_1 with a Uniform going from 0 (chance level) to their own Logd1’ as a prior. Participants with $B<1/3$ were categorized as Metacognitively Unaware, whereas those with a $B>3$ were considered Metacognitively Aware, and the rest as insensitive.

The mean Logd1’ score from those participants Metacognitively Aware was used as a maximum for a uniform to model H_1 in order to test the sensitivity of each participant’s Logd1’. Each corresponding *B* allowed determining their CA as significant (i.e. aware of the stimulus-reward contingencies), sensitively null (i.e. unaware of the contingencies) or insensitive (neither sensitively aware nor unaware)[3].

[3] The MatLab script developped to generate this categorization is available online and can be used not only for learning tasks, but also for visual detection experiments: https://osf.io/p7n9b/?view_only=8295819dae61452ebdf4b3d82ccc61c9.

Out of the 49 participants who completed the study, on the basis of Type II scores, 7 had metacognitive awareness, 12 were definitely Unaware, and 30 showed an insensitive result. On the basis of Type I scores, 16 participants were categorized as being Aware of contingencies, 20 were Unaware, and 13 had an insensitive Bayes factor.

		Type I Awareness			
		Aware	Unaware	Insensitive	Total
Type II Awareness	Aware	7	0	0	7
	Unaware	1	11	0	12
	Insensitive	8	9	13	30
	Total	16	20	13	49

Table 4: *Distribution of participants between Contingency Awareness and Meta-cognition groups*

Thus, participants could be categorized as Aware, Unaware or Insensitive both on a metacognitive and on a contingency awareness level (see Table 4). All analyses reported below were performed on the basis of Type I categorization and only with the 16 Aware and 20 Unaware participants, discarding the 13 participants who showed an Insensitive result.

3.5.4.3 Questionnaires

A series of Independent Samples t-tests compared age and scores on questionnaires (BIS, AUDIT, AUQ, PANAS Positive and BPQ) between Aware and Unaware groups. Due to violations of normality a Wilcoxon test was performed on PANAS Negative and Reverse Digit Span data. A Chi-square was computed to test for differences in gender distribution.

3.5.4.4 Pleasantness

A 2x2 mixed ANOVA with stimulus-type (HR vs. LR) as the within, and CA (Aware vs. Unaware) as the between subjects' factor examined the effects of contingency awareness on pleasantness ratings for HR/LR CS.

A Bayes factor using significant data from Leganes-Fonteneau et al. (2018) as priors was computed for each group separately to quantify differences in pleasantness ratings between HR and LR CS.

3.5.4.5 N-back task

One participant was excluded from all analyses involving the N-back task due to low accuracy in the 2-back condition (<54%). Given the differences in sample size existing between Aware and Unaware conditions analyses are conducted separately for each group (Leganes-Fonteneau et al., 2018).

3.5.4.6 Accuracy:

For each of the three accuracy indices we computed a 2x2 ANOVA with stimulus-type (HR vs. LR) and load (0-back vs. 2-back) as within subjects' factors. Due to the observed violations of normality (Shapiro-Wilk tests, *ps*<.004) that were non-amendable using transformations, these ANOVAs were performed using ARTool non-parametric analyses for non-normal distributions in R (Wobbrock et al., 2011).

3.5.4.7 Latencies

Due to violations of normality, all latency scores were log transformed and analyses were performed on the log transformed data. We examined latencies on all trials as well as latencies of target trials only in separate analyses. These analyses used 2x2 ANOVAs with stimulus-type (HR vs. LR) and load (0-back vs. 2-back) as within subjects' factors, and significant interactions were explored using paired-samples t-Tests.

3.5.4.8 Flanker task

Participants with accuracy deviating by more than 2 SDs from the mean in the congruent Control condition were considered to be outliers. Consequently 3 participants were excluded from all analyses involving Flanker task data. As for the analysis of the N-Back task, all analyses described below were computed in each group separately.

3.5.4.9 Accuracy

Normality was violated for all accuracy scores. Thus, analyses on accuracy data were performed using ARTool. These used 2x2 ANOVAs with congruency (congruent vs.

incongruent) and stimulus-type (HR vs. LR vs. Control) as within subjects' factors. Chi Squares were performed on interactions followed by Wilcoxon tests.

3.5.4.10 Latencies

Latency and flanker effect scores were normally distributed. Analyses of latency data involved 2x2 ANOVAs, with congruency (congruent vs. incongruent) and stimulus-type (HR vs. LR vs. Control) as within subjects' factors. Significant main effects of stimulus-type were followed by post-hoc Bonferroni corrected contrasts. Significant stimulus-type by congruency interactions were explored by running One-way Repeated Measures ANOVAs separately at each level of congruency, and significant effects were followed up further with Bonferroni corrected contrasts. In order to examine differences in the Flanker effect computed for the Control, LR and HR stimulus-type conditions, we used One-way Repeated Measures ANOVAs followed by Bonferroni corrected contrasts.

3.6 Acknowledgements

The authors would like to acknowledge once more Prof. Zoltan Dienes from the University of Sussex for his contribution in the development of Bayesian categorizations of implicit and explicit knowledge.

3.7 References

Anderson BA (2015). Value-driven attentional capture is modulated by spatial context. *Vis cogn* **23**: 67–81.

Anderson BA, Laurent PA, Yantis S (2011). Value-driven attentional capture. *Proc Natl Acad Sci U S A* **108**: 10367–71.

Austin AJJ, Duka T (2010). Mechanisms of attention for appetitive and aversive

outcomes in Pavlovian conditioning. *Behav Brain Res* **213**: 19–26.

Barrett AB, Dienes Z, Seth AK (2013). Measures of metacognition on signal-detection theoretic models. *Psychol Methods* **18**: 535–552.

Berridge KC, Robinson TE (2003). Parsing reward. *Trends Neurosci* **26**: 507–13.

Bonson KR, Grant SJ, Contoreggi CS, Links JM, Metcalfe J, Weyl HL, *et al* (2002). Neural systems and cue-induced cocaine craving. *Neuropsychopharmacology* **26**: 376–86.

Bourgeois A, Neveu R, Vuilleumier P (2016). How Does Awareness Modulate Goal-Directed and Stimulus-Driven Shifts of Attention Triggered by Value Learning? *PLoS One* **11**: e0160469.

Bradley B, Field M, Healy H, Mogg K (2008). Do the affective properties of smoking-related cues influence attentional and approach biases in cigarette smokers? *J Psychopharmacol* **22**: 737–745.

Brown CRH, Duka T, Forster S (2018). Attentional capture by alcohol-related stimuli may be activated involuntarily by top-down search goals. *Psychopharmacology (Berl)* doi:10.1007/s00213-018-4906-8.

Carter BL, Tiffany ST (1999). Meta-analysis of cue-reactivity in addiction research. *Addiction* **94**: 327–40.

Cox W, Fadardi J (2006). The addiction-stroop test: Theoretical considerations and procedural recommendations. *Psychol Bull* **132(3)**: 443.

Desimone R, Duncan J (1995). Neural mechanisms of selective visual attention. *Annu Rev Neurosci* **18**: 193–222.

Dienes Z (2014). Using Bayes to get the most out of non-significant results. *Front Psychol* **5**: 781.

Dienes Z (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. *Behav methods Conscious Res* 199–220

Eriksen CW, Schultz DW (1979). Information processing in visual search: A continuous flow conception and experimental results. *Percept Psychophys* **25**: 249–263.

Failing M, Theeuwes J (2017). Selection history: How reward modulates selectivity of visual attention. *Psychon Bull Rev* 1–25

Field M, Cox WM (2008). Attentional bias in addictive behaviors: a review of its development, causes, and consequences. *Drug Alcohol Depend* **97**: 1–20.

Field M, Mogg K, Mann B, Bennett GA, Bradley BP (2013). Attentional biases in abstinent alcoholics and their association with craving. *Psychol Addict Behav* **27**: 71–80.

Garfinkel SN, Seth AK, Barrett AB, Suzuki K, Critchley HD (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biol Psychol* **104**: 65–74.

Garland EL, Franken IHA, Howard MO (2012). Cue-elicited heart rate variability and

attentional bias predict alcohol relapse following treatment. *Psychopharmacology (Berl)* **222**: 17–26.

Goldstein RZ, Volkow ND (2002). Drug Addiction and Its Underlying Neurobiological Basis: Neuroimaging Evidence for the Involvement of the Frontal Cortex. *Am J Psychiatry* **159**: 1642–1652.

Hester R, Garavan H (2009). Neural mechanisms underlying drug-related cue distraction in active cocaine users. *Pharmacol Biochem Behav* **93**: 270–277.

Hogarth L, Dickinson A, Duka T (2005). Explicit knowledge of stimulus?outcome contingencies and stimulus control of selective attention and instrumental action in human smoking behaviour. *Psychopharmacology (Berl)* **177**: 428–437.

Hogarth L, Dickinson A, Hutton SB, Bamborough H, Duka T (2006a). Contingency knowledge is necessary for learned motivated behaviour in humans: relevance for addictive behaviour. *Addiction* **101**: 1153–66.

Hogarth L, Dickinson A, Hutton SB, Elbers N, Duka T (2006b). Drug expectancy is necessary for stimulus control of human attention, instrumental drug-seeking behaviour and subjective pleasure. *Psychopharmacology (Berl)* **185**: 495–504.

Houwer J De (2006). What are implicit measures and why are we using them. *Handb implicit Cogn Addict*

Jeffs S, Duka T (2017). Predictive but not emotional value of Pavlovian stimuli leads to pavlovian-to-instrumental transfer. *Behav Brain Res* **321**: 214–222.

Kensinger EA, Corkin S (2003). Effect of negative emotional content on working memory and long-term memory. *Emotion* **3**: 378–93.

Krebs RM, Boehler CN, Woldorff MG (2010). The influence of reward associations on conflict processing in the Stroop task. *Cognition* **117**: 341–347.

Kunimoto C, Miller J, Pashler H (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Conscious Cogn* **10**: 294–340.

Ladouceur CD, Silk JS, Dahl RE, Ostapenko L, Kronhaus DM, Phillips ML (2009). Fearful faces influence attentional control processes in anxious youth and adults. *Emotion* **9**: 855–64.

LeDoux J (2002). *Cognitive-emotional interactions: Listen to the brain*. .

Leganes-Fonteneau M, Scott R, Duka T (2018). Attentional responses to stimuli associated with a reward can occur in the absence of knowledge of their predictive values. *Behav Brain Res* **341**: 26–36.

Lovibond PF, Shanks DR (2002). The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *J Exp Psychol Anim Behav Process* **28**: 3–26.

Mehrabian A, Russell JA (1978). A questionnaire measure of habitual alcohol use. *Psychol Rep* **43**: 803–806.

Moeller SJ, Maloney T, Parvaz MA, Dunning JP, Alia-Klein N, Woicik PA, *et al* (2009).

Enhanced Choice for Viewing Cocaine Pictures in Cocaine Addiction. *Biol Psychiatry* **66**: 169–176.

Nikolaou K, Field M, Critchley H, Duka T (2013a). Acute Alcohol Effects on Attentional Bias are Mediated by Subcortical Areas Associated with Arousal and Salience Attribution. *Neuropsychopharmacology* **38**: 1365–1373.

Nikolaou K, Field M, Duka T (2013b). Alcohol-related cues reduce cognitive control in social drinkers. *Behav Pharmacol* **24**: 29–36.

Patton JH, Stanford MS, Barratt ES (1995). Factor structure of the barratt impulsiveness scale. *J Clin Psychol* **51**: 768–774.

Le Pelley ME, Seabrooke T, Kennedy BL, Pearson D, Most SB (2017). Miss it and miss out: Counterproductive nonspatial attentional capture by task-irrelevant, value-related stimuli. *Attention, Perception, Psychophys* **79**: 1628–1642.

Pessoa L, Ungerleider LG (2004). Neuroimaging studies of attention and the processing of emotion-laden stimuli. *Prog Brain Res* **144**: 171–182.

Pollatos O, Schandry R (2008). Emotional processing and emotional memory are modulated by interoceptive awareness. *Cogn Emot* **22**: 272–287.

Pool E, Sennwald V, Delplanque S, Brosch T, Sander D (2016). Measuring wanting and liking from animals to humans: A systematic review. *Neurosci Biobehav Rev* doi:10.1016/j.neubiorev.2016.01.006.

Porges S (1993). Body perception questionnaire. *Lab Dev Assessment, Univ Maryl*

Redick TS, Lindsey DRB (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychon Bull Rev* **20**: 1102–1113.

Sand A, Nilsson ME (2016). Subliminal or not? Comparing null-hypothesis and Bayesian methods for testing subliminal priming. *Conscious Cogn* **44**: 29–40.

Saunders JB, Aasland OG, Babor TF, De JR, Fuente ' L, Grant ' M (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption—II. *Addiction* **88**: 791–804.

Shanks DR (2016). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes.

Stanislaw H, Todorov N (1999). Calculation of signal detection theory measures. *Behav Res methods, instruments.*

Stewart J, Wit H de, Eikelboom R (1984). Role of unconditioned and conditioned drug effects in the self-administration of opiates and stimulants. *Psychol Rev* **91**: 251–268.

Tiffany ST (1990). A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes. *Psychol Rev* **97**: 147–68.

Townshend J, Duka T (2002). Patterns of alcohol drinking in a population of young social drinkers: a comparison of questionnaire and diary measures. *Alcohol*

Alcohol

Townshend JM, Duka T (2001). Attentional bias associated with alcohol cues: differences between heavy and occasional social drinkers. *Psychopharmacology (Berl)* **157**: 67–74.

Uher R, Brooks SJ, Bartholdy S, Tchanturia K, Campbell IC (2014). Increasing Cognitive Load Reduces Interference from Masked Appetitive and Aversive but Not Neutral Stimuli. *PLoS One* **9**: e94417.

Vuilleumier P (2005). How brains beware: neural mechanisms of emotional attention. *Trends Cogn Sci* **9**: 585–594.

Watson D, Clark LA, Tellegen A (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *J Pers Soc Psychol* **54**: 1063–1070.

Wechsler D (San Antonio, TX: The Psychological Corporation.: 2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*. .

Wiers RW, Stacy AW (2006). Implicit Cognition and Addiction. *Curr Dir Psychol Sci* **15**: 292–296.

Wiers RW, Stacy AW, Ames SL, Noll JA, Sayette MA, Zack M, *et al* (2002). Implicit and Explicit Alcohol-Related Cognitions. *Alcohol Clin Exp Res* **26**: 129–137.

Wobbrock JO, Findlater L, Gergle D, Higgins JJ (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. *Proc 2011 Annu Conf Hum factors Comput Syst - CHI '11* 143doi:10.1145/1978942.1978963.

Yokoyama T, Padmala S, Pessoa L (2015). Reward learning and negative emotion during rapid attentional competition. *Front Psychol* **6**: 269.

4 The role of interoception in appetitive conditioning

4.1 Abstract

Interoception, the perception of internal bodily sensations, modulates the development of emotional responses. Two experiments studied the role of interoception in appetitive Pavlovian conditioning, particularly in emotional responses to reward associated stimuli and the accuracy in reward prediction. Using a reward learning task, naturalistic (Experiment 1, *n*=47) and abstract stimuli (Experiment 2, *n*=59) were paired with high or low probabilities of monetary reward. Both experiments demonstrate that individuals with high interoception show elevated emotional responsiveness towards stimuli predicting high reward, whereas individuals with low interoception do not. Experiment 1 also shows enhanced reward prediction in individuals with high versus low interoception, while Experiment 2 did not find an effect on metacognitive contingency awareness. Therefore, only participants with high interoceptive abilities were able to develop stimulus congruent emotional reactions and successfully predict stimulus-outcome contingencies. These findings highlight the role of viscero-afferent responses in reward learning models, and their perception could foster the development of conditioned responses, potentiating drug related cue-reactivity. These findings may have implications for the development of strategies for prevention and treatment of addiction.

4.2 Introduction

The role of viscero-afferent responses in the generation of emotional reactions is integral to 'peripheral' theories of emotion. These propose a basis for emotional feelings in the central representation and perception of changes in bodily physiology (James, 1884). Interoception, the ability to perceive internal bodily sensations (Cameron, 2001; Sherrington, 1948) emanating from inflammatory, thermoregulatory or visceromotor functioning, (e.g. cardiac signalling (Critchley, Wiens, Rotshtein, Ohman, & Dolan, 2004)), is found to modulate the generation of emotional responses (Dunn et al., 2010) and affective experiences (Critchley et al., 2004; Pollatos, Gramann, & Schandry, 2007; Wiens, 2005).

High interoception can also facilitate emotional learning. Interoception modulates bodily responses in the presence of emotional stimuli as well as the recall of those stimuli (Pollatos & Schandry, 2008).

The role of interoception on emotional associative learning was also examined by Pfeifer et al. (2017) using an emotional face learning task, and again providing evidence supporting the ability of highly interoceptive subjects to perform better in emotional learning. Furthermore, experimentally directing attention towards bodily responses can facilitate the development of learning (Raes & De Raedt, 2011); participants performing a heart beat tracking task developed stronger conditioned responses during a subsequent aversive conditioning task compared to their counterparts who did not.

Interoception may also support appraisal of emotional stimuli without conscious experience. Indeed, during a subliminal aversive conditioning task, participants with high interoceptive accuracy were better able to predict the occurrence of an electric shock (Katkin et al., 2001). Thus, increased interoception may facilitate the development of contingency awareness (CA) during aversive conditioning.

Emotions can influence addictive behaviour (Verdejo-Garcia, Perez-Garcia, & Bechara, 2006), thus interoceptive events may also contribute to drug addiction. Indeed, under repeated drug exposure, cues associated with the substance acquire reinforcing properties and incentive salience generating hedonic, attentional and autonomic responses. These processes are equivalent to those observed through appetitive Pavlovian conditioning (Stewart, de Wit, & Eikelboom, 1984) and participate in the maintenance of drug related behaviours (Wiers & Stacy, 2006). Furthermore the link between interoception and drug addiction has been shown using neuro-imaging studies (Antonio Verdejo-Garcia, Clark, & Dunn, 2012), lesion studies (Naqvi, Rudrauf, Damasio, & Bechara, 2007) and behavioural accounts (Sönmez, Kılıç, Çöl, Görgülü, & Çınar, 2016).

The insula is a brain structure that has been isolated as underlying both interoceptive processes and addictive behaviours. The involvement of insula in the processing of afferent body signals and interoception has repeatedly been demonstrated (Critchley et al., 2004; Pollatos et al., 2007; Schulz, 2016) as well as the relevance of this brain area in emotional responses (Garfinkel & Critchley, 2013; Terasawa, Shibata, Moriguchi, & Umeda, 2013). The insula also has a significant role in addictive behaviours (Naqvi & Bechara, 2010). Its function as a mediator of emotional

experience and bodily states could explain processes leading to drug use and cravings (see Garavan, 2010 for a review), as its ability to represent interoceptive effects of substance administration may drive drug seeking behaviours (Naqvi, Gaznick, Tranel, & Bechara, 2014) through the pairing of autonomic and visceral reactions elicited by drug administration with drug related stimuli (Stewart, May, Tapert, & Paulus, 2015). Activation of such autonomic and visceral responses in the presence of drug related stimuli can prime the organism to adopt drug seeking behaviour. However, the mechanisms by which interoception facilitates addiction are not fully understood.

The ability of substance-related cues to elicit conditioned hedonic responses has been demonstrated in a variety of drugs, including cocaine (i.e. Moeller et al., 2009), tobacco (Mogg, Bradley, Field, & De Houwer, 2003) and alcohol (i.e. Field, Mogg, Zetteler, & Bradley, 2004). However, measuring hedonic responses in the laboratory towards naturalistic stimuli intrinsically related with the substance can generate demand awareness in participants, inflating responses towards drug-related cues. For that reason, we have followed appetitive conditioning procedures in the laboratory associating abstract neutral stimuli with monetary rewards (Leganes-Fonteneau, Scott, & Duka, 2018). This way it is possible to isolate and study the generation of conditioned responses without the effect of previous experiences.

Given that reward conditioned stimuli can generate hedonic reactions (e.g. Jeffs & Duka, 2017), and equivalent attentional responses (e.g. Anderson, Laurent, & Yantis, 2011; Leganes-Fonteneau et al., 2018) to naturalistic drug related stimuli, this paper set out to examine the role of interoception in the development of appetitive Pavlovian conditioning using monetary rewards; the objective being to evaluate the

extent to which interoception could contribute to the development of conditioned emotional responses.

Participants completed two well established tasks, the heartbeat tracking (Schandry, 1981) and heartbeat discrimination (Katkin, Reed, & Deroo, 1983; Whitehead, Drescher, Heiman, & Blackwell, 1977) tasks measuring different interoceptive abilities. Out of these tasks, measures of interoceptive accuracy (behavioural performance on these tasks), sensibility (subjective confidence) and metacognitive interoception (relationship between confidence and accuracy) are extracted (Garfinkel, Seth, Barrett, Suzuki, & Critchley, 2015).

For the purpose of this paper we will focus on individual differences in metacognitive interoception (Garfinkel, Manassei, et al., 2016) also termed interoceptive awareness (Garfinkel & Critchley, 2013; Garfinkel et al., 2015) and “interoceptive insight” (Khalsa et al., 2017) as such an approach has been suggested to provide a novel tool in examining the role of interoception in emotional processes (Garfinkel & Critchley, 2013). Moreover, we hypothesise that meta-cognitive interoception may be particularly implicated in the development of contingency awareness as it is the dimension of interoception that specifically pertains to the conscious processing of interoceptive signals. The common denominator between interoceptive awareness, contingency awareness and subjective emotional ratings being consciously perceived information, it might be more useful, and appropriate, to investigate this relationship at an explicit level.

An established conditioning procedure was applied (Leganes-Fonteneau et al., 2018; Yokoyama, Padmala, & Pessoa, 2015). A series of stimuli were paired with high (HR) or

low (LR) probabilities of reward. Reward prediction and CA were measured out of this task.

