Perceiving Agency in a Multi-Agent Context: More Interactions with the Environment yet Reduced Agency in Depression



Riddhi Jain

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Experimental Psychology

Primary Supervisor: Prof. Robin A. Murphy

Secondary Supervisors: Prof. Karl J. Friston, Prof. Nicholas Yeung

Table of Contents

List of Figures 4
List of Tables 6
Acknowledgements 7
Declaration 8
Note to Readers 10
Abstract 11
Chapter 1: Introduction 13
Introduction to Agency 14
Depression and Agency 17
Thesis Overview 18
Chapter 2: Trends in Perceiving Contingencies 23
Experimental Methods 24
Single Action-Outcome Contingency Learning 27
Learning Curves 28
Inter-Trial Intervals 29
The Role of D Events 33
Single and Ready to Mingle: The Role of Others 38
Blocking 40
Relative Validity 41
Summary 43
Depression and Agency 45
Depressive Realism 47
Chapter 3: Mechanisms of Contingency Learning 52
Statistical Accounts 53
Associative Accounts 60
Inferential Accounts 67
Chapter 4: Modifying Contingency Learning through D Event Manipulation (Experiments 1-3) 72
Experiment 1 77
Participants 77
Materials and Procedure 78
Task Design 80
Data Analysis 83
Results 86
Discussion 90
Experiment 2 92
Participants 92

Task Design 92
Data Analysis 93
Results 94
Discussion 97
Experiment 3 99
Participants 99
Task Design 99
Data Analysis 100
Results 100
General Discussion 105
Chapter 5: Explaining Actions 114
Active Inference 120
Chapter 6: Agency Learning in a Multi-Agent Context (Experiments 4-6) 125
Experiment 4 129
Participants 129
Materials, Procedure, and Task Design 130
Data Analysis 134
Results 143
Discussion 156
Experiment 5 158
Participants 158
Materials, Procedure, and Task Design 161
Data Analysis 162
Results 163
Discussion 182
Experiment 6 185
Participants 185
Materials, Procedure, and Task Design 187
Data Analysis 188
Results 190
Discussion 203
Chapter 7: General Discussion 206
References **215**
Supplementary Materials **229**
Supplementary Material 1. 230
Supplementary Material 2. 233
Supplementary Material 3. 235
Supplementary Material 4. 241
Supplementary Material 5. 244
Supplementary Material 6. 245

List of Figures

Figure 1. *How Inter-Trial Intervals Can Increase Perceived Contingency.* 32

Figure 2. *Illustration of A, B, C, and D Trials in a 2x2 Contingency Matrix Framework.* 35

Figure 3. *The Task for Experiments 1-3.* 81

Figure 4. *Results from Experiment 1: Ratings as a Function of D Trial Frequency, D Trial Content, Baseline Contingencies, and Depressive Symptom Levels.* 87

Figure 5. *Results from Experiment 2: Ratings as a Function of D Trial Frequency, D Trial Content, and Depressive Symptom Levels.* 95

Figure 6. *Results from Experiment 3: Ratings as a Function of Baseline Contingencies, D Trial Content, and Depressive Symptom Levels.* 103

Figure 7. *Results from Experiment 3: Ratings as a Function of D Trial Frequency, Block Order, and Depressive Symptom Levels.* 105

Figure 8. *The Task for Experiments 4-6.* 131

Figure 9. *The Active Inference Model for Experiments 4-6.* 141

Figure 10. *Results from Experiment 4: Ratings Across Experimental Conditions by Depressive Symptom Severity Group.* 144

Figure 11. *Results from Experiment 4: Button-Pressing Probability Across Trials by Depressive Symptom Severity Group.* 148

Figure 12. *Results from Experiment 4: Ratings and Experienced ΔPs Across Experimental Conditions, by Depressive Symptom Group.* 150

Figure 13. *Experiment 5: Distribution of Total Depression (BDI-II) Scores for Clinically and Non-Clinically Depressed Groups.* 160

Figure 14. *Results from Experiment 5: Ratings Across Experimental Conditions Where the Self Had Control, by Depression Group.* 165

Figure 15. *Results from Experiment 5: Button-Pressing Probability Across Trials by Depression Group.* 170

Figure 16. *Results from Experiment 5: Mean Ratings and Experienced ΔPs Across Experimental Conditions, by Depression Group.* 173

Figure 17. *Experiment 6: Distribution of Total Depression (BDI-II) Scores for Clinically and Non-Clinically Depressed Groups*. 187

Figure 18. *Results from Experiment 6: Ratings Across Experimental Conditions Where the Self Had Control, by Depression Group.* 193

Figure 19. *Results from Experiment 6: Button-Pressing Probability Across Trials by Depression Group.* 196

Figure 20. *Results from Experiment 6: Ratings and Experienced ΔPs Across Experimental Conditions, by Depression Group.* 198