A previously published paper using this data base allowed categorising participants as Aware or Unaware of the contingencies (see Leganes-Fonteneau et al., 2018), which could have an effect on the development of hedonic responses. To control for a possible such effect we accounted for CA effects in the analyses of pleasantness to better isolate the role of interoception.

Participants were separated according to their levels of metacognitive interoceptive ability in both the heartbeat tracking and discrimination tasks. It was hypothesized that participants with high metacognitive interoception would develop more pleasantness towards stimuli paired with HR compared to those paired with LR. In addition, we examined whether metacognitive interoception could explain the development of knowledge of the outcome that the stimuli predict (i.e. CA) in Experiment 1, and the ability to explicitly express that knowledge (i.e. metacognitive CA) in Experiment 2.

4.3 Experiment 1

4.3.1 Aims

The aim of Experiment 1 was to determine if differences in interoceptive ability affect the development of stimulus congruent emotional responses after a conditioning task. In addition, we evaluated whether interoception can determine the development of CA as learning progresses.

4.3.2 Methods

4.3.2.1 Participants

Forty-eight University of Sussex students took part in this experiment. Interoceptive measurements for two participants were lost due to technical issues, thefore 46 participants were included in the final analysis (26 females, mean age= 23.85, SD=8.06). Participants signed up for the study through an online recruitment system and were compensated financially or with compulsory course credits for their participation.

Participants undergoing medical treatment (except contraceptive pill) or with a history of mental illness were excluded from the experiment. This study was granted ethical approval by the University of Sussex Life Sciences ethics committee.

4.3.2.2 Apparatus and stimuli

A Nonin 8000SM finger pulse oximeter and an XPOD transmitter monitored participants' heartbeats during the interoception tasks at a frequency of 75Hz, its input was processed by a MatLab script. Pictures of Houses (small unifamiliar edifice) and Buildings (multi-story construction) (36 per category) gathered from the internet were selected from an original pool of 50 Houses and 50 Buildings so that they matched for pleasantness ratings given by an independent sample of participants in a pilot study (Leganes-Fonteneau et al., 2018).

Stimuli were presented on a Dell ACPI 64-bit PC, screen refresh rate= 16.6ms using E-prime 2.9. Data was analysed using Matlab 2014 and SPSS 24. During the conditioning

phase 10, 20 and 50 pence coins were used as tangible reinforcers at the end of each block, located in a bank box. An earnings box was also made available.

4.3.2.3 Questionnaires

Participants' demographic characteristics were assessed at entry to the study.

Alcohol consumption patterns were assessed with the Alcohol Use Questionnaire (AUQ, Mehrabian & Russell, 1978). This questionnaire measures the number of drinks consumed per week and the speed of consumption. A Binge Score is also extracted using the number of times participants reported being drunk in the last month, the amount of drinks per hour and the percentage of times intoxicated (Townshend & Duka, 2002). In addition the Alcohol Use Disorders Identification Test (AUDIT) (Saunders et al., 1993) was also administered, this identifies alcohol-related risk behaviours through 10 multiple-choice items.

Impulsivity was also assessed using the Barratt Impulsiveness Scale (BIS-11) (Patton, Stanford, & Barratt, 1995). The BIS-11 provides a 30-item measure of motor, attentional and non-planning impulsivity. BIS-11 questionnaire was administered together with the Behavioural Inhibition and Activation Systems Questionnaire (BIS/BAS) evaluating approach-avoidance behaviours (Carver & White, 1994). The BIS is related to negative reward and punishment avoidance, whereas the BAS reflects the person's disposition to engage in goal-directed behaviours. Current mood was assessed by the PANAS mood questionnaire (Watson, Clark, & Tellegen, 1988) which provided a measure of Positive and Negative affects through 10 mood items associated with each construct. Finally subjective measures of interoception were

gathered using the Porges Body Perception Questionnaire (BPQ, Porges, 1993). This subscale measures bodily sensations (e.g. eye fatigue, muscle tension) through 45 items. Participants are requested to indicate from 1 to 5 how often they feel each sensation.

4.3.2.4 Interoception measurements

Interoceptive Accuracy: Two tasks measuring participants' ability to feel their heartbeat, the tracking and the discrimination task, served as a basis to develop the interoceptive accuracy measurement. Tasks were always presented in the same order. The finger pulse oximeter was placed on the annular finger of the non-dominant hand only for interoceptive measurements.

Heartbeat tracking task: During the heartbeat tracking task (Schandry, 1981) participants heard through a set of speakers the word "Start" and were instructed to count how many heartbeats they felt until they heard "Stop". They then had to report to the experimenter the number of heartbeats felt. At the beginning of the task participants completed a practice trial lasting 20s. The task consisted of 6 trials varying in length (25, 30, 35, 40, 45 and 50s) occurring in random order. Accuracy scores for each trial of the tracking task were computed using the formulae:

$$(|nbeatsreal - nbeatsreported|) \div ((nbeatsreal + nbeatsreported) \div 2)$$

to account for the accuracy biases induced by high counts on longer trials (Hart, McGowan, Minati, & Critchley, 2013).

Heartbeat discrimination task: During the heartbeat discrimination task (Katkin et al., 1983; Whitehead et al., 1977) the emission of ten auditory tones (100 ms, 440Hz) is triggered by participants' heartbeat. The tones sound at the same rate as the heartbeat, however, for half of the trials the tones were synchronized with their heartbeat and for the other half a phasic delay of 300 ms was inserted between each heartbeat and the tone. Therefore, on 50% of trials participants heard 10 tones emitted at the same rate as their heartbeat but unsynchronized with it. Participants had to indicate at the end of each trial whether they thought the tones were synchronous or asynchronous with their heartbeat. The task included 20 sequences altogether, fully randomizing the occurrence of synchronised and non-synchronised trials.

For each trial, independent accuracy scores were computed to form average scores for each task.

Participants were instructed to feel their heartbeat within their whole bodies, and were precluded from manually feeling their pulse on their chest, neck or wrist.

Interoceptive Sensibility: After each trial, for both tasks, participants had to indicate how confident they were in their responses, from "total guess" to "completely confident", using a Visual Analog Scale (VAS; 0 to100) via Inquisit. Average Confidence scores were computed across trials.

Interoceptive Awareness: Metacognitive interoceptive awareness, the level of insight participants have about their ability to detect their own bodily sensations, is evaluated by the correspondence between accuracy and confidence on both interoceptive tasks.

This measures thus determines the extent to which participants are aware of their accuracy in assessing their own heart timing (see Data Analysis section).

4.3.2.5 Conditioning task

Using a task irrelevant conditioning procedure (Yokoyama et al., 2015), stimuli belonging to each of the categories (Houses or Buildings) were paired with 80% or 20% probabilities of getting a reward, the value of which was always 10p. On the rest of the trials participants obtained nothing. The stimulus category paired with High (HR) or Low reward (LR) probability was counterbalanced across participants.

Each trial began with the presentation of a fixation cross for 500 ms followed by the presentation of a HR or LR CS with an overlaid green or yellow Square.

Participants' were instructed to press a green or yellow key depending on the colour of the overlaid Square whilst the stimuli were on screen. Stimuli remained on screen for 2000 ms or until participants effected a response (max recorded time 1499 ms). Inmediately after their response, feedback appeared for 1500 ms indicating whether they had earned 10p. or nothing, depending on the stimulus-outcome contingency. Trials in which participants pressed the wrong button were automatically awarded 0p.

Participants were told they would occasionally obtain money but were not informed about the nature of stimulus-reward contingencies. Expectancy awareness was measured via a Likert scale after responding to the colour of the Square. Participants had to press a key, from 1 to 9, to indicate how likely they thought they were to win 10p. whilst the stimulus compound remained on screen until response. Expectancy measurements occurred only on 25% of the trials so as to perform an online

measurement that would not generate awareness-bias (Lovibond & Shanks, 2002). After the response, feedback about the outcome of the trial appeared on screen, see Figure 1.

Five blocks of 72 trials (36 HR and 36 LR) were presented. After each block, participants received information on screen about the total amount earnt during that block. Participants were instructed to grab and count the equivalent amount in 10, 20 and 50 pence coins from the bank box and transfer it to their earnings box. Finally, feedback about the total amount earnt during the conditioning task appeared.

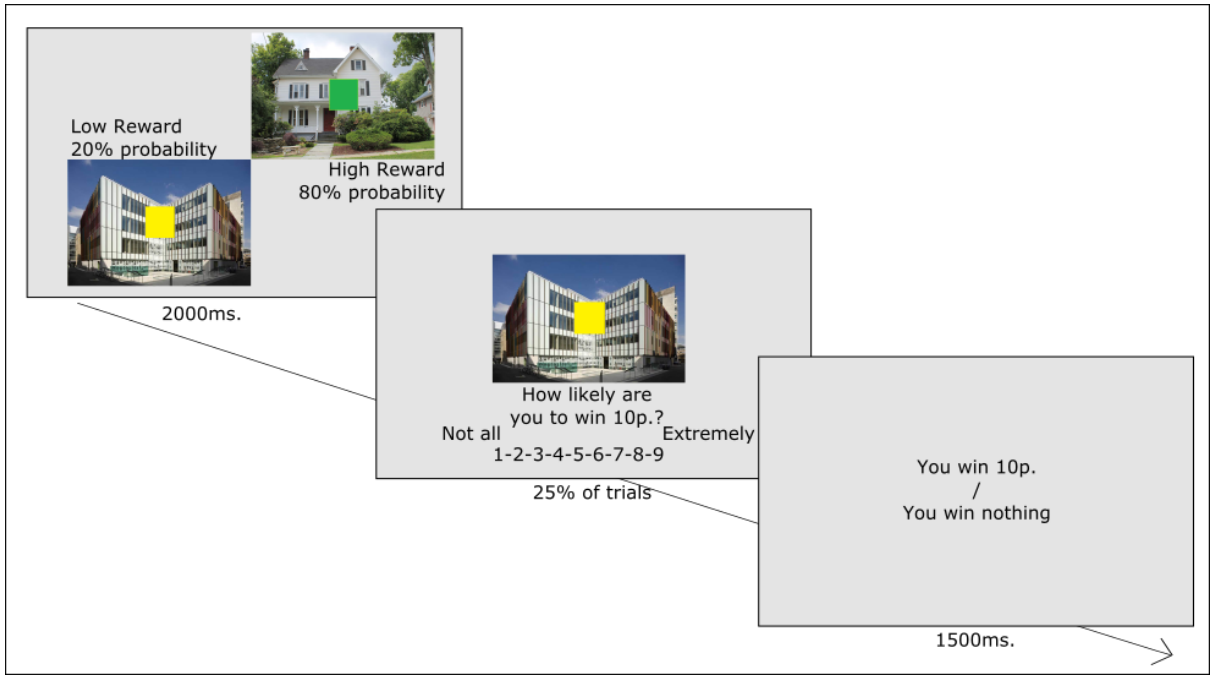


Figure 1: Conditioning procedure used both in experiments 1 and 2. Stimuli depicted were used in experiment 1. *Using a task irrelevant conditioning task stimuli were paired with different probabilities of monetary reward. Expectancy awareness was measured on 25% of the trials.*

The order of trials was pseudorandomized, and the same category of CS (HR/LR) could not appear more than 4 times in a row. The same pseudorandomization was applied to the colour of the overlaid Square. A pseudorandomisation was also applied to CA evaluations so that it could occur every 3, 4 or 5 trials.

4.3.2.6 Pleasantness measurement

Following an Emotional Attentional Blink task (presented elsewhere, (Leganes-Fonteneau et al., 2018)), pleasantness associated with the stimuli was measured. Eighteen pictures of HR and LR from each category were presented in random order and participants had to rate how pleasant each of them was from 1 to 9. Stimuli remained on screen until a response was recorded.

4.3.3 Procedure

After signing the informed consent form, participants completed a series of questionnaires (see questionnaire section) and the two tasks measuring heartbeat interoceptive awareness. The reward-conditioning task was then applied. An Emotional Attentional Blink task was performed at the end of the conditioning in order to examine attentional conditioned responses; however the data from the Emotional Attentional Blink are described elsewhere (Leganes-Fonteneau et al., 2018). Finally, pleasantness ratings for the stimuli (HR and LR) were taken at the end of the experiment. The entire procedure lasted for circa 80 minutes.

4.3.4 Data Analysis

4.3.4.1 Interoceptive awareness

Metacognitive measures are generated by computing the relationship between accuracy and confidence, separately for each of the two tasks.

For the heartbeat tracking task, interoceptive awareness was obtained running a Pearson's correlation between accuracy and confidence for each trial as both accuracy and confidence were measured using a continuous variable (Garfinkel et al., 2015).

For the discrimination task, which provides categorical binary values (accurate/not accurate) for each trial, a receiver operating characteristic (ROC) curve analysis (Green & Swets, 1966) determined the extent to which confidence responses were related to task accuracy (Garfinkel et al., 2015). The ROC curve plots over all possible detection thresholds the hit vs. false alarm rate, providing a numerical measure (area under the ROC) of interoceptive awareness irrespective of positive or negative confidence biases.

4.3.4.2 Group categorizations

As the aim of this experiment was to determine the role of individual differences in metacognitive interoceptive awareness on aspects of Pavlovian conditioning, participants were classified as high or low in interoceptive awareness using a median split (as performed for example on Garfinkel et al., 2015; Suzuki, Garfinkel, Critchley, & Seth, 2013; Tsakiris, Tajadura-Jiménez, & Costantini, 2011). Two distinct categorizations were created, low meta-trackers (LMT) versus high meta-trackers (HMT), and low meta-discriminators (LMD) versus high meta-discriminators (HMD).

The median score in the tracking task was .318. On the basis of the median score HMT (*n*=23, *mean*=.66, *SD*=.20) and LMT (*n*=23, *mean*=-.03, *SD*=.27) groups were generated.

The median score in the discrimination task was .505. On the basis of the median score HMD (*n*=23, *mean*=.63, *SD*=.09) and LMD (*n*=23, *mean*=.43, *SD*=.06) groups were created.

4.3.4.3 Questionnaires and Interoceptive Accuracy and Sensibility

A series of one-way ANOVAS compared scores on the different questionnaires between HMD and LMD groups. That same comparison was carried-out between HMT and LMT groups. A Chi-Square test explored the overlap of participants belonging to both groups.

One-way ANOVAS compared interoceptive accuracy and sensibility and heart rate on both interoception tasks between the groups regarding tracking (HMT vs. LMT) and discrimination (LMD vs. HMD) respectively.

4.3.4.4 Pleasantness

Previous analyses using the same dataset (Leganes-Fonteneau et al., 2018) showed that stimulus category (Houses vs. Buildings) seems to influence the development of pleasantness towards HR stimuli. Houses were consistently evaluated as more pleasant than Buildings during the conditioning task, although an independent sample of participants who rated the pictures for pleasantness in each category did not find differences between them. For that reason, in a Mixed 3-way ANOVA analysing the effect of interoceptive awareness on pleasantness ratings, we included stimulus category (whether for each participant a HR was a House or a Building) and

discrimination awareness group [(HMD vs. LMD) or (HMT and LMT)] as between factors with stimulus type (HR vs. LR) as within factor. Mean heartbeat was included as covariate for the between group interactions. Difference scores between expectancy towards HR and LR stimuli were included also as a covariate to account for CA effects.

4.3.4.5 Contingency Awareness

A Mixed 2-way ANOVA analysed the effect of interoceptive awareness on expectancy ratings, with stimulus type (HR vs. LR) as within, and discrimination awareness group (HMD vs. LMD) as between subjects' factors. Mean heartbeat was included as covariate for the interaction. An equivalent ANOVA was run for the HMT and LMT groups.

Post-hoc paired samples t-test compared scores towards HR and LR stimuli within group when necessary.

4.3.5 Results

4.3.5.1 Questionnaires and Interoceptive Accuracy and Sensibility

The Chi-Square test did not show any significant association between both classifications, $\chi^2(1)=0.087$, $p=.768$. See Table 1 for participant distribution between groups.

		Meta-discrimination		
		High	Low	Total
Meta-tracking	High	12	11	23
	Low	11	12	23
	Total	23	23	46

Table 1: *Categorization of participants in different groups depending on their scores on meta-tracking and meta-discrimination, Experiment 1.*

There were no age or gender differences either between LMT and HMT or LMD and HMD. The groups also were not significantly different in terms of their scores on alcohol consumption, AUDIT, impulsivity or mood. Only the BAS scores were different between LMD and HMD, however this difference did not survive corrections for multiple comparisons (see Table 2).

Discrimination task	**High-Meta group**		***n*=23**	**Low- Meta group**		***n*=23**	
	Mean	*SD*		*Mean*	*SD*	*F*(1,46)	*p*
One-way ANOVA							
Age	23.04	7.56		24.65	8.63	0.452	.51
BIS-11	2.04	0.39		2.20	0.44	1.513	.23
AUDIT	6.57	4.52		9.04	7.28	1.922	.17
Binge score	18.19	13.48		20.70	20.76	0.236	.63
AUQ score	25.79	17.32		32.53	30.03	0.87	.36
PANAS Positive	2.82	0.77		2.87	0.89	0.053	.82
PANAS Negative	1.53	0.51		1.49	0.45	0.06	.81
BAS	2.81	0.37		3.07	0.42	4.88	.03
BIS	2.80	0.32		2.83	0.32	0.107	.75
Porges	2.74	0.52		2.76	0.59	0.018	.89

Tracking Task	**High-Meta group**		***n*=23**	**Low Meta group**		***n*=23**	
	Mean	*SD*		*Mean*	*SD*	*F*(1,45)	*p*
Age	25.43	9.64		22.26	5.89	1.82	.18
BIS-11	2.05	0.36		2.19	0.46	1.21	.28
AUDIT	6.78	5.60		8.83	6.57	1.29	.26
Binge Score	17.98	17.87		20.91	17.10	0.32	.57
AUQ score	27.63	26.54		30.69	22.72	0.18	.68

PANAS Positive	2.90	0.84	2.80	0.82	0.17	.69
PANAS Negative	1.44	0.47	1.57	0.49	0.85	.36
BAS	2.89	0.44	2.99	0.39	0.72	.40
BIS	2.83	0.33	2.81	0.31	0.04	.84
Porges	2.74	0.52	2.76	0.58	0.02	.88

Table 2: *Results of questionnaire analyses for median splits on metacognitive interoception for discrimination and tracking tasks, Experiment 1.*

There were no significant differences in terms of accuracy, confidence and heart rate between high and low meta trackers or discriminators, see Table 3.

		High-meta group		**Low-meta group**			
		Mean	*SD*	Mean	*SD*	*F*(1,45)	*p*
Tracking	***Accuracy***	0.61	.20	0.51	.26	2.109	.154
	Confidence	0.47	.23	0.40	.20	1.425	.239
	Heart rate	77.43	10.59	82.07	9.66	2.399	.129
						F(1,45)	*p*
Discrimination	***Accuracy***	0.50	0.15	0.48	0.14	0.162	.69
	Confidence	0.52	0.15	0.53	0.19	0.023	.88
	Heart rate	78.34	8.79	78.51	9.91	0.004	.95

Table 3: *Descriptives and statistics for the comparison of interoceptive accuracy and confidence as well as heart-rate between their respective metacognitive groups in Experiment 1.*

4.3.5.2 Pleasantness

There was no significant main effect of stimulus type on pleasantness measurements, $F(1,41)=0.503$, $p=.482$.

As predicted, a significant interaction was found between meta-discrimination group and stimulus type, $F(1,41)=5.000$, $p=.031$, Fig. 2, with HMD experiencing more pleasantness towards HR than LR, and LMD showing the opposite pattern (although post-hoc paired t-tests comparing HR and LR stimuli within groups were non-significant *ps*>.13).

There was no significant 3-way interaction between stimulus type, category and meta-discrimination group, $F(1,41)=0.951$, $p=.335$, no main effect of group, $F(1,41)=0.175$, $p=.678$ and no interaction between stimulus type and the heart rate covariate, $F(1,41)=0.593$, p=.446.

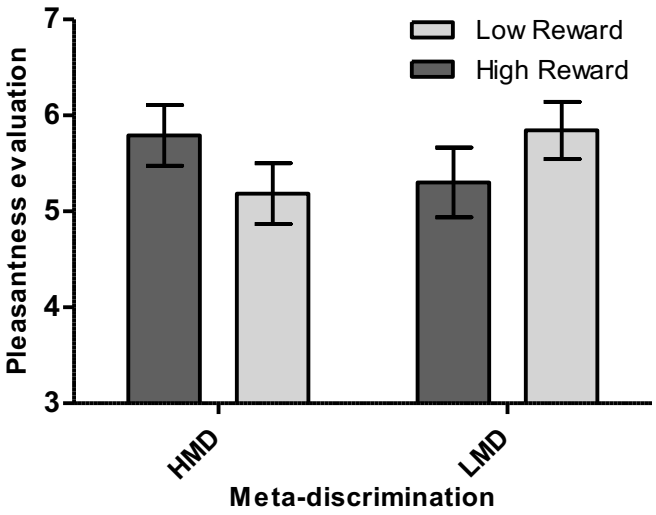


Figure 2: *Pleasantness ratings towards High reward and Low reward stimuli depending on meta-discrimination group, F(1,41)=5.000, p=.031, Experiment 1.*

With regard to the tracking task, contrary to our prediction, there was no significant interaction between stimulus type and group (HMT vs LMT; $F(1,41)=0.717$, $p=.402$), or between stimulus type, group and stimulus category, $F(1,41)=0.121$, $p=.730$. No main effect of group was found either, $F(1,41)=1.223$ $p=.275$, and no interaction between stimulus type and the heart rate covariate, $F(1,41)=0.858$, p=.360.

As previously reported (Leganes-Fonteneau et al., 2018), there was a Two-way interaction between stimulus type and stimulus category, $F(1,41)=11.415$, $p=.002$, with Houses when paired with high probabilities of reward, HR (*mean*=6.05, *SD*=1.47) being rated more pleasant than LR (*mean*=5.02, *SD*=1.29), $t(20)=2.687$, $p=.014$; Buildings on the other hand were rated as marginally less pleasant when paired with high probabilities of reward HR [HR (*mean*=5.24, *SD*=1.79), $t(26)=-1.933$, $p=.064$ than LR (*mean*=6.00, *SD*=.335)].

4.3.5.3 Contingency Awareness

There was no main effect of stimulus type on expectancy ratings, $F(1,43)=0.349$, $p=.558$, but

regarding the tracking split there was a significant interaction between group and stimulus type, $F(1,43)=4.231$, $p=.046$, with HMT showing higher expectancy ratings for HR than for LR, $t(22)=2.212$, $p=.038$. LMT showed no significant differences in their ratings between HR and LR, $t(22)=0.176$, $p=.862$, see Figure 3.

No main effect of group was found, $F(1,43)=0.322$, $p=.573$, and no interaction between stimulus type and the heart rate covariate, $F(1,43)=0.725$, p=.399.

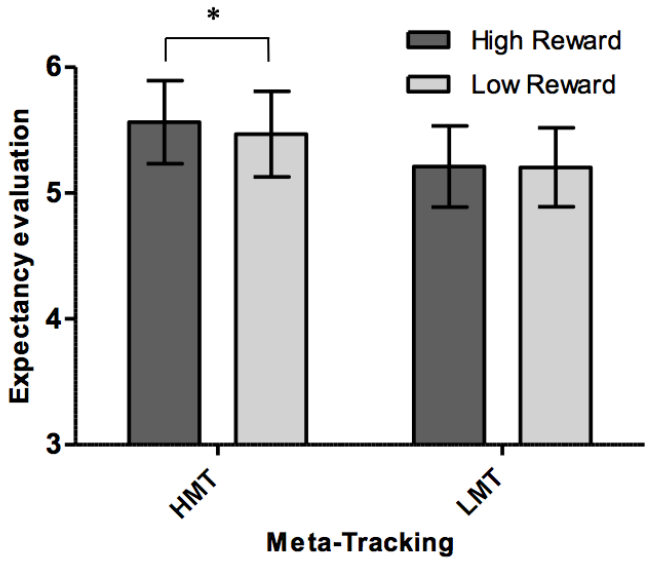


Figure 3: *Expectancy ratings towards HR and LR stimuli depending on meta-tracking group. * High meta-trackers show increased expectancy evaluations towards HR stimuli compared to LR, t(22)=2.212, p=.038, Experiment 1.*

There was no significant interaction between discrimination awareness group and stimulus type, $F(1,43)=.444$, $p=.509$. There was a marginal main effect of group, $F(1,43)=3.815$, $p=.057$, and no interaction between stimulus type and the heart rate covariate, $F(1,43)=0.443$, p=.509.

4.3.6 Discussion Experiment 1

Experiment 1 shows, in accordance with our predictions, that metacognitive interoceptive awareness in the heartbeat discrimination task has a role in the development of pleasantness towards the stimulus associated with high reward (HR). Contrary to our predictions, such an effect was not found with the metacognitive interoceptive awareness in the heartbeat tracking task. On the other hand,

participants with HMT compared to LMT, were better able to discriminate outcome-contingencies correctly, showing the relevance of interoceptive abilities in reward prediction. These findings point to possible differences in interoceptive abilities as measured with the heartbeat tracking and discrimination task.

It is worth noting that the conditioning task itself failed to generate increased pleasantness for HR stimuli over LR stimuli. The fact that Houses when associated with high reward were found more pleasant compared to LR stimulus may be explained as follows: It is possible that pre-existing differences in the intrinsic nature of the stimuli (Houses as naturalistic conditioned stimuli may be preferable to Buildings), permitted the repeated association of Houses with high reward to increase pleasantness, while the lower pre-exiting preference for Buildings prevented the same effect occurring for those stimuli.

4.4 Experiment 2

4.4.1 Aims

In order to address the limitations on Experiment 1, specifically the lack of a main effect of stimulus type on pleasantness evaluation, the second experiment used abstract stimuli instead of Houses and Buildings. Using two different categories of geometrical shapes, Squares and Octagons, we aimed to prevent the development of pleasantness ratings from being modulated by the intrinsic differential characteristic of naturalistic stimuli and thus to more accurately examine the role of interoception in pleasantness development.

Subjective measurements of expectancy assess whether participants have developed knowledge about the rules governing stimulus-outcome contingencies. However, within this measure, it is not evaluated whether participants have developed conscious awareness about those rules. For that purpose, we added confidence ratings after each expectancy awareness evaluation to develop a measure of metacognitive contingency awareness (Barrett, Dienes, & Seth, 2013; Dienes & Perner, 1999); measurements of CA were changed to a dichotomous scale to avoid response transfers between confidence and expectancy ratings. The aim was to investigate in detail whether interoceptive awareness explains CA development, metacognitive-awareness development, or both. Probabilities of reward were changed in order to facilitate the development of metacognitive contingency awareness.

4.4.2 Methods

4.4.2.1 Participants

Sixty participants took part in this experiment; however, data from one of the participants was discarded due to a failure in interoceptive measurements. A total of 59 Sussex University Students (52 females, *mean* age= 22.7, *SD*= 3.78) were considered for data analysis. Participants completed all the questionnaires as in Experiment 1. This study was granted ethical approval by the University of Sussex Life Sciences ethics committee.

4.4.2.2 Apparatus and stimuli

A set of 72 geometrical stimuli belonging to two different categories were developed using InkScape software. These geometrical stimuli were created in order to obtain stimuli devoid of any intrinsic value properties and hence to maximise the generation of emotional reactions congruent solely with the conditioning procedure. Stimuli consisted of 36 Squares and 36 Octagons and were filled with different motifs of stripes, see Figure 4, for a detailed description of stimuli, see (Leganes-Fonteneau et al., 2018). The stimuli appeared overlaid on neutral landscape pictures for all tasks.

Figure 4: Stimuli used for the conditioning in experiment 2. *36 Octagons (A) and 36 Squares (B) were chosen for Experiment 2 to replace the Houses and Buildings used in Experiment 1. All stimuli appeared overlaid on a picture showing a neutral landscape. This was set in order to match the visual characteristics of the stimuli series presented*

during the Emotional Attentional Blink protocol administered between the two experiments.

4.4.3 Procedure

4.4.3.1 Conditioning task

A task equal (besides the changes reported) to Experiment 1 was used with geometrical stimuli as CS. The probability of obtaining 10p. was increased for HR trials from 80% to 90%, and decreased for LR from 20% to 10%.

Expectancy awareness evaluation was also modified. A dichotomous question ("Will you get money?", Yes/No) appeared on the same number of trials (25%) as the outcome expectancy question in experiment 1. In addition, participants had to indicate their level of confidence in their judgment via a Likert scale (1. "completely guessing", 2. "more or less guessing", 3. "fairly sure", 4. "almost certain", 5. "completely certain"). Measuring expectancy awareness this way allowed indexes of metacognitive contingency awareness to be determined.

4.4.3.2 Pleasantness measurement

Pleasantness ratings for HR and LR stimuli were obtained following the same procedure as Experiment 1. Pleasantness was measured immediately after the conditioning task.

4.4.4 Data analysis

4.4.4.1 Interoceptive Awareness

Following the procedure of Experiment 1, participants were separated into two groups according to their levels of metacognitive interoceptive awareness. Median values were comparable to those obtained in Experiment 1. For the heartbeat discrimination task, 29 participants were labelled as HMD if their scores were above the median 0.5 (*mean*=.63 *SD*=.098) and 30 as LMD (*mean*=.36, *SD*=.108). Using scores from the heartbeat tracking task, 28 participants with scores higher than .23 (median) were categorized as HMT (*mean*=.59, *SD*=.196) and the rest as LMT (*mean*=-.24, *SD*=.36).

4.4.4.2 Questionnaires and Interoceptive Accuracy and Sensibility

As in Experiment 1, one-way ANOVAS were applied to compare the relevant data between HMD and LMD as well as between HMT and LMT.

4.4.4.3 Pleasantness

The role of interoception on the development of pleasantness was examined using a 2-way mixed ANOVA with stimulus type (HR vs. LR) as within, and interoceptive-awareness group (HMT vs. LMT and HMD vs LMD) as between subjects' factors. Mean heartbeat was included as a covariate in the analysis of between group interactions. CA, measured with Type I d' scores (see below), was introduced as a covariate in the effect of interoception in hedonic responses.

4.4.4.4 Contingency Awareness

CA was determined from the overall accuracy on the predictive task; this was necessary as, in this experiment, expectancy awareness was measured via a binary forced choice test (Yes/No). Type I d' scores considering the rate of Hits (accurately predicting a HR trial), Correct Rejections (accurately predicting a LR trial), False Alarms and Misses were calculated for each participant, indicating their ability to anticipate trial outcomes accounting for response biases.

In order to obtain numerical estimates of metacognitive CA levels, ROC scores were computed for each participant (Fleming & Lau, 2014). ROC analyses indicate the extent to which confidence ratings on expectancy measurements are predictive of accuracy in contingency discrimination. These scores were calculated separately for HR and LR trials to examine metacognition for both types of stimuli.

A 2-way ANOVA with stimulus type (HR vs. LR) as within and interoceptive-awareness (HMT vs. LMT and HMD vs LMD) as between subjects' factors examined the development of metacognitive awareness of stimulus-outcome contingencies based on ROC scores. Furthermore, one-way ANOVAS compared Type I d' scores between interoceptive groups (HMT vs. LMT and HMD vs LMD).

Mean heartbeat during interoception measurement was included as a covariate for all between group interactions.

4.4.5 Results

4.4.5.1 Questionnaires and Interoceptive Accuracy and Sensibility

The Chi-Square test did not show any significant association between both classifications, $\chi^2(1)=0.416$, $p=.519$, See table 4 for frequency categorization of each group.

		Meta-discrimination		
		High	Low	Total
Meta-tracking	High	15	13	28
	Low	14	17	31
	Total	29	30	59

Table 4: *Categorization of participants in different groups depending on their scores on meta-tracking and meta-discrimination, Experiment 2.*

There were no significant differences between groups in any of the questionnaire scores or on accuracy and confidence ratings in the interoceptive tasks, (see Table 5 and 6 respectively).

Discrimination task	**High-Meta group**		***n*=29**	**Low-Meta group**		***n*=30**	
	Mean	*SD*		*Mean*	*SD*	*F*(1,58)	*p*
One-way ANOVA							
Age	20.38	3.88		20.73	2.97	.156	.69
BIS-11	2.08	0.28		2.04	0.29	.386	.54
AUDIT	6.62	4.59		6.30	5.13	.064	.80
Binge Score	19.13	16.97		17.28	16.21	.104	.67
AUQ score	30.19	25.14		30.08	25.99	.000	.99
PANAS Positive	2.75	0.71		3.13	0.83	3.507	.07
PANAS Negative	1.43	0.43		1.58	0.57	1.296	.26
BAS	2.07	0.27		2.06	0.32	.001	.98
BIS	1.63	0.54		1.88	0.51	3.063	.09
Porges	2.60	0.72		2.43	0.60	.925	.34

Tracking Task	**High-Meta group**		***n*=28**	**Low-Meta group**		***n*=31**	
	Mean	*SD*		*Mean*	*SD*	*F*(1,58)	*p*
Age	20.46	2.86		20.65	3.90	.040	.84
BIS-11	2.01	0.33		2.11	0.23	2.064	.16
AUDIT	5.50	5.43		7.32	4.13	2.133	.16
Binge Score	16.89	17.42		19.36	15.77	.327	.57
AUQ score	26.56	28.20		33.36	22.46	1.059	.31
PANAS Positive	3.03	0.86		2.86	0.73	.599	.44
PANAS Negative	1.45	0.50		1.55	0.51	.677	.41
BAS	2.10	0.25		2.03	0.33	.935	.34
BIS	1.74	0.66		1.77	0.41	.054	.82
Porges	2.47	0.73		2.55	0.61	.216	.64

Table 5: *Results of questionnaire analyses for median splits on Meta-cognitive interoception for discrimination and tracking tasks, Experiment 2.*

		High-meta group		Low-meta group			
		Mean	*SD*	Mean	*SD*	*F*(1,58)	*p*
Tracking	***Accuracy***	.49	.32	0.31	1.03	-.794	.38
	Confidence	0.41	.19	0.38	.21	.356	.55
	Heart rate	77.29	14.58	85.52	14.76	4.625	.036
						F(1,58)	p
Discrimination	***Accuracy***	0.49	.12	0.52	.15	.704	.41
	Confidence	0.50	.16	0.52	.18	.083	.78
	Heart rate	76.83	11.64	79.43	12.69	.728	.40

Table 6: *Descriptives and statistics for the comparison of interoceptive accuracy and confidence and heart-rate between their respective metacognitive groups in Experiment 2.*

4.4.5.2 Pleasantness

There was a main effect of stimulus type on pleasantness development, $F(1,56)=5.087$, $p=.028$, with HR (*mean*=.53, *SD*=.182) rated as more pleasant than LR (*mean*=.49, *SD*=.187). There was also a significant interaction between meta-discrimination group and stimulus type $F(1,55)=5.305$, $p=.025$, with HMD experiencing more pleasantness towards HR than LR (post-hoc t-test, $t(28)=2.689$, $p=.012$) and LMD showing no differential response between HR and LR, $t(29)=0.768$, $p=.449$, see Figure 5. There was no main effect of discrimination group on overall pleasantness, $F(1,55)=.810$, $p=372$ and no interaction between stimulus type and the heart rate covariate, $F(1,55)=0.676$, p=.414 or Type 1d' scores, $F(1,55)=0.981$, p=.326.

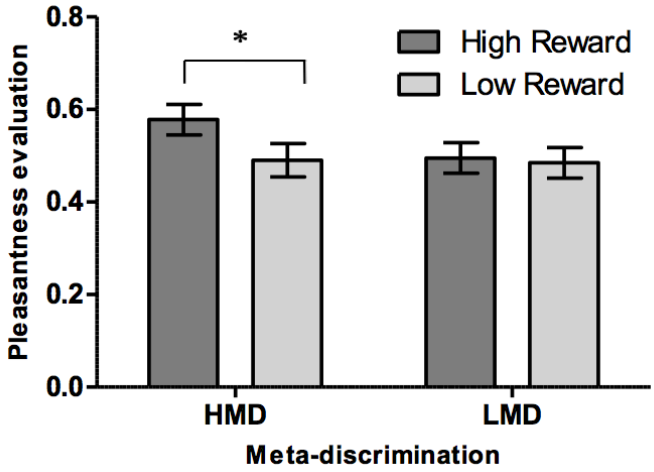


Figure 5: *Pleasantness evaluations of High reward and Low reward stimuli depending on meta-discrimination group. * for High meta-discriminators, High reward stimuli were more pleasant than LR, t(28)=2.689, p=.012.*

There was no main effect $F(1,55)=.786$, $p=.379$ of group, or interaction, $F(1,55)=1.507$, $p=.225$, involving the meta-tracking groups, and no interaction between stimulus type and the heart rate covariate, $F(1,55)=0.189$, p=.666, or Type 1d' scores, $F(1,55)=0.750$, p=.390.

4.4.5.3 Contingency Awareness

There was no main effect of stimulus type on ROC scores, $F(1,56)=0.855$, $p=.359$. There was no significant interaction between stimulus type and meta-discrimination, $F(1,56)=.251$, $p=.618$, no main effect of group, $F(1,56)=.052$, $p=.820$, and no interaction between stimulus type and HR as a covariate, $F(1,56)=.867$, $p=.356$.

Regarding the meta-tracking categorization, there was no significant interaction between stimulus type and group, $F(1,56)=.112$, $p=.739$, no main effect of group,

F(1,56)=.030, *p*=.862, and no interaction between stimulus type and HR as a covariate, *F*(1,56)=.809, *p*=.732.

In addition, comparisons of Type-I d' scores between groups showed no effect of interoceptive categorization according to meta-discrimination, *F*(1,56)=0.35, *p*=.852, or meta-tracking groups, *F*(1,56)=0.054, *p*=.817.

4.4.6 Discussion Experiment 2

The finding in the present experiment that HR conditioned stimuli were evaluated as more pleasant than LR supports the proposal that abstract stimuli used in a conditioning procedure can refine the development of conditioned responses. The finding that participants with high scores in metacognitive interoceptive awareness for the heartbeat discrimination task experience more pleasantness for HR stimuli compared to LR, replicates the finding in Experiment 1 and supports the principle that interoceptive awareness can modulate emotional conditioned responses. The lack of an effect of metacognitive interoceptive awareness for the heartbeat tracking task on pleasantness ratings, as found in Experiment 1, serves as further support that these two interceptive tasks may be accessing different aspects of interoceptive functioning, subserved by different neural structures (Schulz, 2016). Indeed, in a normative sample (*n*=80) metacognitive performance across these two interoceptive tasks was not found to correlate (Garfinkel et al., 2015).

In the present experiment, interoception did not modulate the development of metacognitive contingency awareness; this could be attributed to the fact that very few participants (*n*=3 out of 60) (Leganes-Fonteneau et al., 2018) developed

metacognitive contingency awareness, likely due to the use of a task-irrelevant conditioning.

4.5 General Discussion

The aim of the present study was to investigate the role interoception has on appetitive Pavlovian conditioning, specifically, to determine how high interoception modulates the development of hedonic responses towards stimuli associated with a monetary reward. We also examined how interoceptive abilities facilitate accurate predictions of reward. In line with our hypotheses, the current findings highlight the relevance of interoception in the development of associative learning and conditioned emotional responses.

Although previous research had investigated the role of interoception in aversive conditioning (i.e. Katkin, Wiens, & Ohman, 2001) and emotional learning (i.e. Pfeifer et al., 2017; Pollatos & Schandry, 2008) using an implicit approach (Werner, Peres, Duschek, & Schandry, 2010), the present report is the first, to our knowledge, that examines individual differences in interoception and their relationship to appetitive conditioning.

In both experiments, our findings show that participants with high metacognitive interoceptive awareness in the heartbeat discrimination task rate HR stimuli as more pleasant compared to LR. This demonstrates the role of interoception in the development of emotional responses, in line with our predictions based upon work demonstrating the modulation of emotional processing as a function of interoception (Pollatos & Schandry, 2008). Importantly, these results are obtained accounting for

knowledge of contingencies, showing the unique role that interoception has in the development of emotional responses towards CS.

In addition, participants with high metacognitive interoceptive awareness on the tracking task were better able to predict the occurrence of rewards (albeit only in Experiment 1 in which reward outcome expectancy was evaluated with a contiunuous Likert scale and not with a yes/no dichotomous choice). Thus, when probing reward expectancies using a continuous measure, a more sensitive measure of CA (Lovibond & Shanks, 2002), metacognitive interoception on the heartbeat tracking task can be shown to modulate reward prediction. This supports previous data documenting that interoception is related with enhanced ability to predict the occurrence of aversive stimuli (Katkin et al., 2001).

Our second experiment also investigated the role of interoception in metacognitive CA for the outcome predicted by the stimuli. Our results, however, did not bring support for this hypothesis. Several factors could account for this. In particular, the levels of metacognitive CA may have not been high enough to study the effect of interoception, thus a floor effect may have obscured any potential relationship. Previous analyses of the data show very low metacognitive CA in participants, irrespective of their ability to predict rewards (Leganes-Fonteneau et al., 2018).

On the other hand, no effects on Type I d' scores were found. This may be due to changes in the way CA was evaluated, from a Likert scale to a dichotomous 'yes/no' measurement accompanied by an evaluation of confidence. That factor might have engaged cognitive processes unrelated to interoception. Most importantly, the

probabilities of reward were also more extreme on Experiment 2 than on Experiment 1 (90-10% instead of 80-20%) making the development of Type I CA much easier (Leganes-Fonteneau et al., 2018). The fact that predictions were easier to develop in this experiment may decrease the relevance of the interoceptive appraisal of autonomous reactions to CS, explaining why in Experiment 2, HMT and LMT groups did not differ in their ability to predict rewards.

Our data suggest that different measures of interoception may correspond with different aspects of learning. In this case, metacognitive discrimination modulated the development of hedonic responses whereas metacognitive tracking modulated the accuracy in reward prediction. The nature of the interoceptive tasks differs, with the discrimination task requiring the integration of an external stimulus (the beep) with cardiac functioning, whereas the tracking task is self-focused as participants concentrate on counting their own heartbeat (Garfinkel, Tiley, et al., 2016; Garfinkel et al., 2015). It is possible that assessing whether a stimulus will predict rewards or not requires the perception of autonomic states to guide “gut-feelings” (Kandasamy et al., 2016). Such self-focused attention is augmented during tracking measures. On the other hand, evaluating stimulus pleasantness requires observing the stimulus (being attentive) and coupling its characteristics with its effects on the autonomic nervous system, and this internal-external integration is a core feature of the discrimination task. This explanation is supported by recent data indicating a differential effect of Oxytocin on the two tasks (Betka et al, 2018). Importantly, the current report demonstrates that metacognitive dimensions of interoception play a crucial role in emotional responses to appetitive conditioned stimuli.

Our findings add to previous reports demonstrating the role of interoception in other types of emotional responses. For instance, Werner and colleagues (2010) found that successful completion of previously presented emotional word stems was enhanced in highly interoceptive participants. They also found a correlation between skin conductance responses to emotional words and interoceptive accuracy. Furthermore, Pollatos and Schandry (2008) showed that interoception modulates bodily responses in the presence of emotional stimuli. Thus, during the appetitive conditioning task, autonomous responses elicited by reward presentation may generate specific physiological states which become associated with CS, and high interoceptive abilities would serve to amplify the perception of those physiological changes, hence impacting the hedonic attributes of stimuli (Garfinkel et al., 2014; Paulus, Tapert, & Schulteis, 2009). When stimuli are presented for pleasantness evaluation, those values may be retrieved again more easily due to interoceptive amplification (Gray & Critchley, 2007).

This interpretation goes in line with early learning theories integrating physiological responses within the development of hedonic reactions (Bindra, 1978; Toates, 1986).

Reward learning is capital for the understanding of addictive processes (Berridge, 2000; Stewart et al., 1984), and numerous research points towards a series of interoceptive mechanisms explaining addiction (Paulus & Stewart, 2014; Verdejo-Garcia et al., 2012).

It seems therefore that interoception plays a role in the development of conditioned responses, which could be translated to the involvement of bodily sensations in

maladaptive learning processes, a key characteristic of addiction (Koob & Volkow, 2016).

A limitation of our study is that we did not measure the aforementioned physiological reactivity in the presence of stimuli. Examining the relationship between interoception, autonomic responses generated by CS, and emotional reactivity would strengthen and improve our understanding of the role of physiological states in appetitive conditioning and their relevance to drug-addiction. Recent research (Stewart et al., 2015) shows that individuals who had recently progressed to problematic stimulant use presented with increased insular activation during pleasant interoceptive stimulation (applying soft touch). This enhanced interoceptive reactivity possibly leading to maladaptive learning could partially explain the aetiology and maintenance of addictive behaviours.

Moreover, findings in Experiment 1 also highlight the importance of using neutral abstract stimuli (like Octagons and Squares) instead of Houses or Buildings. Neutral stimuli devoid of any intrinsic value may prevent uncontrollable associations with reward outcomes.

Interestingly, the described effects were obtained using the metacognitive measure of interoception. This is a relatively understudied dimension of interoception (Canales-Johnson et al., 2015; Forkmann et al., 2016; Garfinkel et al., 2015), though our data suggest that it might display a heightened relationship to mechanisms underlying reward Pavlovian conditioning.

The present research adds to our understanding of the interoceptive mechanisms underlying emotional appraisals, with a particular focus on appetitive conditioning; supporting further the relationship between interoception and addictive behaviours. More research is needed to delineate the exact processes underlying this relationship, to better understand the complex interaction between interoception and addiction.

4.6 Declaration of interests

This work was supported by the University of Sussex, UK. No potential conflict of interest was reported by the authors.

4.7 References

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–71. http://doi.org/10.1073/pnas.1104047108

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. http://doi.org/10.1037/a0033268

Berridge, K. C. (2000). Reward learning: Reinforcement, incentives, and expectations. *Psychology of Learning and Motivation*, *40*, 223–278. http://doi.org/10.1016/S0079-7421(00)80022-5

Bindra, D. (1978). How adaptive behavior is produced: a perceptual-motivational alternative to response reinforcements. *Behavioral and Brain Sciences*, *1*(01), 41. http://doi.org/10.1017/S0140525X00059380

Cameron, O. G. (2001). Interoception: The Inside Story—a Model for Psychosomatic Processes. *Psychosomatic Medicine*, *63*(5), 697–710.

Canales-Johnson, A., Silva, C., Huepe, D., Rivera-Rei, Á., Noreika, V., Garcia, M. del C., … Bekinschtein, T. A. (2015). Auditory Feedback Differentially Modulates Behavioral and Neural Markers of Objective and Subjective Performance When Tapping to Your Heartbeat. *Cerebral Cortex*, *25*(11), 4490–4503. http://doi.org/10.1093/cercor/bhv076

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333. http://doi.org/10.1037/0022-3514.67.2.319

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*(2), 189–95. http://doi.org/10.1038/nn1176

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*(05). http://doi.org/10.1017/S0140525X99002186

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., … Dalgleish, T. (2010). Listening to Your Heart. *Psychological Science*, *21*(12), 1835–1844. http://doi.org/10.1177/0956797610389191

Field, M., Mogg, K., Zetteler, J., & Bradley, B. P. (2004). Attentional biases for alcohol cues in heavy and light social drinkers: the roles of initial orienting and maintained attention. *Psychopharmacology*, *176*(1), 88–93. http://doi.org/10.1007/s00213-004-1855-1

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. http://doi.org/10.3389/fnhum.2014.00443

Forkmann, T., Scherer, A., Meessen, J., Michal, M., Schächinger, H., Vögele, C., & Schulz, A. (2016). Making sense of what you sense: Disentangling interoceptive awareness, sensibility and accuracy. *International Journal of Psychophysiology*, *109*, 71–80. http://doi.org/10.1016/J.IJPSYCHO.2016.09.019

Garavan, H. (2010). Insula and drug cravings. *Brain Structure & Function*, *214*(5–6), 593–601. http://doi.org/10.1007/s00429-010-0259-8

Garfinkel, S. N., & Critchley, H. D. (2013). Interoception, emotion and brain: new insights link internal physiology to social behaviour. Commentary on:: “Anterior insular cortex mediates bodily sensibility and social anxiety” by Terasawa et al. (2012). *Social Cognitive and Affective Neuroscience*, *8*(3), 231–4. http://doi.org/10.1093/scan/nss140

Garfinkel, S. N., Manassei, M. F., Hamilton-Fletcher, G., In den Bosch, Y., Critchley, H. D., & Engels, M. (2016). Interoceptive dimensions across cardiac and respiratory axes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708), 20160014. http://doi.org/10.1098/rstb.2016.0014

Garfinkel, S. N., Minati, L., Gray, M. A., Seth, A. K., Dolan, R. J., & Critchley, H. D. (2014). Fear from the heart: sensitivity to fear stimuli depends on individual heartbeats. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(19), 6573–82. http://doi.org/10.1523/JNEUROSCI.3507-13.2014

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. http://doi.org/10.1016/j.biopsycho.2014.11.004

Garfinkel, S. N., Tiley, C., O’Keeffe, S., Harrison, N. A., Seth, A. K., & Critchley, H. D. (2016). Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety. *Biological Psychology*, *114*, 117–126. http://doi.org/10.1016/J.BIOPSYCHO.2015.12.003

Gray, M. A., & Critchley, H. D. (2007). Interoceptive Basis to Craving. *Neuron*, *54*(2), 183–186.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *New York: Wiley*.

Hart, N., McGowan, J., Minati, L., & Critchley, H. D. (2013). Emotional Regulation and Bodily Sensation: Interoceptive Awareness Is Intact in Borderline Personality Disorder. *Journal of Personality Disorders*, *27*(4), 506–518. http://doi.org/10.1521/pedi_2012_26_049

James, W. (1884). II.—What is an emotion? *Mind*.

Jeffs, S., & Duka, T. (2017). Predictive but not emotional value of Pavlovian stimuli leads to pavlovian-to-instrumental transfer. *Behavioural Brain Research*, *321*, 214–222. http://doi.org/10.1016/j.bbr.2016.12.022

Kandasamy, N., Garfinkel, S. N., Page, L., Hardy, B., Critchley, H. D., Gurnell, M., & Coates, J. M. (2016). Interoceptive Ability Predicts Survival on a London Trading Floor. http://doi.org/10.1038/srep32986

Katkin, E., Reed, S., & Deroo, C. (1983). A methodological analysis of 3 techniques for the assessment of individual-differences in heartbeat detection.

Katkin, E. S., Wiens, S., & Ohman, A. (2001). Nonconscious Fear Conditioning, Visceral Perception, and the Development of Gut Feelings. *Psychological Science*, *12*(5), 366–370. http://doi.org/10.1111/1467-9280.00368

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., … Zucker, N. (2017). Interoception and Mental Health: A Roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. http://doi.org/10.1016/J.BPSC.2017.12.004

Koob, G. F., & Volkow, N. D. (2016). Neurobiology of addiction: a neurocircuitry analysis. *The Lancet Psychiatry*, *3*(8), 760–773. http://doi.org/10.1016/S2215-0366(16)00104-8

Leganes-Fonteneau, M., Scott, R., & Duka, T. (2018). Attentional responses to stimuli associated with a reward can occur in the absence of knowledge of their predictive values. *Behavioural Brain Research*, *341*, 26–36. http://doi.org/10.1016/j.bbr.2017.12.015

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26.

Mehrabian, A., & Russell, J. A. (1978). A questionnaire measure of habitual alcohol use. *Psychological Reports*, *43*(3), 803–806. http://doi.org/10.2466/pr0.1978.43.3.803

Moeller, S. J., Maloney, T., Parvaz, M. A., Dunning, J. P., Alia-Klein, N., Woicik, P. A., … Goldstein, R. Z. (2009). Enhanced Choice for Viewing Cocaine Pictures in Cocaine Addiction. *Biological Psychiatry*, *66*(2), 169–176. http://doi.org/10.1016/j.biopsych.2009.02.015

Mogg, K., Bradley, B. P., Field, M., & De Houwer, J. (2003). Eye movements to smoking-related pictures in smokers: relationship between attentional biases and implicit and explicit measures of stimulus valence. *Addiction*, *98*(6), 825–836. http://doi.org/10.1046/j.1360-0443.2003.00392.x

Naqvi, N. H., & Bechara, A. (2010). The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. *Brain Structure & Function*, *214*(5–6), 435–50. http://doi.org/10.1007/s00429-010-0268-7

Naqvi, N. H., Gaznick, N., Tranel, D., & Bechara, A. (2014). The insula: a critical neural substrate for craving and drug seeking under conflict and risk. *Annals of the New York Academy of Sciences*, *1316*(1), 53–70. http://doi.org/10.1111/nyas.12415

Naqvi, N. H., Rudrauf, D., Damasio, H., & Bechara, A. (2007). Damage to the insula disrupts addiction to cigarette smoking. *Science (New York, N.Y.)*, *315*(5811), 531–4. http://doi.org/10.1126/science.1135926

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774. http://doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Paulus, M. P., & Stewart, J. L. (2014). Interoception and drug addiction. *Neuropharmacology*, *76 Pt B*, 342–50. http://doi.org/10.1016/j.neuropharm.2013.07.002

Paulus, M. P., Tapert, S. F., & Schulteis, G. (2009). The role of interoception and alliesthesia in addiction. *Pharmacology, Biochemistry, and Behavior*, *94*(1), 1–7. http://doi.org/10.1016/j.pbb.2009.08.005

Pfeifer, G., Garfinkel, S. N., Gould van Praag, C. D., Sahota, K., Betka, S., & Critchley, H. D. (2017). Feedback from the heart: Emotional learning and memory is controlled by cardiac cycle, interoceptive accuracy and personality. *Biological Psychology*, *126*, 19–29. http://doi.org/10.1016/j.biopsycho.2017.04.001

Pollatos, O., Gramann, K., & Schandry, R. (2007). Neural systems connecting interoceptive awareness and feelings. *Human Brain Mapping*, *28*(1), 9–18. http://doi.org/10.1002/hbm.20258

Pollatos, O., & Schandry, R. (2008). Emotional processing and emotional memory are modulated by interoceptive awareness. *Cognition & Emotion*, *22*(2), 272–287. http://doi.org/10.1080/02699930701357535

Porges, S. (1993). Body perception questionnaire. *Laboratory of Developmental Assessment, University of Maryland*.

Raes, A. K., & De Raedt, R. (2011). Interoceptive awareness and unaware fear conditioning: are subliminal conditioning effects influenced by the manipulation of visceral self-perception? *Consciousness and Cognition*, *20*(4), 1393–402. http://doi.org/10.1016/j.concog.2011.05.009

Saunders, J. B., Aasland, O. G., Babor, T. F., De, J. R., Fuente ', L., & Grant ', M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO

Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption—II. *Addiction*, *88*, 791–804.

Schandry, R. (1981). Heart Beat Perception and Emotional Experience. *Psychophysiology*, *18*(4), 483–488. http://doi.org/10.1111/j.1469-8986.1981.tb02486.x

Schulz, S. M. (2016). Neural correlates of heart-focused interoception: a functional magnetic resonance imaging meta-analysis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708), 20160018. http://doi.org/10.1098/rstb.2016.0018

Sherrington, C. S. (1948). The integrative action of the nervous system, (Cambridge Univ. Pess, Cambridge, UK).

Sönmez, M. B., Kılıç, E. K., Çöl, I. A., Görgülü, Y., & Çınar, R. K. (2016). Decreased interoceptive awareness in patients with substance use disorders. *Http://Dx.Doi.Org/10.3109/14659891.2016.1143048*.

Stewart, J., de Wit, H., & Eikelboom, R. (1984). Role of unconditioned and conditioned drug effects in the self-administration of opiates and stimulants. *Psychological Review*, *91*(2), 251–268. http://doi.org/10.1037/0033-295X.91.2.251

Stewart, J. L., May, A. C., Tapert, S. F., & Paulus, M. P. (2015). Hyperactivation to pleasant interoceptive stimuli characterizes the transition to stimulant addiction. *Drug and Alcohol Dependence*, *154*, 264–270. http://doi.org/10.1016/J.DRUGALCDEP.2015.07.009

Suzuki, K., Garfinkel, S. N., Critchley, H. D., & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, *51*(13), 2909–2917. http://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2013.08.014

Terasawa, Y., Shibata, M., Moriguchi, Y., & Umeda, S. (2013). Anterior insular cortex mediates bodily sensibility and social anxiety. *Social Cognitive and Affective Neuroscience*, *8*(3), 259–66. http://doi.org/10.1093/scan/nss108

Toates, F. M. (Frederick M. . (1986). *Motivational systems*. Cambridge University Press.

Townshend, J., & Duka, T. (2002). Patterns of alcohol drinking in a population of young social drinkers: a comparison of questionnaire and diary measures. *Alcohol and Alcoholism*.

Tsakiris, M., Tajadura-Jiménez, A., & Costantini, M. (2011). Just a heartbeat away from one's body: interoceptive sensitivity predicts malleability of body-representations. *Proceedings. Biological Sciences*, *278*(1717), 2470–6. http://doi.org/10.1098/rspb.2010.2547

Verdejo-Garcia, A., Clark, L., & Dunn, B. D. (2012). The role of interoception in addiction: A critical review. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1857–1869. http://doi.org/10.1016/j.neubiorev.2012.05.007

Verdejo-Garcia, A., Perez-Garcia, M., & Bechara, A. (2006). Emotion, Decision-Making

and Substance Dependence: A Somatic-Marker Model of Addiction. *Current Neuropharmacology*, *4*(1), 17–31. http://doi.org/10.2174/157015906775203057

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. http://doi.org/10.1037/0022-3514.54.6.1063

Werner, N. S., Peres, I., Duschek, S., & Schandry, R. (2010). Implicit memory for emotional words is modulated by cardiac perception. *Biological Psychology*, *85*(3), 370–376. http://doi.org/10.1016/J.BIOPSYCHO.2010.08.008

Whitehead, W. E., Drescher, V. M., Heiman, P., & Blackwell, B. (1977). Relation of heart rate control to heartbeat perception. *Biofeedback and Self-Regulation*, *2*(4), 371–392. http://doi.org/10.1007/BF00998623

Wiens, S. (2005). Interoception in emotional experience. *Current Opinion in Neurology*, *18*(4), 442–7.

Wiers, R. W., & Stacy, A. W. (2006). Implicit Cognition and Addiction. *Current Directions in Psychological Science*, *15*(6), 292–296. http://doi.org/10.1111/j.1467-8721.2006.00455.x

Yokoyama, T., Padmala, S., & Pessoa, L. (2015). Reward learning and negative emotion during rapid attentional competition. *Frontiers in Psychology*, *6*, 269. http://doi.org/10.3389/fpsyg.2015.00269

5 Interoceptive awareness is associated with acute alcohol-induced changes in mood states

5.1 Abstract

Interoception, the sensing of bodily signals, is related to emotional reactivity and may contribute to the pathophysiology of addiction. Evidence is accumulating that individuals with alcohol use disorders and other substance-dependences show altered interoceptive abilities, however little is known about the acute effects of alcohol on interoception and how this may influence the perception of drug induced effects.

In a double-blind design, fifty (30 females) healthy young participants were given a beverage containing either a low (0.4g/kg, n=18) or high (0.6g/kg, n=15) alcohol dose or a placebo (n=17). After alcohol administration, participants completed two interoceptive paradigms, the heart-beat tracking and heart-beat discrimination tasks, both assessing different accuracy and metacognitive measures of interoception. Subjective feelings elicited by alcohol administration were also measured.

Participants under the low alcohol dose had decreased metacognitive interoceptive awareness on the discrimination task compared to placebo. Participants under alcohol experienced feelings of light-headedness, which were positively associated with increased interoceptive awareness in the cardiac discrimination task.

These findings provide evidence for an interplay between interoceptive processing and the perception of drug-induced mood changes. Interoceptive awareness of cardiac

discrimination was shown to have an independent contribution in the appraisal of subjective states generated by alcohol administration, suggesting that interoception may play a role in the perception of positive alcohol effects thus contributing to alcohol abuse.

5.2 Introduction

Interoception refers to the neural and mental representation of internal bodily signals (Craig, 2002; Sherrington, 1948). The processing of this information is implicated in the formation of emotional responses (Critchley & Garfinkel, 2017; Dunn et al., 2010). Internal bodily signals are communicated to the brain via afferent pathways and integrated within the insular cortex (Craig, 2002; Critchley & Harrison, 2013; Schulz, 2016). The insular cortex is associated with addictive processes, as demonstrated using a range of techniques including lesion (Naqvi, Rudrauf, Damasio, & Bechara, 2007) and imaging studies (Naqvi & Bechara, 2010).

The role of interoception in addictive behaviours is hypothesized to relate to the perception of bodily sensations induced by substance consumption (Paulus, Tapert, & Schulteis, 2009), where neural areas subserving interoception may also contribute to craving states (Gray & Critchley, 2007). Insula activation reflects the sensing of internal bodily states and physiological changes elicited by drug administration (Verdejo-Garcia, Clark, & Dunn, 2012). This information is then used to extract conscious information about the effects of the drug (Garavan, 2010; Naqvi & Bechara, 2010). Altered interoceptive processes in the context of emotional appraisals could in turn contribute to the development of addictive disorders (Stewart, May, Tapert, & Paulus,

2015). In addition, research shows that individuals with heavy alcohol use disorders have impaired interoceptive accuracy compared to a control group, as demonstrated using a heartbeat tracking task (Ateş Çöl, Sönmez, & Vardar, 2016).

Drug effects encompass strong sensory and mood changes, which can be transformed into interoceptive cues associated with the rewarding properties of drugs. Drug discrimination tasks are used to identify the type of sensations generated by drugs. During drug discrimination procedures, participants initially learn to discriminate a drug given at a low but effective dose from placebo. Once learning is achieved, participants' ability to generalise this discrimination at lower doses of the same drug is tested and the drug effects associated with this ability are evaluated. In an alcohol discrimination task (Duka, Stephens, Russell, & Tasker, 1998) it was shown that administration of low alcohol doses generates subjective feelings of light-headedness, which facilitates the discrimination (and generalisation to lower doses) of the drink consumed. The mechanisms by which drug-discrimination is established may therefore originate in interoceptive processes (Duka, Jackson, Smith, & Stephens, 1999).

In fact, it has long been posited that interoception may mediate the detection of reward effects (i.e. food, substances) (Paulus et al., 2009), determining their hedonic value even in healthy participants (Cabanac, 1979; Toates, 1986), indicating its possible involvement in addictive processes. However, the experimental evidence in this regard is limited. This study constitutes one of the first examinations of the acute effects of alcohol on interoception and its relationship to the perception of subjective drug-induced effects.

Using a double-blind alcohol-placebo experiment we explored the role of interoceptive awareness of cardiac functioning (heartbeat) in the appraisal of alcohol effects. Interoception was measured using the tracking (Schandry, 1981) and discrimination (Katkin, Reed, & Deroo, 1983; Whitehead, Drescher, Heiman, & Blackwell, 1977) tasks, which evaluate different facets of interoceptive processing (Garfinkel *et al*, 2015, 2016; Garfinkel and Critchley, 2013); the tracking task testing the perception of a subject's own heartbeat, and the discrimination task testing the ability of the subject to assess whether a tone is synchronised or not with their own heartbeat (Betka et al., 2018).

It seems that these two interoceptive tasks tap into different cognitive processes (Garfinkel & Critchley, 2013), with the tracking task being based on the observation of internal physiological information, amenable to higher order influences such as knowledge about heartrate (Ring & Brener, 1996); and the discrimination task requiring coupling information proceeding from exteroceptive (the tone) and interoceptive channels (Garfinkel, Tiley, et al., 2016; Garfinkel et al., 2015). Out of these two tasks, indexes of interoceptive accuracy and sensibility (confidence on the response) were extracted (Garfinkel et al., 2015). Importantly, there is a lack of correlation between the measures (Garfinkel et al., 2015; Garfinkel, Manassei, et al., 2016) and, it seems that not all the information required for an accurate performance on the task reaches consciousness (Garfinkel et al., 2015). The study of metacognitive interoceptive awareness, also termed interoceptive insight (Khalsa et al., 2017), can provide information about conscious interoceptive abilities (Garfinkel et al., 2016). Conscious metacognitive interoceptive awareness might constitute a more suitable predictor of the perception of subjective (conscious) drug effects.

In order to control for subjective interoceptive sensibility, we measured also the general ability to perceive bodily functions using the awareness section of the Body Perception Questionnaire (BPQ) (Porges, 1993). Alexithymia, the deficit in the perception of one's own emotions, is related to addictive processes (Kopera et al., 2015; Thorberg, Young, Sullivan, & Lyvers, 2009), partially mediating the relationship between interoception and alcohol consumption (Betka et al., 2018), and was therefore also assessed. Impulsivity relates also to alcohol consumption and alcohol administration can affect impulsivity measures (Caswell, Morgan, & Duka, 2013); for that reason impulsivity as a trait was also measured.

We aimed to observe differences in interoception induced by alcohol administration in a low and a high dose compared to a placebo group. We also aimed to observe a dose-dependent induction of subjective feelings of light-headedness by alcohol, replicating previous results (Duka et al., 1998). We hypothesised that a high ability to consciously perceive internal bodily sensations (metacognitive interception) would facilitate the detection of subjective drug effects.

5.3 Materials and methods

5.3.1 Participants

Fifty students from the University of Sussex (30 females, age range 18-48, mean age 21.79) took part in this experiment. Exclusion criteria were: being below the legal drinking age, extreme Body Mass Index (BMI < 18 or BMI > 28), symptoms of mental illness, current prescribed regular medication and pregnancy or breastfeeding. Asian participants were excluded as well due to high probabilities of aldehyde

dehydrogenase isoenzyme deficiencies in this population (Wall et al., 1997), which can trigger aversive reactions to alcohol intake. All participants included in the experiment consumed more than six units of alcohol a week (1 unit = 8g of alcohol).

In addition, participants refrained from drinking alcohol for at least 12 hours prior to the test session and were breathalysed at the start of the session to ensure a blood alcohol concentration (BAC) of 0. They also refrained from taking illicit drugs for at least seven days, as well as caffeinated drinks and cigarettes an hour before the test. Participants were also required to have a low-fat meal the evening before testing and a low-fat breakfast on the day of testing.

Ethical approval was granted by the BSMS ethics committee at the University of Sussex.

5.3.2 Methods

5.3.2.1 Questionnaires

Participants completed the Alcohol Use Questionnaire (AUQ) (Mehrabian & Russell, 1978) for evaluating drinking habits. The AUQ measures, via 12 items, the amount of alcohol consumed per week as well as the frequency and speed of drinking to obtain an alcohol Binge score (Townshend & Duka, 2002).

Impulsivity traits were measured using Barratt's Impulsiveness Scale (BIS-11) (Patton, Stanford, & Barratt, 1995). This 30-item questionnaire assesses different constructs related with impulsivity, namely attentional, motor and non-planning impulsiveness, in addition to providing an overall impulsivity score.

The awareness subscale of the BPQ (Porges, 1993) measures trait sensitibility to bodily changes with 45 items, as ascertained via self-report, by asking participants to rate on a Likert scale the extent to which they feel different bodily sensations (i.e. facial twitches).

Finally, the ability to process emotions was assessed using the Toronto Alexithymia Scale (TAS-20) (Bagby, Parker, & Taylor, 1994), which measures, via a Likert scale, difficulties in describing feelings, difficulty identifying feelings and the propensity to engage in externally oriented thinking.

5.3.2.2 Current affect and subjective alcohol effects

Affects and subjective alcohol effects were measured using the Positive and Negative Affect Scale, (PANAS) (Watson, Clark, & Tellegen, 1988) and Subjective Alcohol-induced Effects Visual Analogue Scales (Alcohol VAS) (Duka et al., 1998). For the PANAS, participants evaluate their positive and negative affect rating 10 words for each construct. On the Alcohol VAS, participants had to indicate the extent to which they were experiencing a range of states (e.g. 'light-headed', 'stimulated', 'alert', 'relaxed' and 'contented').

5.3.3 Alcohol administration

Breath alcohol levels we measured using a breathalyser (Lion alcolmeter SD-400, Lion Laboratories Ltd., UK). Following baseline measurements, participants were randomly allocated to receive either an alcoholic or a non-alcoholic beverage in a double-blind design. Two different doses of alcohol were used on this experiment, either a low dose (0.4g/kg, n=18, 12 females) or a high dose (0.6g/kg, n=15, 8 females), with 90% v/v

alcohol, diluted with sugar-free tonic water (Schweppes, Uxbridge, UK) to make up a 500ml beverage mixed with 6 drops of Angostura bitters (Garfinkel, Dienes, & Duka, 2006). The placebo group (n=17, 11 females), was given a beverage consisting of 500 ml of tonic water mixed with an equivalent measure of Angostura bitters. The drink was divided into 10 portions of 50 ml and participants were instructed to consume them at 3 min intervals.

5.3.4 Interoception tasks

Interoceptive accuracy, operationalized as the objective ability to accurately detect internal bodily sensations using behavioural testing, was measured using the heartbeat discrimination (Katkin et al., 1983; Whitehead et al., 1977) and tracking (Schandry, 1981) tasks. For both tasks, participants' pulse was monitored using an 8000SM finger pulse oximeter (Nonin Medical, Inc., Minnesota, USA).

In the heartbeat tracking task, participants are instructed to count their heartbeats within their whole body, without putting their hands on their chest or neck. The task started with a practice trial of 20s after which the 6 experimental trials of different time-windows (25, 30, 35, 40, 45 and 50s) occurred in a randomized order. Through a set of speakers, participants heard the word "start" and had to count heartbeats until they heard "stop". At the end of each trial, they indicated to the experimenter the amount of heartbeats they had felt and completed a computerised visual analogue scale to evaluate how confident they are in their responses (0 not confident – 100 extremely confident).

Participants were then administered the heartbeat discrimination task. On each trial, ten auditory tones (100 ms, 440Hz) were presented either synchronized or asynchronously with the participant's own heartbeat. On non-synchronized trials, a 300ms delay was introduced between each heartbeat and the tone. After each trial, participants indicated whether tones were synchronized or not with their heartbeat and again indicated confidence in their responses using a visual analogue scale. In total, 20 trials were presented, randomly allocating synchronised and non-synchronised trials.

The order of the tasks was fixed for all participants.

Three dimensions of interoception, incorporating interoceptive accuracy, sensibility and metacognitive awareness, were computed for each task (Garfinkel et al., 2015). Interoceptive accuracy is based upon the overall performance on each of the tasks. For the discrimination task, interoceptive accuracy is the percentage of correct responses (hits and correct rejections). For the tracking task, scores are computed based upon the ratio of reported to actual heartbeats, using a formula that accounts for the effect of longer trials (Hart *et al*, 2013):

$$1 - \left(|\, nbeatsreal \; - \; nbeatsreported \,|\right) / \left(\left(nbeatsreal \; + \; nbeatsreported\right) / 2\right)$$

Interoceptive sensibility is a subjective measure computed from the average confidence in responses stated for both tasks.

For the tracking task, metacognitive awareness was calculated as the relationship between confidence and accuracy using Pearsons' correlations. A high correlation

implies increased metacognitive awareness. In the discrimination task, an Area Under Receiving Operating Curve (AUROC) (Green & Swets, 1966; Hajian-Tilaki, 2013) provided a measure of the extent to which confidence predicts accuracy accounting for participants' propensity to indicate high levels of confidence. Both these metacognitive measures provide accounts of individual differences in 'interoceptive insight' (Khalsa et al., 2017).

5.3.5 Procedure

Participants came into the lab after 12 pm. Once having read and signed a consent form they completed the AUQ, BIS-11, TAS-20 and BPQ questionnaires. Participants then filled PANAS and Alcohol VAS at baseline (t_0) and were breathalysed. Next, they were administered the drink depending on the group they had been assigned to (placebo, low or high dose). After a 10-minute resting period, breath alcohol levels were measured (t_1), together with PANAS and Alcohol VAS. Interoceptive measurements (tracking and discrimination tasks) were finally taken followed by a measurement of breath alcohol levels (t_2). After the experiment, participants were debriefed and remained in a calm area within the lab until their breath alcohol level had fallen below 0.18mg/L, half the legal driving limit in England. Participants also agreed not to drive or operate any machinery for at least 4 hours following the experiment.

5.4 Data analysis

5.4.1 Questionnaires, subjective alcohol effects and blood alcohol concentration

Questionnaire scores were compared between groups (placebo vs. low vs. high dose) with a series of One-way ANOVAs. We also compared heartrate scores both during the tracking and discrimination tasks between groups.

Positive and negative affect (PANAS) and Alcohol VAS scores were analysed using Two-way mixed ANOVAs with time (t_0 vs. t_1) as a within subjects' factor and group (placebo vs. low vs. high-dose) as a between subjects' factors.

Blood alcohol concentration (BAC) was calculated from breath alcohol measurements by multiplying breath alcohol levels by 2.3 and dividing them by 10. BAC levels were compared between groups (low vs. high-dose) and time (t_1 vs. t_2) with a Two-way ANOVA.

5.4.2 Interplay between alcohol and interoception on subjective alcohol effects

For participants who consumed alcohol, a linear regression examined light-headedness at t_1 as DV, with BAC at t_1, metacognitive interoceptive awareness, interoceptive accuracy and sensibility on the discrimination task, age and mean heartrate as predictors. An equivalent analysis was also performed using the tracking task. The regression aimed at providing evidence for the role of interoception in the perception of alcohol induced effects.

5.4.3 Effects of alcohol on interoception

A series of One-way ANOVAs examined group differences (placebo vs. low vs. high-dose) in interoceptive performance after alcohol consumption, incorporating as dependent measures metacognitive interoceptive awareness, interoceptive accuracy and sensibility for both the discrimination and tracking tasks. Interoceptive awareness has been seen to decrease with age (Khalsa, Rudrauf, & Tranel, 2009) and heartrate can be affected by alcohol administration (Conrod, Peterson, & Pihl, 2001; Sayette, 1993), for that reason these variables were included as covariates. Age data for one participant was missing and hence not accounted for on the interactions with list-wise deletion.

5.4.4 Exploratory analysis on gender effects

Data published during the write up of this manuscript indicated that alcohol administration decreased accuracy in the tracking task, albeit only in males (Abrams et al., 2018).

A post-hoc analysis explored this with a Two-way ANOVA with gender and group (placebo vs. low vs. high-dose) as a between subjects' factors on tracking accuracy.

5.5 Results

5.5.1 Questionnaires, subjective alcohol effects and BAC

Regarding questionnaire scores, there were no significant differences between groups (Fs<2.100, ps>.05).

In terms of the Alcohol VAS, a significant interaction between group and time was observed for ratings of light-headedness, $F(2,47)=11.067$, $p<.001$, $\eta^2=.320$, with participants in the low-dose group having lower levels of light-headedness than those in the high-dose group, $t(20.992)=3.369$, $p=.003$, $d=1.47$, who were also experiencing more light-headedness than the placebo group, $t(22.823)=6.489$, $p<.001$, $d=2.72$, post alcohol consumption. As expected there was a dose-dependent effect of alcohol on light-headedness.

The expected main effect of group, $F(1,31)=142.086$, $p<.001$, $\eta^2=.821$ and of time, $F(1,31)=61.257$, $p<.001$, $\eta^2=.664$, on BAC was also found.

No other significant effects were found. See Table 1 for full descriptive statistics and results.

	Placebo n=17		04ml/kg n=18		0.6ml/kg n=15				
	Mean	SD	Mean	SD	Mean	SD			
One-way ANOVA							*F*(2,49)	*p*	η^2
Age	21.06	1.60	21.06	2.38	23.47	6.15	2.098	.134	.08
BIS-11	68.12	9.42	74.33	5.81	71.40	15.33	1.502	.233	.06
Binge score	22.24	14.44	21.28	11.60	29.33	15.85	1.576	.218	.06
AUQ score	29.35	17.14	30.00	14.43	43.40	22.40	3.045	.057	.11
Porges	2.79	0.74	2.97	0.70	2.89	0.88	0.240	.788	.01
TAS-20	49.88	11.82	49.56	8.93	52.93	11.18	0.483	.620	.02
Tracking HR	73.11	11.77	81.08	15.57	71.11	12.46	2.617	.084	.10
Discrimination HR	72.47	10.98	79.67	13.41	70.47	12.44	2.608	.084	.10
2-way mixed ANOVA time x group							*F*(2,47)	*p*	η^2
PANAS Positive t_0	29.59	8.12	31.56	7.31	28.93	7.76	0.085	.918	.004
PANAS Positive t_1	27.12	7.86	30.11	8.90	26.93	9.19			
PANAS Negative t_0	16.65	4.85	20.78	7.47	15.73	5.50	0.519	.599	.022
PANAS Negative t_1	12.94	3.23	15.61	5.83	12.27	4.45			
Light-headedness t_0	0.20	0.20	0.16	0.25	0.17	0.17	11.067	.001	.320
Light-headedness t_1	0.38	0.25	0.52	0.35	0.81	0.11			
Irritability t_0	0.14	0.15	0.13	0.18	0.12	0.17	0.451	.640	.019
Irritability t_1	0.17	0.19	0.11	0.17	0.07	0.13			
Stimulated t_0	0.38	0.19	0.37	0.24	0.53	0.20	1.611	.210	.064
Stimulated t_1	0.45	0.24	0.54	0.20	0.54	0.25			
Alertness t_0	0.53	0.24	0.45	0.18	0.57	0.19	2.064	.138	.081
Alertness t_1	0.58	0.24	0.52	0.16	0.45	0.26			
Relaxed t_0	0.58	0.20	0.60	0.20	0.50	0.21	0.710	.497	.029
Relaxed t_1	0.64	0.25	0.66	0.23	0.65	0.27			
Content t_0	0.63	0.24	0.60	0.17	0.51	0.19	1.698	.194	.067
Content t_1	0.64	0.19	0.64	0.22	0.65	0.23			
							F(1,31)	*p*	η^2
BAC t_1		In %	0.053	0.015	0.094	0.005	0.010	.922	.000
BAC t_2		In %	0.045	0.010	0.085	0.005			

Table 1: Descriptive statistics and results comparing questionnaire scores between groups and state changes due to alcohol administration. The only significant interaction was between time and dose in light-headedness*, p*=.001*, n*=50.

5.5.2 Interplay between alcohol and interoception on subjective alcohol effects

The regression examining the factors contributing to subjective ratings of light-headedness at t_1 was significant, R^2=.463, $F(6,31)=3.805$, p=.008, see Table 2.

Predictor	*B*	*SE B*	*β*	*t*	*p*
Blood Alcohol Concentration t_1	4.65	2.28	0.34	2.04	.052
Metacognitive Discrimination	0.82	0.29	0.47	2.86	.008
Accuracy Discrimination	0.32	0.48	0.11	0.67	.507
Sensibility Discrimination	-0.16	0.24	-0.10	-0.65	.522
Age	0.00	0.00	0.04	0.26	.798
Mean Heart-rate	0.00	0.01	0.02	0.13	.900

R^2=.477. Dependent variable: Light-headedness at t_1

***Table 2*:** Regression table for Light-headedness at t_1. Reported light-headedness was significantly explained by Metacognitive interoceptive awareness for the Discrimination task and marginally by Blood Alcohol Concentration, *n*=33.

Metacognitive interoceptive awareness for the discrimination task was the best predictor of light-headedness following alcohol administration within the model accounting for BAC, age, mean heartrate and accuracy and sensibility on the discrimination task. Figure 1 (a-b) presents the relationship of Light-headedness scores

with BAC (Figure 1a) and with metacognitive interoceptive awareness for the discrimination task (Figure 1b).

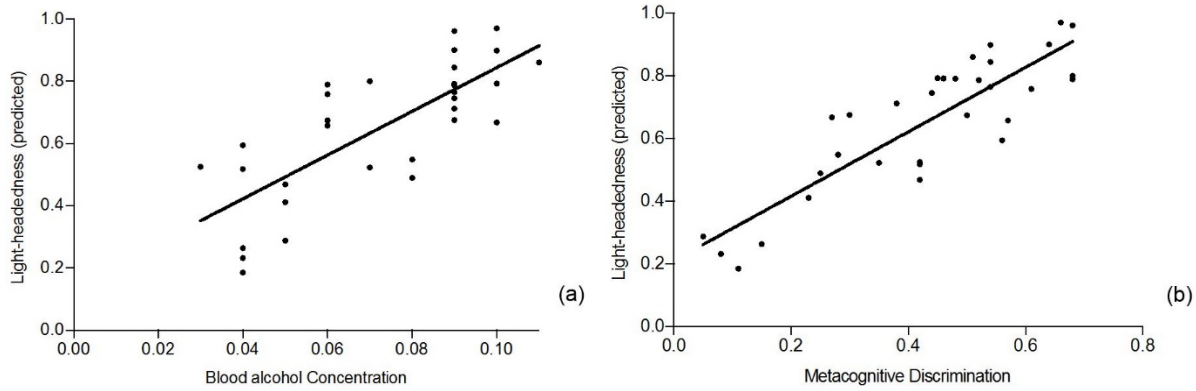


Figure 1: Scattergramms depicting the relationship of Unstandardized Predicted Light-headedness scores at t_1 with Blood Alcohol Concentration (a) and Metacognitive Interoceptive Awareness for the Discrimination task (b). Increased levels of Blood Alcohol Concentration and Metacognitive Discrimination positively correlate with feelings of light-headedness, *n*=33.

The regression using tracking scores was significant, R^2=.415, $F(6,31)=2.959$, p=.025, albeit the only significant predictor was BAC, p=.031.

5.5.3 Effects of alcohol on interoception

When examining the effect of alcohol on metacognitive interoceptive awareness for the discrimination task, a marginal main effect of dose was observed $F(2,48)=3.144$ p=.053, η^2=.125. Accounting for covariates, metacognitive interoceptive awareness under the low dose of alcohol was significantly reduced relative to the placebo group,

$F(1,33)=5.479$, $p=.026$, $\eta^2=.154$, demonstrating a deleterious effect of alcohol that was not found in the high-dose group, $F(1,31)=0.506$, $p=.483$, $\eta^2=.018$, see Figure 2(a).

There were no significant effects of group regarding discrimination accuracy, $F(2,48)=0.314$, $p=.732$, $\eta^2=.014$, or sensibility $F(2,48)=1.432$, $p=.250$, $\eta^2=.061$, see Figure 2(b-c).

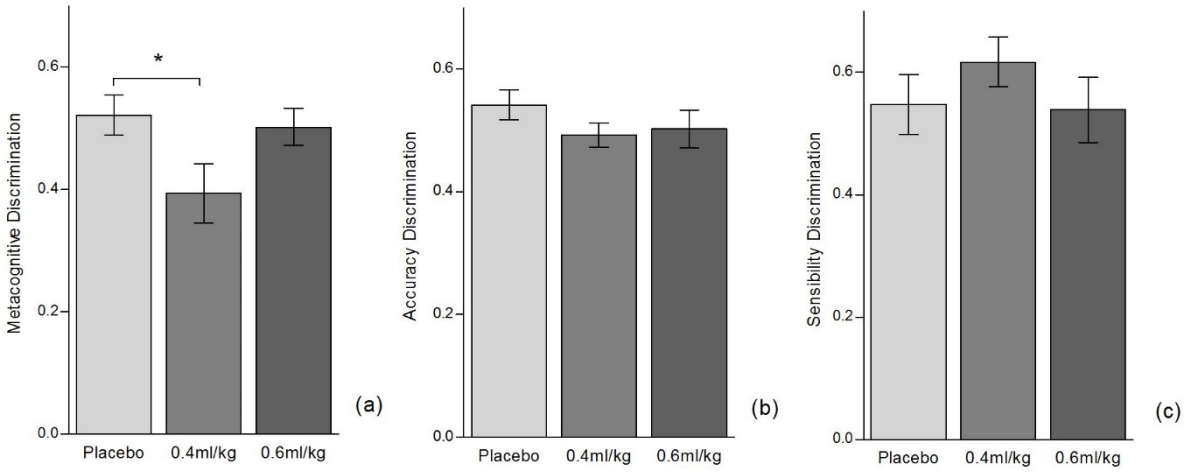


Figure 2: Metacognitive Interoceptive awareness (a), accuracy (b) and sensibility (c) obtained in the discrimination task across experimental groups (mean scores and SEM). *Metacognitive interoceptive awareness was higher for the placebo group than for the 0.4ml/kg group, $p=.026$, $n=50$.

Regarding the tracking task, there were no significant group effects on metacognitive interoceptive awareness, $F(2,48)=0.347$, $p=.709$, $\eta^2=.016$, accuracy, $F(2,48)=1.321$, $p=.277$, $\eta^2=.057$, or sensibility, $F(2,48)=0.588$, $p=.560$, $\eta^2=.026$ (data not shown).

5.5.4 Exploratory analysis on gender effects

Post-hoc analyses, including gender, show a marginal Two-way interaction between gender and dose for tracking accuracy, $F(2,48)=3.186$, $p=.052$, $\eta^2=.135$. This was explained by decreases in accuracy for males (*mean*=.42, *SD*=.44) compared to females (*mean*=.72, *SD*=.16) in the high-dose group, $F(1,14)=11.044$, $p=.007$, $\eta^2=.501$, which were not found in the other groups ($Fs<0.8$, $ps>.39$). No other effects including gender were found, $Fs<1$, $ps>.4$.

5.6 Discussion

This report assesses the role of interoceptive processes in the appraisal of drug-induced effects on mood states. Empirical evidence is provided for the effect of acute alcohol administration on the perception of internal bodily sensations and their relationship to drug effect experiences.

As predicted, there was a relationship between metacognitive interoceptive awareness and the subjective states alcohol induces in participants, specifically light-headedness. According to previous research, alcohol discrimination at very low doses is based on the influence the drink has on subjective light-headedness, an effect resembling “high” (Duka et al., 1998). Our findings demonstrate that metacognitive cardiac awareness correlates with higher acuity in the perception of substance-induced responses, meaning the greater one’s ability to recognise how well they perceive their internal bodily sensations, the more they experience substance effects. Insight in Interoceptive abilities can therefore constitute the basis of substance discrimination, which in turn can act as an interoceptive substance-related cue.

Such a relationship should play a relevant role in the development of associations between stimuli and drug effects supporting conditioning models of addiction (Stewart *et al*, 1984). Furthermore, it is possible that increased interoceptive awareness enhances the detection of low-intensity physiological responses, exemplifying again the function interoception has on the processing of emotional cues which are not explicitly accessible with ease (Damasio, 2000, Leganes-Fonteneau *et al,* 2018). This could be crucial in the understanding of the mechanisms underlying drug discrimination (Duka et al., 1999) and alcohol priming effects (Rose & Duka, 2006), as well as emotional biases to alcohol related stimuli.

In the present study, a decrease in metacognitive interoceptive awareness in the discrimination task was found for the low-dose group, highlighting that a low dose of alcohol may leave interoceptive accuracy and sensibility relatively unimpaired, but instead influence the capacity for metacognitive interoceptive insight. This effect was revealed after accounting for age differences and heartrate during the tasks.

It is possible that only the low dose of alcohol impaired interoceptive metacognition as participants at that dose did not yet have insight into their own intoxicated state. Under this low dose, confidence seemed boosted, in line with general stimulant effects of alcohol (Earleywine & Martin, 1993; Ray, McGeary, Marshall, & Hutchison, 2006) while performance too had a mild tendency to drop. While neither of these results were significant in their own right, it is possible that their modulation at this low dose resulted in a disruption of confidence-accuracy mapping, leading to a selective impairment in metacognitive interoception.

It is worth noting as well that acute administration of low alcohol doses affects general error monitoring (Ridderinkhof et al., 2002), a type of metacognitive ability, thus supporting the deleterious effect of the low dose on metacognitive interoception. Lack of such an effect in the high dose may be due to a compensatory mechanism mobilised when drug effects are experienced (e.g Marczinski and Fillmore, 2005) or expected (Caswell *et al*, 2013).

We did not find significant effects of alcohol administration on overall accuracy or sensibility for the discrimination task; or on any of the interoceptive indexes for the tracking task. An exploratory analysis did, however, replicate recently published data (Abrams et al., 2018). Males had lower accuracy in the tracking task than females, albeit results were restricted to the high-dose group. This replication highlights once more the role of interoception in addictive processes and brings further evidence towards the effects of acute alcohol administration in proprioception and other forms of perception (Stock, Mückschel, & Beste, 2017).

Physiological disparities between males and females could explain the differences observed in the tracking task (Ehlers, Mayou, Sprigings, & Birkhead, 2000). In males, alcohol administration may have affected interoceptive pathways, leading to the effects observed in tracking accuracy, which were not altered in females. Gender differences in cardiac functioning, notably in heart-rate variability (Bates et al., 2011; Koenig & Thayer, 2016) have already been reported, explaining disparities in emotional processing after alcohol administration (Udo et al., 2009). Further research should therefore examine the role of interoception in emotional responses and their relationship with heart-rate variability and gender differences.

The use of metacognitive measures of interoceptive awareness over simple indices of accuracy brings a novel approach towards the study of interoceptive processes (Garfinkel, Manassei, et al., 2016; Garfinkel & Critchley, 2013) that is relatively unexplored (see Canales-Johnson et al., 2015; Forkmann et al., 2016 and Garfinkel et al., 2015 for notable exceptions).The way metacognitive interoceptive awareness is computed, particularly using AUROC for the discrimination task, creates a measure which is less affected by individuals' dispositional or situational interoceptive sensibility (i.e. confidence) on that particular task (Fleming & Lau, 2014), and thus may provide an unbiased account of their interoceptive ability in the metacognitive domain. In our case, metacognitive discrimination did not differ between males and females but was affected by alcohol administration. Moreover, the predictive power of metacognitive interoception on light-headedness was present accounting for accuracy and confidence scores, pointing towards the unique role of metacognitive interoception in the appraisal of drug effects. Metacognitive indices of interoception might therefore constitute a better measure of the interoceptive correlates of addiction.

As explained before, different tasks and measures assessing interoception share similar and distinct functional architecture (Schulz, 2016) and reflect different cognitive processes (Garfinkel et al., 2016). Although observing alcohol effects on interoceptive processing across the whole range of measures would strengthen our conclusions, given that each of the tasks and measures evaluates different aspects of interoception (Garfinkel et al., 2016) it is not surprising to observe results limited to one or two indexes.

The present results suggest that the effects of alcohol appear to be more sensitive to an interoception paradigm that requires internal-external integration of stimuli. Interestingly both oxytocin (Betka et al., 2018) and stress (Schulz & Vögele, 2015) also selectively affect interoception as measured with the discrimination task, though the effects on this task were seen on accuracy rather than on the metacognition of interoception.

5.7 Limitations

The lack of a baseline measurement of interoception in the present study prevents the clear assertion that our findings are solely due to the direct effect of alcohol on interoception and not influenced also by individual differences on interoception. A baseline measurement of interoception, before the administration of any substance, would provide a clearer account of the effects of alcohol on interoception. However repeated administrations of the tasks could lead to learning effects, which could be a confound for the influence of alcohol. Future research should examine the role of interoception as a trait, and not as the result of an experimental manipulation, on alcohol discrimination abilities.

The present study did not directly assess participants' knowledge about the nature of the substance administered (placebo or alcohol) as a single administration does not allow sensitive measures of drug-discrimination accounting for chance identifications (50% probabilities of being accurate) (Jackson, 2001). We also did not assess physiological responses (e.g. heart-rate, blood-pressure) after alcohol administration.

Future studies should expand and improve these novel findings by incorporating these additional measures.

5.8 Conclusions

Taken together, our findings show that alcohol can alter some interoceptive processes. Our findings also show that interoceptive abilities can lead to differences in the perception of the effects elicited by a substance, ultimately constituting a risk factor in substance abuse disorders. Thus, the present study brings further evidence for the interplay between interoceptive processing and the perception of drug-induced mood changes, and opens a series of pathways for future research. Uncovering the interoceptive correlates of alcohol administration could shed light onto the link between bodily responses and different phenomena associated with alcohol and addiction, the implications of which could shape novel intervention programs.

5.9 Funding and Disclosure

This research was funded by the University of Sussex, United Kingdom. Authors report no conflict of interest.

5.10 References

Abrams, K., Cieslowski, K., Johnson, S., Krimmel, S., La Rosa, G. B.-D., Barton, K., & Silverman, P. (2018). The effects of alcohol on heartbeat perception: Implications for anxiety. *Addictive Behaviors*, *79*, 151–158. http://doi.org/10.1016/J.ADDBEH.2017.12.023

Ateş Çöl, I., Sönmez, M. B., & Vardar, M. E. (2016). Evaluation of Interoceptive Awareness in Alcohol-Addicted Patients. *Noro Psikiyatri Arsivi*, *53*(1), 17–22. http://doi.org/10.5152/npa.2015.9898

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23–32. http://doi.org/10.1016/0022-3999(94)90005-1

Bates, M. E., Buckman, J. F., Vaschillo, E. G., Fonoberov, V. A., Fonoberova, M., Vaschillo, B., … Mezić, I. (2011). The Redistribution of Power: Neurocardiac Signaling, Alcohol and Gender. *PLoS ONE*, *6*(12), e28281. http://doi.org/10.1371/journal.pone.0028281

Betka, S., Gould Van Praag, C., Paloyelis, Y., Bond, R., Pfeifer, G., Sequeira, H., … Critchley, H. (2018). Impact of intranasal oxytocin on interoceptive accuracy in alcohol users: An attentional mechanism? *Social Cognitive and Affective Neuroscience*.

Betka, S., Pfeifer, G., Garfinkel, S., Prins, H., Bond, R., Sequeira, H., … Critchley, H. (2018). How Do Self-Assessment of Alexithymia and Sensitivity to Bodily Sensations Relate to Alcohol Consumption? *Alcoholism: Clinical and Experimental Research*, *42*(1), 81–88. http://doi.org/10.1111/acer.13542

Cabanac, M. (1979). Sensory Pleasure. *The Quarterly Review of Biology*, *54*(1), 1–29. http://doi.org/10.1086/410981

Canales-Johnson, A., Silva, C., Huepe, D., Rivera-Rei, Á., Noreika, V., Garcia, M. del C., … Bekinschtein, T. A. (2015). Auditory Feedback Differentially Modulates Behavioral and Neural Markers of Objective and Subjective Performance When Tapping to Your Heartbeat. *Cerebral Cortex*, *25*(11), 4490–4503. http://doi.org/10.1093/cercor/bhv076

Caswell, A. J., Morgan, M. J., & Duka, T. (2013). Acute alcohol effects on subtypes of impulsivity and the role of alcohol-outcome expectancies. *Psychopharmacology*, *229*(1), 21–30. http://doi.org/10.1007/s00213-013-3079-8

Conrod, P., Peterson, J., & Pihl, R. (2001). Reliability and validity of alcohol-induced heart rate increase as a measure of sensitivity to the stimulant properties of alcohol. *Psychopharmacology*, *157*(1), 20–30. http://doi.org/10.1007/s002130100741

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews. Neuroscience*, *3*(8), 655–66. http://doi.org/10.1038/nrn894

Critchley, H. D., & Harrison, N. A. (2013). Visceral influences on brain and behavior. *Neuron*, *77*(4), 624–38. http://doi.org/10.1016/j.neuron.2013.02.008

Critchley, H., & Garfinkel, S. (2017). Interoception and emotion. *Current Opinion in Psychology*, *17*, 7–14. http://doi.org/10.1016/J.COPSYC.2017.04.020

Damasio, A. R. (2000). *The Feeling of what Happens: Body, Emotion and the Making of Consciousness*.

Duka, T., Jackson, A., Smith, D. C., & Stephens, D. N. (1999). Relationship of Components of an Alcohol Interoceptive Stimulus to Induction of Desire for Alcohol in Social Drinkers. *Pharmacology Biochemistry and Behavior*, *64*(2), 301–309. http://doi.org/10.1016/S0091-3057(99)00080-5

Duka, T., Stephens, D. N., Russell, C., & Tasker, R. (1998). Discriminative stimulus properties of low doses of ethanol in humans. *Psychopharmacology*, *136*(4), 379–389. http://doi.org/10.1007/s002130050581

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., … Dalgleish, T. (2010). Listening to Your Heart. *Psychological Science*, *21*(12), 1835–1844. http://doi.org/10.1177/0956797610389191

Earleywine, M., & Martin, C. S. (1993). Anticipated Stimulant and Sedative Effects of Alcohol Vary with Dosage and Limb of the Blood Alcohol Curve. *Alcoholism: Clinical and Experimental Research*, *17*(1), 135–139. http://doi.org/10.1111/j.1530-0277.1993.tb00738.x

Ehlers, A., Mayou, R. A., Sprigings, D. C., & Birkhead, J. (2000). Psychological and perceptual factors associated with arrhythmias and benign palpitations. *Psychosomatic Medicine*, *62*(5), 693–702. http://doi.org/10.1097/00006842-200009000-00014

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. http://doi.org/10.3389/fnhum.2014.00443

Forkmann, T., Scherer, A., Meessen, J., Michal, M., Schächinger, H., Vögele, C., & Schulz, A. (2016). Making sense of what you sense: Disentangling interoceptive awareness, sensibility and accuracy. *International Journal of Psychophysiology*, *109*, 71–80. http://doi.org/10.1016/J.IJPSYCHO.2016.09.019

Garavan, H. (2010). Insula and drug cravings. *Brain Structure & Function*, *214*(5–6), 593–601. http://doi.org/10.1007/s00429-010-0259-8

Garfinkel, S. N., & Critchley, H. D. (2013). Interoception, emotion and brain: new insights link internal physiology to social behaviour. Commentary on:: “Anterior insular cortex mediates bodily sensibility and social anxiety” by Terasawa et al. (2012). *Social Cognitive and Affective Neuroscience*, *8*(3), 231–4. http://doi.org/10.1093/scan/nss140

Garfinkel, S. N., Dienes, Z., & Duka, T. (2006). The effect of alcohol and repetition at encoding on implicit and explicit false memories. *Psychopharmacology*, *188*(4), 498–508. http://doi.org/10.1007/s00213-006-0480-6

Garfinkel, S. N., Manassei, M. F., Hamilton-Fletcher, G., In den Bosch, Y., Critchley, H. D., & Engels, M. (2016). Interoceptive dimensions across cardiac and respiratory axes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708), 20160014. http://doi.org/10.1098/rstb.2016.0014

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. http://doi.org/10.1016/j.biopsycho.2014.11.004

Garfinkel, S. N., Tiley, C., O’Keeffe, S., Harrison, N. A., Seth, A. K., & Critchley, H. D.

(2016). Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety. *Biological Psychology*, *114*, 117–126. http://doi.org/10.1016/J.BIOPSYCHO.2015.12.003

Gray, M. A., & Critchley, H. D. (2007). Interoceptive Basis to Craving. *Neuron*, *54*(2), 183–186.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *New York: Wiley*.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–35.

Hart, N., McGowan, J., Minati, L., & Critchley, H. D. (2013). Emotional regulation and bodily sensation: interoceptive awareness is intact in borderline personality disorder. *Journal of Personality Disorders*, *27*(4), 506–518.

Jackson, A., D., S., & T., D. (2001). A low dose alcohol drug discrimination in social drinkers: relationship with subjective effects. *Psychopharmacology*, *157*(4), 411–420. http://doi.org/10.1007/s002130100817

Katkin, E., Reed, S., & Deroo, C. (1983). A methodological analysis of 3 techniques for the assessment of individual-differences in heartbeat detection.

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., … Zucker, N. (2017). Interoception and Mental Health: A Roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. http://doi.org/10.1016/J.BPSC.2017.12.004

Khalsa, S. S., Rudrauf, D., & Tranel, D. (2009). Interoceptive awareness declines with age. *Psychophysiology*, *46*(6), 1130–1136. http://doi.org/10.1111/j.1469-8986.2009.00859.x

Koenig, J., & Thayer, J. F. (2016). Sex differences in healthy human heart rate variability: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *64*, 288–310. http://doi.org/10.1016/J.NEUBIOREV.2016.03.007

Kopera, M., Jakubczyk, A., Suszek, H., Glass, J. M., Klimkiewicz, A., Wnorowska, A., … Wojnar, M. (2015). Relationship Between Emotional Processing, Drinking Severity and Relapse in Adults Treated for Alcohol Dependence in Poland. *Alcohol and Alcoholism*, *50*(2), 173–179. http://doi.org/10.1093/alcalc/agu099

Mehrabian, A., & Russell, J. A. (1978). A questionnaire measure of habitual alcohol use. *Psychological Reports*, *43*(3), 803–806. http://doi.org/10.2466/pr0.1978.43.3.803

Naqvi, N. H., & Bechara, A. (2010). The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. *Brain Structure & Function*, *214*(5–6), 435–50. http://doi.org/10.1007/s00429-010-0268-7

Naqvi, N. H., Rudrauf, D., Damasio, H., & Bechara, A. (2007). Damage to the insula

disrupts addiction to cigarette smoking. *Science (New York, N.Y.)*, *315*(5811), 531–4. http://doi.org/10.1126/science.1135926

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774. http://doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Paulus, M. P., Tapert, S. F., & Schulteis, G. (2009). The role of interoception and alliesthesia in addiction. *Pharmacology, Biochemistry, and Behavior*, *94*(1), 1–7. http://doi.org/10.1016/j.pbb.2009.08.005

Porges, S. (1993). Body perception questionnaire. *Laboratory of Developmental Assessment, University of Maryland*.

Ray, L. A., McGeary, J., Marshall, E., & Hutchison, K. E. (2006). Risk factors for alcohol misuse: Examining heart rate reactivity to alcohol, alcohol sensitivity, and personality constructs. *Addictive Behaviors*, *31*(11), 1959–1973. http://doi.org/10.1016/J.ADDBEH.2006.01.010

Ridderinkhof, K. R., de Vlugt, Y., Bramlage, A., Spaan, M., Elton, M., Snel, J., & Band, G. P. H. (2002). Alcohol Consumption Impairs Detection of Performance Errors in Mediofrontal Cortex. *Science*, *298*(5601), 2209–2211. http://doi.org/10.1126/science.1076929

Ring, C., & Brener, J. (1996). Influence of beliefs about heart rate and actual heart rate on heartbeat counting. *Psychophysiology*, *33*(5), 541–546. http://doi.org/10.1111/j.1469-8986.1996.tb02430.x

Rose, A. K., & Duka, T. (2006). Effects of dose and time on the ability of alcohol to prime social drinkers. *Behavioural Pharmacology*, *17*(1), 61–70. http://doi.org/10.1097/01.fbp.0000189814.61802.92

Sayette, M. A. (1993). Heart Rate as an Index of Stress Response in Alcohol Administration Research: A Critical Review. *Alcoholism: Clinical and Experimental Research*, *17*(4), 802–809. http://doi.org/10.1111/j.1530-0277.1993.tb00845.x

Schandry, R. (1981). Heart Beat Perception and Emotional Experience. *Psychophysiology*, *18*(4), 483–488. http://doi.org/10.1111/j.1469-8986.1981.tb02486.x

Schulz, A., & Vögele, C. (2015). Interoception and stress. *Frontiers in Psychology*, *6*, 993. http://doi.org/10.3389/fpsyg.2015.00993

Schulz, S. M. (2016). Neural correlates of heart-focused interoception: a functional magnetic resonance imaging meta-analysis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708), 20160018. http://doi.org/10.1098/rstb.2016.0018

Sherrington, C. S. (1948). The integrative action of the nervous system, (Cambridge Univ. Pess, Cambridge, UK).

Stewart, J. L., May, A. C., Tapert, S. F., & Paulus, M. P. (2015). Hyperactivation to pleasant interoceptive stimuli characterizes the transition to stimulant addiction. *Drug and Alcohol Dependence*, *154*, 264–270. http://doi.org/10.1016/J.DRUGALCDEP.2015.07.009

Stock, A.-K., Mückschel, M., & Beste, C. (2017). Reversal of alcohol-induced effects on response control due to changes in proprioceptive information processing. *Addiction Biology*, *22*(1), 246–256. http://doi.org/10.1111/adb.12296

Thorberg, F. A., Young, R. M., Sullivan, K. A., & Lyvers, M. (2009). Alexithymia and alcohol use disorders: A critical review. *Addictive Behaviors*, *34*(3), 237–245. http://doi.org/10.1016/J.ADDBEH.2008.10.016

Toates, F. M. (Frederick M. . (1986). *Motivational systems*. Cambridge University Press.

Townshend, J., & Duka, T. (2002). Patterns of alcohol drinking in a population of young social drinkers: a comparison of questionnaire and diary measures. *Alcohol and Alcoholism*.

Udo, T., Bates, M. E., Mun, E. Y., Vaschillo, E. G., Vaschillo, B., Lehrer, P., & Ray, S. (2009). Gender differences in acute alcohol effects on self-regulation of arousal in response to emotional and alcohol-related picture cues. *Psychology of Addictive Behaviors*, *23*(2), 196–204. http://doi.org/10.1037/a0015015

Verdejo-Garcia, A., Clark, L., & Dunn, B. D. (2012). The role of interoception in addiction: A critical review. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1857–1869. http://doi.org/10.1016/j.neubiorev.2012.05.007

Wall, T. L., Peterson, C. M., Peterson, K. P., Johnson, M. L., Thomasson, H. R., Cole, M., & Ehlers, C. L. (1997). Alcohol Metabolism in Asian-American Men with Genetic Polymorphisms of Aldehyde Dehydrogenase. *Annals of Internal Medicine*, *127*(5), 376. http://doi.org/10.7326/0003-4819-127-5-199709010-00007

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. http://doi.org/10.1037/0022-3514.54.6.1063

Whitehead, W. E., Drescher, V. M., Heiman, P., & Blackwell, B. (1977). Relation of heart rate control to heartbeat perception. *Biofeedback and Self-Regulation*, *2*(4), 371–392. http://doi.org/10.1007/BF00998623

6 Discussion

6.1 Summary of results

6.1.1 Stimuli conditioned with rewards generate preferential responses in the absence of outcome-expectancies

The experiments presented in sections 2 and 3 examined implicit aspects of PC.

The core focus of this series of experiments was to provide substantial evidence, with different samples and paradigms, for the existence of implicit appetitive PC.

Using task irrelevant conditioning procedures, we were able to associate stimuli with different probabilities of reward outcomes. With the use of Bayes factors, it was possible to determine whether participants had gained the ability to predict the occurrence of rewards or not.

Although it seems from our data that CA is necessary to develop explicit CResp (e.g. hedonic reactions) towards CS, stimuli associated with HR probabilities generated preferential responses compared to LR stimuli in the absence of CA.

First, these observations were derived using an Emotional Attentional Blink task. Participants had to detect the presence of CS in a RSVP stream. Emotional aversive distractors were embedded in the stream, decreasing accuracy in the detection of target CS. However, for Unaware participants, HR stimuli overcame the effect of the distractors, suggesting that the incentive value acquired by the HR CS can direct attentional responses in the absence of outcome-expectancies. The way this task was designed required participants to engage in goal-directed mental processes as they had to detect a target stimulus. The interference by aversive distractors however, and

the ability of CS to overcome these distractors, would reflect implicit attentional responses towards CS. It is possible therefore that the examination of implicit attentional components allows detecting CResp in the absence of outcome-expectancies.

Furthermore, using CS as distractors in a Flanker task, we found that HR CS generated stronger cognitive interference than LR or control stimuli only for participants Unaware of contingencies. This time, CS were completely task irrelevant, devoid of any goal directedness, and yet, they interfered in cognitive control in a similar way as other incentive salient stimuli (i.e. alcohol pictures (Nikolaou, Field, & Duka, 2013)).

We have been therefore able to provide evidence that CS can generate preferential responses in the absence of CA. Such automatic attentional allocation responses can be elicited both in situations where the salient stimulus is a target (requiring a goal-directed focus of attention), or in situations where it interferes with an irrelevant distractor (as in the Flanker task); in both cases the stimulus outcome contingencies were not explicitly recognised.

Interestingly, using an n-back task in which CS were supposed to interfere with working memory in a similar way as in the Flanker task, we did not find any effects in participants Unaware of contingencies, but only in those Aware, albeit the effect was marginal. As discussed on that chapter, such a differential contribution to attentional biases by CA may reflect different cognitive processes underlying the two tasks, the one being more susceptible to automatic interferences by CS (in the Flanker task) and the other one requiring explicit access to the incentive value of CS (in the n-back task).

Regarding pleasantness, we hypothesised that Unaware participants would show preferential responses towards CS as per recent research (Jeffs & Duka, 2017). However, we only found those responses in participants Aware of contingencies. This might be partially due to hedonic responses being assessed via explicit pleasantness ratings. The fact that these explicit hedonic responses were previously found in Unaware participants (Jeffs and Duka, 2017) is difficult to understand. One explanation could be that the separation of Aware versus Unaware participants in that study was not as rigorous as in the present experiments, leading to an overestimate of learning effects in Unaware participants (Shanks, 2016).

Implicit hedonic reactions might not be readily accessible to conscious observation, that is, a subjective measure of pleasantness might not be sensitive to the detection of implicit learning, given that the former belongs to the realm of conscious experience and the latter is unconscious. On the other hand, attentional responses, reflecting automatic reactivity, could better grasp implicit correlates of learning, therefore explaining the differences between measurements (De Houwer, 2006).

A Bayes factor showed that for Unaware participants, results of pleasantness evaluations were insensitive, whereas for Aware participants, the increased pleasantness towards HR CS compared to LR was fully confirmed. This can, again, reflect the insensitivity of subjective measurements when it comes to evaluating implicit learning. It is also possible that factors other than outcome-expectancies can participate in the development of those responses. The study of interoceptive correlates of learning can shed light on this matter, as high interoceptive awareness

seems to facilitate hedonic responses and reward prediction, key elements in the generation of Pavlovian effects.

6.1.2 Interoceptive awareness facilitates the development of conditioned responses and potentiates the perception of subjective alcohol effects

We designed two experiments targeting different facets of learning and their relationship with interoception (sections 4 and 5). Interoceptive processes have been seen to participate in different aspects of addiction (Paulus & Stewart, 2014; Verdejo-Garcia, Clark, & Dunn, 2012), and therefore understanding their relationship with reward processing is necessary.

We measured the development of hedonic responses and outcome-expectancies in an appetitive learning task and how interoceptive awareness mediates this process. We found that participants with high metacognitive interoceptive awareness developed higher hedonic responses towards HR CS. They also had a higher ability to predict the occurrence of rewards. Participants with low interoceptive abilities on the other hand were not able to develop either of those responses. These results go in line with research pointing at the relationship of interoception with emotional processing and learning (Critchley & Garfinkel, 2017; Garfinkel et al., 2013; Pfeifer et al., 2017; Pollatos & Schandry, 2008); but also with reward prediction (Kandasamy et al., 2016) or aversion to monetary losses (Sokol-Hessner et al., 2015).

A series of possible underlying mechanisms will be discussed later towards a theoretical conceptualisation of this relationship. However, in order to better

understand this matter, we examined as well how interoception shapes the perception of substance related rewards (e.g. feelings of high, stimulant effects).

We administered an acute alcohol dose (0.4ml/kg or 0.6ml/kg) or a control drink to participants and measured interoceptive indexes. Our main finding was that high interoceptive abilities (as measured with metacognitive discrimination) correlate with perceived light-headedness (a feeling of “high”) after alcohol administration, indicating that interoception may amplify the perception of drug effects. Additionally, alcohol decreased metacognitive interoceptive awareness for the discrimination task in participants being administered the low dose. The fact that low (but not high) doses of alcohol decreased metacognitive interoception could reflect an impairment in metacognitive processing due to the unawareness of substance effects in the low dose. In the high dose, the awareness of substance effects may engage compensatory mechanisms responsible for maintaining an accurate monitoring of interoceptive performance.

The high dose of alcohol on the other hand decreased accuracy on the tracking task, albeit only on male participants, replicating previous results (Abrams et al., 2018). This replication of differential effects in gender might reflect differences in the effect of alcohol administration on cardiac responses between men and women (Bates et al., 2011).

6.2 Theoretical implications

6.2.1 Relationship of implicit conditioning with appetitive learning and addiction

6.2.1.1 Learning theories

The main finding of this thesis, that CS can generate attentional responses in the absence of outcome-contingencies, has implications for the understanding of incentive theories of learning. Bolles (1972) argued that the source of reinforcement was not S-R associations, but rather S-R or S-S expectancies. The main determinant of learning then would be the strength of the association and the generation or positive or aversive expectancies driving behaviour. The results presented in this thesis do not completely discredit his hypothesis, but at the very least should nuance the corpus of appetitive learning theories based on outcome-expectancies. Dickinson (1989) and Shanks (2007) stressed as well the role of expectancies in the development of CResp, which are supposed to guide behaviours.

Without a doubt, outcome-expectancies play a capital role in the development of conditioned responses. We found in three different experiments that explicit hedonic responses could only occur in participants Aware of stimulus-outcome contingencies. However, given Bayesian analyses show insensitive results for Unaware participants, we cannot conclude that subjective pleasantness does never develop in the absence of CA.. Jeffs and Duka (2017) also found that CA was necessary for CS to guide motivational transfer on a PIT task. However, not only us, as reported in Chapters 2 and 3, but also others (e.g. Anderson, 2015; Bourgeois, Neveu, & Vuilleumier, 2016) recently found that reward CS can generate attentional responses, representing

substantial evidence for the existence of implicit PC, i.e. in the absence of outcome-expectancies.

Bindra's (1978) view on learning mechanisms can provide a framework, in which these results can be integrated. Bindra claimed that the motivational properties of rewards are transferred to CS after repeated concomitant presentations. It would be this motivational transfer, occurring regardless of outcome-expectancies, that drives learning effects observable through approach behaviours towards CS, and not the knowledge about outcome-contingencies. In our case, the motivational transfer can be observed through the attentional interference and allocation generated by CS in the absence of CA.

Moreover, Bindra (1974) posited that R-S associations were not necessary for learning to occur, and that learning can be ultimately derived from S-S pairings. He proposed a theoretical procedure by which S-S associations could be separated from the corresponding R-S (or in his own words, "isolation of the observation of lever stimuli from the lever-pressing response" (Bindra, 1974, p. 207)). In terms of animal conditioning, such a procedure might have been difficult to design. However, in humans, the use in our experiments of task irrelevant conditioning procedures, for which the nature of the response (i.e. pressing a key depending on the colour of an irrelevant square) is foreign to the contingencies between CS (Octagons or Squares) and rewards, appears to have supported his ideas.

The perspective embraced by Toates (1986) reflects the importance of both the hedonic value of US and of outcome-expectancies in appetitive learning, providing a

theoretical account of learning that reconciles both views. Results showing on one hand that CA is necessary for increased explicit hedonic reactions and behavioural motivation, and on the other hand that attentional responses can occur regardless of outcome-expectancies would support Toates's theory.

6.2.1.2 Addiction theories

As explained before, many of the advancements in appetitive learning theories have been incorporated within addiction perspectives. Our findings, therefore, have implications for understanding drug-addiction. For instance, Tiffany's (1990) conceptualisation of addictive processes as two discrete mechanisms, based on automatic physiological responses or on non-automatic approach behaviours, would be supported by our findings. Although we did not evaluate physiological reactivity towards CS, the fact that CS can drive automatic attentional responses implicitly but not affect behavioural outcomes or explicit hedonic responses would match to some extent his proposal. CS in the absence of CA could partly trigger some of the "drug (reward in the present experiments)-action plans" mentioned by Tiffany.

The experiments presented here do not allow an accurate exploration of the role of implicitly CS in the establishment of habitual responses (Everitt & Robbins, 2005). This theory posits that stimuli associated with a substance can trigger habitual responses, and that once those stimuli are presented in extinction they will be more resistant to outcome devaluation (in contrast to stimuli associated with a sweet solution for instance) (Miles, Everitt, & Dickinson, 2003). Hogarth and Chase (2011) however, showed that drug expectancies (or more precisely, the awareness that a CS is no longer followed by a substance) can override existing habitual responses in humans,

discrediting habit theories. It would be possible to integrate our methodologies within outcome-devaluations tasks in order to provide some evidence consistent with animal research in favour of habit theories. For instance, we could assess whether participants are able to develop CResp towards stimuli predicting a reward (i.e. cigarette puffs), and whether after extinction participants Unaware of changes in contingencies can maintain attentional or behavioural responses towards CS or not.

Results obtained by Hogarth and colleagues (Hogarth & Chase, 2011; Hogarth, Dickinson, Hutton, Bamborough, & Duka, 2006; Hogarth, Dickinson, Wright, Kouvaraki, & Duka, 2007) represented the adaptation to addiction theories of Dickinson's postulates, that reward processing varies according both to the incentive value of a reward and to the generation of outcome-contingencies (Dickinson & Balleine, 1994), although Hogarth and colleagues always stressed the importance of CA over implicit incentive values. It is therefore possible that two distinct mechanisms associated with CS, one requiring explicit knowledge about contingencies and high predictive value, and another one, based on implicit associations, can drive drug related behaviours in different ways. Implicit associations might generate preparatory or anticipatory responses (in the form of attentional or autonomous reactivity) whereas explicit learning is necessary for overt behavioural responses.

Although the present work did not target the distinction between "wanting" and "liking" as conceptualised by Robinson and Berridge (2016), it did indeed examine hedonic components of appetitive learning (i.e. pleasantness), finding that the occurrence of such phenomenon seems to be dependent on CA. Interestingly, Robinson and Berridge previously suggested the role of implicit processes in this

regard (Berridge, 1999; Berridge & Winkielman, 2003; Winkielman, Berridge, & Wilbarger, 2005). Particularly, they posited that motivation ("wanting"), learning, and emotion ("liking") are supported by implicit mechanisms (Berridge & Robinson, 2003), and that subjective measures of pleasantness might not be able to reflect implicit "liking" effects. This might explain again why we did not find subjective hedonic responses in the absence of CA, as pleasantness measurements might be insensitive to implicit manifestations of "liking". The experiments presented here however do not allow us to determine whether the attentional responses obtained in the absence of CA correspond to implicit correlates of "wanting" or "liking".

6.2.1.3 Dual process theories of addiction

Dual process theories of addiction offer a framework in which CS can drive automatic drug approach behaviours and interfere with executive functioning (Wiers & Stacy, 2006). In this context, it is reasonable to assume that the automaticity of such effects can reflect their unconscious nature. Usually, such models consider that these CS are stimuli explicitly associated with the substance (e.g. alcohol pictures), and it is true that drug-related stimuli can generate automatic interferences when used as distractors on cognitive tasks (e.g. Hester & Garavan, 2009; Nikolaou, Field, Critchley, & Duka, 2013), fitting the model.

Our findings support this position, and add further that CS predicting rewards in the absence of CA are able to elicit such interferences. It seems that conscious outcome-expectancies are not a requisite for CS to generate the same kind of behavioural response as drug-cues (as shown by the parallel effects on the Flanker task between implicit CS in our case, and alcohol pictures on Nikolaou at al.'s experiment (2013)).

However, even if drug cues can generate automatic responses, they are also providing explicit information about their relationship to the outcome (i.e. a bottle of alcohol is explicitly associated with the beverage). Thus, an experimental approach using automatic responses towards CS without CA allows the examination of the implicit correlates of reward processing, detached from explicit associations that might have formed between drug-related stimuli and substances through drug consumption experiences (Wiers et al., 2002). In our case, the use of abstract shapes associated with rewards in the absence of CA provides a "purer" automatic account of reward processing.

Moreover, if implicit CS can generate such behaviours, it would imply that a whole range of stimuli not readily accessible to consciousness can trigger addictive processes. This would not only apply to drug cues presented subliminally (e.g. Wetherill et al., 2014), to the influence of implicit emotional stimuli (e.g. Winkielman et al., 2005), or to the effect of a substance administered without explicit awareness (e.g. Hart, Ward, Haney, Foltin, & Fischman, 2001), but to stimuli predicting a reward and readily observable in plain sight. In our conditioning paradigms, CS were always present on screen, and the source of contingency Unawareness relied on the cognitive effort necessary to perform the dual task and on the uncertainty introduced by reward probabilities. It is easy to imagine how in real life scenarios, stimuli belonging to different sensory modalities (a sight, smell, place, person or even an interoceptive sensation) might have predicted substance effects in the past, remotely from explicit drug related cues and without conscious awareness of contingencies. This lack of awareness could be either due to the cognitive demands of other mental processes at

the time of the association (in a way similar to our dual task effect), to forgetfulness, or to the unpredictable probabilities of a given stimulus to be paired with a reward. Relapse prevention therapies in addiction (Marlatt & Donovan, 2005) are based on identifying possible risk situations, such as external or emotional stimuli (e.g. walking in front of bar, seeing a pack of cigarettes on a table), towards which individuals have conscious access. Based on that, coping strategies are implemented. But if stimuli or associations below conscious detection thresholds can trigger risk situations, other interventions, aiming at increasing cognitive and emotional control systems (with which reward related stimuli interfere), might provide a set of skills adapted to the implicit nature of those stimuli.

6.2.2 Relationship of interoceptive processing with appetitive learning and addiction

6.2.2.1 Learning theories

The role of interoception in appetitive learning but also in the way rewarding effects of substances are experienced was investigated in this thesis.

Regarding appetitive learning, our results show the importance of interoceptive awareness on emotional reactivity. We found that metacognitive discrimination mediates hedonic responses towards HR CS, in a similar way as others (Füstös, Gramann, Herbert, & Pollatos, 2013; Pfeifer et al., 2017; Pollatos & Schandry, 2008), but also, that CA can be explained by metacognitive tracking, in line with previous research pointing to the link between interoception and outcome prediction (Katkin, Wiens, & Ohman, 2001). Finally, the value or detection of US effects (as shown by our results on acute alcohol administration) can be explained by interoceptive awareness.

Drive reduction theories already posited the relevance of internal states in the establishment of learnt responses (Hull, 1943). Although long discredited, within that model interoceptive processes would be necessary to observe the internal bodily states generating drives (e.g. hunger) and also to quantify their satiation.

Bindra (1968) proposed that appetitive states depend on the presence of an incentive stimulus (e.g. food) concomitantly with a compatible organismic state (e.g. hunger) that can together generate a central motive state able to drive consummatory reactions. Here the focus was on incentive stimuli and their hedonic properties, but also on the "sensory inflow" (a composite of exteroceptive and interoceptive information (Bindra, 1978, p.88)) providing information on the organismic state of the subject. That sensory inflow, according to our results, could be determined by interoceptive abilities. Higher interoceptive awareness would facilitate the detection of physiological states compatible with reward perception and appraisal, fostering a central motive state able to explain the development of increased hedonic responses towards CS in highly interoceptive participants.

Toates's perspective was that both hedonic responses and outcome-expectancies are responsible for PC effects. According to our findings, interoceptive awareness would be related to both factors, pleasantness and outcome-expectancies, responsible for learning effects.

Regarding the mechanisms by which this occurs, alliesthetic views of learning propose that reward values depend on the "internal milieu" of the subject (Cabanac, 1979; Toates, 1986) and whether the reward is compatible with it. In a situation in which the

subject is craving or needing a substance or food, the value of that reward would be higher than once satiety is achieved. The administration of small quantities of a rewarding stimulus (e.g. food) or the presentation of CS can also trigger alliesthetic mechanisms, increasing the value of the reward and hedonic responses. In our case, reward CS may have triggered an alliesthetic response, increasing their perceived value. This response would be amplified in highly interoceptive participants as shown particularly by the heightened emotional reactivity towards HR CS.

Finally, cognitive expectancy theories (Dickinson, 1989; Dickinson & Balleine, 1995) do not particularly stress the role of physiological responses or interoceptive signalling, but interoceptive processing and its facilitator effects on reward prediction could be integrated in classical outcome-expectancy accounts of learning.

6.2.2.2 The role of interoception in Addiction

Theoretical and experimental work has pointed towards the link between interoception and addiction (Gray & Critchley, 2007; Paulus & Stewart, 2014; Verdejo-Garcia et al., 2012), positing that the appraisal of physiological responses plays a role at several stages of the aetiology (Naqvi and Bechara, 2010).

In a similar way as with Hullian drive reduction perspectives, withdrawal theories of addiction (Jellinek, 1955) have long been discredited. However, Siegel (1975) attributed to compensatory physiological mechanisms the occurrence of drug craving and drives. The preparatory responses generated by drug related stimuli (i.e. anxiety in the presence of a bottle of alcohol) are meant to generate withdrawal responses. We did not directly design any experiment targeting those preparatory responses, but the

finding that interoceptive awareness correlates with reward prediction, hedonic responsiveness and detection of substance effects could support Siegel's theoretical account. In accordance to these ideas, if an individual is better able to perceive preparatory responses triggered by reward CS due to increased interoceptive abilities, then it would make sense to display heightened responsiveness towards those stimuli or towards rewards.

Perhaps, the most well-grounded hypothesis is the one proposing an alliesthetic role of interoception in addiction (Paulus, Tapert, & Schulteis, 2009) reflecting Toates's learning theory. Again, testing such hypothesis would require paradigms targeting alliesthetic processes, for example administering small doses of alcohol and measuring a possible increase in the expected value of the substance.

In any case, the finding that metacognitive interoceptive awareness modulates the perception of subjective alcohol effects, found in the present thesis, implies that the perception of physiological states is a crucial factor in the development of addiction. In that sense for example, sensitivity to interoceptive stimulation predicts the transition from abuse to dependence in meta-amphetamine addicts (Stewart, May, Tapert, & Paulus, 2015). It is possible that being more attuned with one's bodily sensations provides amplified access to the physiological states generated by a substance. This higher awareness of substance effects, particularly of those associated with being "high", would intensify reward salience, provoking an increase in conditioned responses at the origin of addictive behaviours.

Finally, Gray and Critchley (2007) proposed that craving responses or drug urges would depend on the detection of physiological reactions elicited by drug-related stimuli. This would be supported by our results indicating increased responsiveness towards CS in highly interoceptive participants.

In light of the results obtained and incorporating as well the ideas of Paulus et al. (2009) and Gray and Critchley (2007), it is possible to propose a model of interoceptive incentive appetitive learning.

First, interoception facilitates the detection of the physiological effects generated by a substance, as shown with alcohol administration and the perception of light-headedness (a positive effect similar to feeling “high”). The experience of a positive response amplifies the salience of rewarding effects, which are then transferred to CS. Interoception, as shown in our appetitive PC experiment, also facilitates reward prediction, improving outcome-expectancy awareness, which would trigger reward approach behaviours (Hogarth et al., 2007). The internal state or milieu of the subject will shape alliesthetic effects (Cabanac, 1979; Toates, 1986), together with the presentation of CS or the priming by small doses of reward (i.e. alcohol micro-dosing - Duka & Townshend, 2004)). These alliesthetic processes will in turn be amplified by interoceptive awareness (Paulus, 2009) to determine the value of rewards.

Once CS acquire the incentive value associated with the reward, their mere presentation will trigger in turn physiological responses amplified by interoceptive awareness able to provoke positive hedonic responsiveness, drug urges and cravings

(Gray & Critchley, 2007), and maybe approach behaviours (e.g. behavioural responses, PIT effects or attentional biases) compatible with the nature of CS, see Fig. 1.

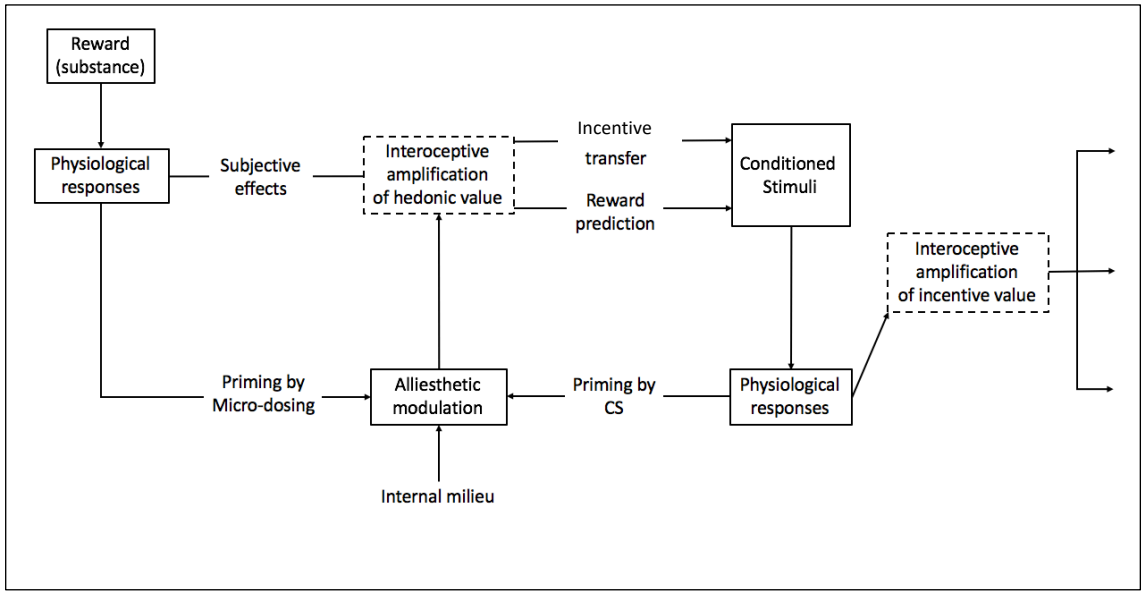


Figure 1: theoretical model describing the role of interoceptive awareness in appetitive learning.

This model of appetitive learning is based on the results obtained in our experiments, also integrating interoceptive models of alliesthesia and craving. Substance administration generates physiological responses deriving into subjective effects. These effects are amplified in highly interoceptive participants, increasing the perception of the hedonic value of rewards and the transfer of incentive values to conditioned stimuli. Interoception, by mediating reward prediction, could also participate in reward approach behaviours. The hedonic component of rewards is also modulated by alliesthetic mechanisms susceptible to interoceptive amplification. This alliesthetic process will be contingent to the internal state of the subject, but can also be primed either by small doses of a reward or by the physiological reactions elicited by conditioned stimuli. The detection and amplification of internal signals generated by

conditioned stimuli will finally drive approach behaviours and drug seeking, hedonic evaluations of conditioned stimuli, and drug urges.

This hypothesized model represents our results, but more evidence is necessary in order to strengthen it. A further discussion on that respect will be found in the limitations section.

6.3 Advances in implicit learning and conditioning

6.3.1 Measures of explicit knowledge

Our experiments had to overcome the existing criticisms in the field of implicit learning, and particularly in PC regarding measurements of stimulus-outcome contingency knowledge (CA) (Lovibond & Shanks, 2002). The first experiment carried out used a simpler methodology in that regard. Participants were categorised as Aware or Unaware of contingencies based on t-tests comparing expectancy ratings towards HR and LR stimuli to chance level. With that methodology, yielding a high number of Unaware participants, we were able to obtain results showing implicit PC effects. Nevertheless, the overrepresentation of implicit states due to lax measurements of awareness is a problem pointed out in the implicit literature (Dienes, 2015), as more participants might be categorised as Unaware of contingencies than there actually are. We therefore improved the methodology used to measure CA, including measurements of metacognitive knowledge. We also performed a two-step Bayesian analysis, with which we could determine the sensitivity of metacognitive states (or Type II d' scores) for each participant using their Type I d' score as a prior. We then

used the mean Type I d' scores of metacognitively aware participants as a prior to determine the individual sensitivity of Type I d' scores of the rest of participants. This allowed us to successfully categorise participants as Aware or Unaware of outcome-contingencies based on sensitive evidence (Dienes, 2014, 2015). The rest of the participants, yielding insensitive results, were not considered. Our approach didn't completely solve Shanks's concerns about post-hoc categorisations of consciousness (2016), but at least considered his recommendation of using Bayes factors in a participant-by-participant basis.

This statistical approach can be applied to other paradigms used in the field of implicit learning, but also to the study of consciousness in general, for example to ascertain the ability or inability to detect subliminal stimuli, offering a more reliable account of unconscious processes.

6.3.2 Measures of implicit knowledge

The main rationale for including attentional paradigms, such as the Flanker task, Emotional Attentional Blink or n-back task, was to examine implicit learning with different procedures able to evaluate automatic responses. We assumed from the beginning that pleasantness measurements, reflecting explicit appraisals, might not be the most sensitive way to measure implicit learning, and we were right. We only managed to obtain significant increases in pleasantness towards HR stimuli in Aware participants. For Unaware participants, all evidence was insensitive, as shown with Bayes factors.

These differential findings exemplify how important it is to use measures of learning sensitive to the kind of information or construct targeted (Houwer, 2006). We were able to find preferential attentional responses towards implicitly CS in two tasks, one based on goal-directed target detection and the effect of implicit values of CS (Emotional Attentional Blink - Raymond, Shapiro, & Arnell, 1992), and another one on task irrelevant cognitive interferences by CS (Flanker task - Eriksen, 1995). This implies that different paradigms need to be used to study attentional reactivity in implicit learning (see also Anderson, 2015). Further clarification of how implicit attentional biases are generated might be obtained by the use of other kinds of tasks.

6.4 Limitations and future directions

Although the experiments presented in this thesis apply innovative solutions, allowing the attainment of novel results, they also indicate the need for improvements to be addressed in future studies.

For example, we found that interoception mediates hedonic reactions, subjective substance effects and reward prediction. However, we did not report physiological responses during any of these procedures at any point. Having done so would have allowed us to better understand the relationship between interoception and physiological responses towards reward CS. Classical experiments on interoception and emotion show that heart-rate changes depend on interoceptive abilities (e.g. Pollatos, Herbert, Matthias, & Schandry, 2007). Replicating those results in the context of PC would increase our understanding of the three-way relationship between emotion, interoception and physiological responses.

That same problem is present in our experiments examining implicit learning. The addition of any physiological measures (heart-rate, skin conductance or pupilometry) would have provided us with a supplementary measure of automatic reactivity, independently from explicit learning measures.

Future studies should examine learning effects in a PC task in parallel with eye-gaze and pupilometry measurements so that objective physiological state changes accompanying learning can be assessed. We have carried out such a study (also another one measuring heart rate changes), which are still undergoing data analysis.

Bayesian analyses, included in our CA evaluations, provided a much more reliable awareness categorisation technique, although this implied that a large proportion of the sample was discarded due to their results being insensitive. We attempted to tackle that problem by increasing the amount of expectancy measurements taken, but without success. Maybe a more sensitive way to measure Type I expectancies and confidence (using visual analogue or Likert scales) would solve the matter, however, this would make it very difficult to obtain categorical measures of accuracy or confidence. On the positive side, mean differences or effect sizes obtained in previous experiments can also be introduced as priors on Bayesian analyses. Now that significant results are available using this paradigm, future experiments might not need complicated two-step Bayesian categorisations using meta-cognitive awareness, but could just use the scores obtained previously as priors for new categorisations based on continuous or Likert scale expectancy ratings, respecting one of the main proposals of Lovibond and Shanks (2002).

The results obtained in the n-back task, showing that CA was necessary for CS to interfere with cognitive control, and the results of the Emotional Attentional Blink, in which Aware participants seemed to pay preferential attention to LR CS compared to HR, are puzzling. Moreover, on the Flanker task, Aware participants did not show any learning effects. We attempted to provide a series of explanations for both cases, but their theoretical or experimental support is limited. If anything, the fact that we were able to obtain in three instances very clear learning effects in Unaware participants, and somewhat weaker results in Aware, supports the necessity to explore automatic learning processes with implicitly CS rather than with stimuli explicitly related to a reward.

Perhaps the most daunting piece of evidence that needs clarification is the role of attentional responses in implicit PC. If implicit PC cannot generate subjective hedonic responses or behavioural effects (i.e. PIT), why would CS trigger such attentional biases? And what could be the particular role they might play in driving addiction? One possibility is that they participate in the generation of preparatory responses. In that case, the link between attention, physiological responses and reward appraisal or effects should be observed. Maybe a more careful examination of implicit learning effects in PIT paradigms could bring evidence supporting the role of CS in overt consumption behaviours. Another possibility is that CS could drive attention depending on the development of CA. If a participant is not explicitly aware of contingencies, an implicit mechanism based on the transfer of incentive values would drive attentional responses towards CS predicting rewards until outcome-contingencies become conscious (i.e. the subject “realises” the relationship between

CS and reward). This possibility was outlined in our first experiments and would reconcile the views of Pearce and Mackintosh (2010) on associative learning. Pearce and Hall (1980) posited that attention towards a stimulus is only required whilst still learning about its consequences, afterwards attention would be allocated towards other stimuli in search for novel contingencies. Mackintosh (1974) proposed on the other hand that stimuli predicting rewards will always drag more attention. In our Emotional Attentional Blink, we found that Unaware participants paid more attention to HR CS, following Mackintosh's idea. On the contrary, Aware participants showed preferential responses towards LR CS, disregarding CS predicting high probabilities of reward, in accordance to Pearce and Hall. This might point towards the role of CA in the allocation of attention in associative learning, but such an idea requires further examination.

As much as the conclusions extracted from the alcohol administration experiment were interesting, the results regarding subjective light-headedness and interoception could benefit from the examination of the awareness subjects had of the type of drink consumed, something impossible to achieve reliably with a single trial (one drink administration). For that reason, we did not obtain this measure. Getting to know whether interoception can facilitate the discrimination of alcohol drinks compared to placebo would provide us with a mechanism explaining substance discrimination abilities. Light-headedness is the main factor explaining alcohol discrimination in social drinkers (Duka, Stephens, Russell, & Tasker, 1998), the hypotheses seems to make sense, but an experimental demonstration would involve lengthy procedures, requiring 2 or 3 day-long sessions per participant.

We found discrepancies between the different measures of interoceptive awareness and the way the interacted with learning or alcohol administration. Meta-cognitive discrimination explained emotional reactivity towards CS and light-headedness during alcohol administration, meta-cognitive tracking on the other hand explained reward prediction. Alcohol administration also decreased meta-cognitive discrimination and tracking accuracy only for males. These inconsistencies follow the lack of evidence for a correlation between interoceptive factors (Garfinkel et al., 2016; Garfinkel, Seth, Barrett, Suzuki, & Critchley, 2015) added to recent criticisms to the use of tracking tasks (Zamariola, Maurage, Luminet, & Corneille, 2018). Further experiments should seek either for control tasks capable of overcoming some of the limitations related to interoceptive measurements, or for novel techniques able to measure interoceptive awareness in a more consistent way.

Finally, although the model of interoceptive appetitive learning is partly supported by our results and previous theoretical accounts, more evidence regarding the effects of CS and micro-dosing in alliesthesia and their link to addiction mediated by interoception is necessary. That could be studied by administering very low doses of alcohol to social drinkers, assessing the effect of that priming on craving and attentional biases (i.e. Schoenmakers, Wiers, & Field, 2008), and evaluating the mediator effect of interoception. It would be relevant as well to test Gray and Critchley's (2007) hypothesis concerning drug cravings and to examine the role of interoception in attentional or behavioural responses towards reward-related stimuli.

6.5 Conclusions

The research presented here aimed at disentangling the implicit correlates of appetitive PC. For that purpose, we developed a series of novel techniques able to measure both subjective and objective learning in a sensitive manner. Our main finding, that implicit CS can generate responses, is a novel characterisation of PC, but has also multiple implications for the field of addiction. Using the methodologies designed here we can achieve a better understanding of the implicit correlates of addictive processes, which we may ultimately be able to translate to addiction therapies.

The examination of interoceptive correlates of appetitive learning also has important implications. Our results provide some evidence for the role of bodily sensations in appetitive learning; such a finding can be integrated in learning theories and provide a more novel conceptualisation of addiction.

6.6 References

Abrams, K., Cieslowski, K., Johnson, S., Krimmel, S., La Rosa, G. B.-D., Barton, K., & Silverman, P. (2018). The effects of alcohol on heartbeat perception: Implications for anxiety. *Addictive Behaviors*, *79*, 151–158. http://doi.org/10.1016/J.ADDBEH.2017.12.023

Anderson, B. A. (2015a). Value-driven attentional capture is modulated by spatial context. *Visual Cognition*, *23*(1–2), 67–81. http://doi.org/10.1080/13506285.2014.956851

Anderson, B. A. (2015b). Value-driven attentional priority is context specific. *Psychonomic Bulletin & Review*, *22*(3), 750–756. http://doi.org/10.3758/s13423-014-0724-0

Bates, M. E., Buckman, J. F., Vaschillo, E. G., Fonoberov, V. A., Fonoberova, M., Vaschillo, B., … Mezić, I. (2011). The Redistribution of Power: Neurocardiac Signaling, Alcohol and Gender. *PLoS ONE*, *6*(12), e28281. http://doi.org/10.1371/journal.pone.0028281

Berridge, K. C. (1999). Pleasure, pain, desire, and dread: Hidden core processes of

emotion. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations ofhedonic psychology* (pp. 525–557). New York: Russell Sage Foundation.

Berridge, K. C., & Robinson, T. E. (2003). Parsing reward. *Trends in Neurosciences*, *26*(9), 507–13. http://doi.org/10.1016/S0166-2236(03)00233-9

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *The American Psychologist*, *71*(8), 670–679. http://doi.org/10.1037/amp0000059

Berridge, K., & Winkielman, P. (2003). What is an unconscious emotion?(The case for unconscious “liking”). *Cognition & Emotion*, *17*(2), 181–211. http://doi.org/10.1080/02699930302289

Bindra, D. (1968). Neuropsychological interpretation of the effects of drive and incentive-motivation on general activity and instrumental behavior. *Psychological Review*, *75*(1), 1–22. http://doi.org/10.1037/h0025306

Bindra, D. (1974). A motivational view of learning, performance, and behavior modification. *Psychological Review*, *81*(3), 199–213. http://doi.org/10.1037/h0036330

Bindra, D. (1978). How adaptive behavior is produced: a perceptual-motivational alternative to response reinforcements. *Behavioral and Brain Sciences*, *1*(01), 41. http://doi.org/10.1017/S0140525X00059380

Bolles, R. C. (1972). Reinforcement, expectancy, and learning. *Psychological Review*, *79*(5), 394–409. http://doi.org/10.1037/h0033120

Bourgeois, A., Neveu, R., & Vuilleumier, P. (2016). How Does Awareness Modulate Goal-Directed and Stimulus-Driven Shifts of Attention Triggered by Value Learning? *PLOS ONE*, *11*(8), e0160469. http://doi.org/10.1371/journal.pone.0160469

Cabanac, M. (1979). Sensory Pleasure. *The Quarterly Review of Biology*, *54*(1), 1–29. http://doi.org/10.1086/410981

Critchley, H., & Garfinkel, S. (2017). Interoception and emotion. *Current Opinion in Psychology*, *17*, 7–14. http://doi.org/10.1016/J.COPSYC.2017.04.020

Dickinson, A. (1989). *Expectancy theory in animal conditioning.*

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18. http://doi.org/10.3758/BF03199951

Dickinson, A., & Balleine, B. (1995). Motivational Control of Instrumental Action. *Current Directions in Psychological Science*, *4*(5), 162–167. http://doi.org/10.1111/1467-8721.ep11512272

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. http://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental

states are unconscious. *Behavioural Methods in Consciousness Research*, 199–220.

Duka, T., Stephens, D. N., Russell, C., & Tasker, R. (1998). Discriminative stimulus properties of low doses of ethanol in humans. *Psychopharmacology*, *136*(4), 379–389. http://doi.org/10.1007/s002130050581

Duka, T., & Townshend, J. M. (2004). The priming effect of alcohol pre-load on attentional bias to alcohol-related stimuli. *Psychopharmacology*, *176*(3–4), 353–61. http://doi.org/10.1007/s00213-004-1906-7

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, *2*(2–3), 101–118. http://doi.org/10.1080/13506289508401726

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. http://doi.org/10.1038/nn1579

Füstös, J., Gramann, K., Herbert, B. M., & Pollatos, O. (2013). On the embodiment of emotion regulation: interoceptive awareness facilitates reappraisal. *Social Cognitive and Affective Neuroscience*, *8*(8), 911–917. http://doi.org/10.1093/scan/nss089

Garfinkel, S. N., Barrett, A. B., Minati, L., Dolan, R. J., Seth, A. K., & Critchley, H. D. (2013). What the heart forgets: Cardiac timing influences memory for words and is modulated by metacognition and interoceptive sensitivity. *Psychophysiology*, *50*(6), 505–512. http://doi.org/10.1111/psyp.12039

Garfinkel, S. N., Manassei, M. F., Hamilton-Fletcher, G., In den Bosch, Y., Critchley, H. D., & Engels, M. (2016). Interoceptive dimensions across cardiac and respiratory axes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1708), 20160014. http://doi.org/10.1098/rstb.2016.0014

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. http://doi.org/10.1016/j.biopsycho.2014.11.004

Gray, M. A., & Critchley, H. D. (2007). Interoceptive Basis to Craving. *Neuron*, *54*(2), 183–186.

Hart, C., Ward, A., Haney, M., Foltin, R., & Fischman, M. (2001). Methamphetamine self-administration by humans. *Psychopharmacology*, *157*(1), 75–81. http://doi.org/10.1007/s002130100738

Hester, R., & Garavan, H. (2009). Neural mechanisms underlying drug-related cue distraction in active cocaine users. *Pharmacology Biochemistry and Behavior*, *93*(3), 270–277. http://doi.org/10.1016/J.PBB.2008.12.009

Hogarth, L., & Chase, H. W. (2011). Parallel goal-directed and habitual control of human drug-seeking: implications for dependence vulnerability. *Journal of Experimental Psychology. Animal Behavior Processes*, *37*(3), 261–76.

http://doi.org/10.1037/a0022913

Hogarth, L., Dickinson, A., Hutton, S. B., Bamborough, H., & Duka, T. (2006). Contingency knowledge is necessary for learned motivated behaviour in humans: relevance for addictive behaviour. *Addiction (Abingdon, England)*, *101*(8), 1153–66. http://doi.org/10.1111/j.1360-0443.2006.01459.x

Hogarth, L., Dickinson, A., Wright, A., Kouvaraki, M., & Duka, T. (2007). The role of drug expectancy in the control of human drug seeking. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(4), 484–496. http://doi.org/10.1037/0097-7403.33.4.484

Houwer, J. De. (2006). What are implicit measures and why are we using them. *The Handbook of Implicit Cognition and Addiction*. R

Hull, C. L. (1943). Principles of Behavior. An Introduction to Behavior Theory. *The Journal of Philosophy*, *40*(20), 558. http://doi.org/10.2307/2019960

Jeffs, S., & Duka, T. (2017). Predictive but not emotional value of Pavlovian stimuli leads to pavlovian-to-instrumental transfer. *Behavioural Brain Research*, *321*, 214–222. http://doi.org/10.1016/j.bbr.2016.12.022

Jellinek, E. M. (1955). The craving for alcohol. *Quarterly Journal of Studies on Alcohol*, *16*(1), 35–8.

Kandasamy, N., Garfinkel, S. N., Page, L., Hardy, B., Critchley, H. D., Gurnell, M., & Coates, J. M. (2016). Interoceptive Ability Predicts Survival on a London Trading Floor. http://doi.org/10.1038/srep32986

Katkin, E. S., Wiens, S., & Ohman, A. (2001). Nonconscious Fear Conditioning, Visceral Perception, and the Development of Gut Feelings. *Psychological Science*, *12*(5), 366–370. http://doi.org/10.1111/1467-9280.00368

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26.

Mackintosh, N. (1974). The psychology of animal learning.

Marlatt, G., & Donovan, D. (Eds.). (2005). *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors*. Guildford press.

Miles, F. J., Everitt, B. J., & Dickinson, A. (2003). Oral cocaine seeking by rats: action or habit? *Behavioral Neuroscience*, *117*(5), 927–38. http://doi.org/10.1037/0735-7044.117.5.927

Naqvi, N. H., & Bechara, A. (2010). The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. *Brain Structure & Function*, *214*(5–6), 435–50. http://doi.org/10.1007/s00429-010-0268-7

Nikolaou, K., Field, M., Critchley, H., & Duka, T. (2013). Acute Alcohol Effects on Attentional Bias are Mediated by Subcortical Areas Associated with Arousal and Salience Attribution. *Neuropsychopharmacology*, *38*(7), 1365–1373. http://doi.org/10.1038/npp.2013.34

Nikolaou, K., Field, M., & Duka, T. (2013). Alcohol-related cues reduce cognitive control in social drinkers. *Behavioural Pharmacology*, *24*(1), 29–36. http://doi.org/10.1097/FBP.0b013e32835cf458

Paulus, M. P., & Stewart, J. L. (2014). Interoception and drug addiction. *Neuropharmacology*, *76 Pt B*, 342–50. http://doi.org/10.1016/j.neuropharm.2013.07.002

Paulus, M. P., Tapert, S. F., & Schulteis, G. (2009). The role of interoception and alliesthesia in addiction. *Pharmacology, Biochemistry, and Behavior*, *94*(1), 1–7. http://doi.org/10.1016/j.pbb.2009.08.005

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. http://doi.org/10.1037/0033-295X.87.6.532

Pearce, J., & Mackintosh, N. (2010). Two theories of attention: A review and a possible integration. *Attention and Associative Learning:*

Pfeifer, G., Garfinkel, S. N., Gould van Praag, C. D., Sahota, K., Betka, S., & Critchley, H. D. (2017). Feedback from the heart: Emotional learning and memory is controlled by cardiac cycle, interoceptive accuracy and personality. *Biological Psychology*, *126*, 19–29. http://doi.org/10.1016/j.biopsycho.2017.04.001

Pollatos, O., Herbert, B. M., Matthias, E., & Schandry, R. (2007). Heart rate response after emotional picture presentation is modulated by interoceptive awareness. *International Journal of Psychophysiology*, *63*(1), 117–124. http://doi.org/10.1016/J.IJPSYCHO.2006.09.003

Pollatos, O., & Schandry, R. (2008). Emotional processing and emotional memory are modulated by interoceptive awareness. *Cognition & Emotion*, *22*(2), 272–287. http://doi.org/10.1080/02699930701357535

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849–860. http://doi.org/10.1037/0096-1523.18.3.849

Schoenmakers, T., Wiers, R. W., & Field, M. (2008). Effects of a low dose of alcohol on cognitive biases and craving in heavy drinkers. *Psychopharmacology*, *197*(1), 169–178. http://doi.org/10.1007/s00213-007-1023-5

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology*, *60*(3), 291–309. http://doi.org/10.1080/17470210601000581

Shanks, D. R. (2016). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. http://doi.org/10.3758/s13423-016-1170-y

Siegel, S. (1975). Evidence from rats that morphine tolerance is a learned response. *Journal of Comparative and Physiological Psychology*, *89*(5), 498–506.

Sokol-Hessner, P., Hartley, C. A., Hamilton, J. R., & Phelps, E. A. (2015). Interoceptive ability predicts aversion to losses. *Cognition and Emotion*, *29*(4), 695–701. http://doi.org/10.1080/02699931.2014.925426

Stewart, J. L., May, A. C., Tapert, S. F., & Paulus, M. P. (2015). Hyperactivation to pleasant interoceptive stimuli characterizes the transition to stimulant addiction. *Drug and Alcohol Dependence*, *154*, 264–270. http://doi.org/10.1016/J.DRUGALCDEP.2015.07.009

Tiffany, S. T. (1990). A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes. *Psychological Review*, *97*(2), 147–68. http://doi.org/10.1037/0033-295X.97.2.147

Toates, F. M. (Frederick M. . (1986). *Motivational systems*. Cambridge University Press.

Verdejo-Garcia, A., Clark, L., & Dunn, B. D. (2012). The role of interoception in addiction: A critical review. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1857–1869. http://doi.org/10.1016/j.neubiorev.2012.05.007

Wetherill, R. R., Childress, A. R., Jagannathan, K., Bender, J., Young, K. A., Suh, J. J., … Franklin, T. R. (2014). Neural responses to subliminally presented cannabis and other emotionally evocative cues in cannabis-dependent individuals. *Psychopharmacology*, *231*(7), 1397–1407. http://doi.org/10.1007/s00213-013-3342-z

Wiers, R. W., & Stacy, A. W. (2006). Implicit Cognition and Addiction. *Current Directions in Psychological Science*, *15*(6), 292–296. http://doi.org/10.1111/j.1467-8721.2006.00455.x

Wiers, R. W., Stacy, A. W., Ames, S. L., Noll, J. A., Sayette, M. A., Zack, M., & Krank, M. (2002). Implicit and Explicit Alcohol-Related Cognitions. *Alcoholism: Clinical and Experimental Research*, *26*(1), 129–137. http://doi.org/10.1111/j.1530-0277.2002.tb02441.x

Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality & Social Psychology Bulletin*, *31*(1), 121–35. http://doi.org/10.1177/0146167204271309

Zamariola, G., Maurage, P., Luminet, O., & Corneille, O. (2018). Interoceptive accuracy scores from the heartbeat counting task are problematic: Evidence from simple bivariate correlations. *Biological Psychology*, *137*, 12–17. http://doi.org/10.1016/J.BIOPSYCHO.2018.06.006