List of Tables

Table 1. *2x2 Matrix Illustrating Four Types of Information: A, B, C and D.* 25

Table 2. *Task Design for Experiment 1.* 83

Table 3. *Task Design for Experiment 2.* 93

Table 4. *Task Design for Experiment 3.* 100

Table 5. *Results from Experiment 4: Active Inference Model Fitting Results for Low and High Depressive Symptom Groups.* 154

Table 6. *Experiment 5: Demographic Information from the Three Sample Groups.* 159

Table 7. *Results from Experiment 5: Active Inference Model Fitting Results for Non-Clinically and Clinically Depressed Groups.* 180

Table 8. *Experiment 6: Demographic Information from the Three Sample Groups.* 186

Table 9. *Results from Experiment 6: Active Inference Model Fitting Results for Non-Clinically and Clinically Depressed Groups.* 201

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Robin, my academic supervisor and life mentor. Thank you for teaching me how to navigate curiosity while enjoying every part of the journey. I am honoured to be part of your academic family, and I hope to maintain a bond as strong as the one you share with Andy, while carrying on traditions like Gregynog. I am also deeply grateful to Karl and Tim for their mentorship, friendship, and of course, unconditional positivity. The three of you have shown me how to balance intellectual sharpness with humour, making even the most challenging moments enjoyable. Your guidance has left a lasting impression on me, and I aspire to approach both my work and life with the same spirit.

I am especially thankful to Maxwell for introducing me to active inference and its community, and to Nick for balancing this with his persistent and invaluable questioning of "why active inference". To my friends and colleagues Ozan, Toon, Dimi, and Conor, thank you for always being approachable and patient, and generously sharing your insights and knowledge. Knowing such incredible people reminds me of how much there is yet to learn, and I am excited to continue this journey with such a supportive community by my side.

Next, I would like to thank Oxford for being my first home away from home. I am immensely grateful for all it has given me – cherished memories, and personal and professional growth. A home isn't complete without family, and I am lucky to have found mine in Camille, Nilgoun, Vedang, and Will – thank you for being such an integral part of my life here and for filling my days with such good quality companionship. To my dear friends, Sam W, Malin, Ramani, Sam R, Romain, and Katja, thank you for being such special people who have each played a significant role in my time at Oxford.

Finally, to my family: Papa, thank you for teaching me the value of hard work, discipline, and balancing it all with masti. Mumma, thank you for teaching me the importance of kindness and care. Vaibhav, thank you for showing me what it means to be a go-getter. Advait, thank you for being a constant source of support, no matter the distance, and for your drive and resilience, which have truly inspired me.

P.S. I'd also like to thank The XX for creating the Intro and for whoever created the 10-hour seamless version of the track because it has gotten me through GCSEs, IB, Undergrad exams, and now, this PhD thesis.

Declaration

This thesis comprises work conducted by Riddhi Jain in fulfilment of the degree of Doctor of Philosophy in Experimental Psychology, University of Oxford. Riddhi Jain led the conceptualisation, design, recruitment, data collection, data analyses, visualisations, interpretation of results, discussions, and writing of all work presented in this thesis. The contributions of collaborators and relevant organisations are detailed as follows:

- Chapter 4 includes experiments where Prof. Ralph Miller contributed to the conceptualisation of the experimental task and the interpretation of results.
- In Experiment 6, as part of Chapter 6, Dr. Paolo Ossola and his research assistant team in Ferrara, Italy, assisted with data collection and result interpretation.
- Prof. Karl Friston provided expertise in designing the active inference model used in Experiments 4-6 (Chapter 6).
- Dr. Santiago Castiello assisted with the conceptualisation of the task used in Experiments 4-6 (Chapter 6).
- Prof. Robin Murphy has supervised all work across all chapters.
- VERSES Inc. funded the second to fourth years of the PhD.

Additional contributions are captured in the publication list below.

Relevant publications (in chronological order):

- Castiello, S., Miller, R. R., Witnauer, J. E., Alcaide, D. M., Fung, E., Pitliya, R. J., Morrissey, D. K. C., & Murphy, R. A. (2022). Benefiting from trial spacing without the cost of prolonged training: Frequency, not duration, of trials with absent stimuli enhances perceived contingency. *J Exp Psychol Gen*, *151*(8), 1772-1792. https://doi.org/10.1037/xge0001166

- Albarracin, M., & Pitliya, R. J. (2022). The nature of beliefs and believing [Opinion]. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.981925
- Pitliya, R.J., Murphy, R.A. (2023). A Model of Agential Learning Using Active Inference. In: Buckley, C.L., *et al.* Active Inference. IWAI 2023. Communications in Computer and Information Science, vol 1915. Springer, Cham. https://doi.org/10.1007/978-3-031-47958-8_8
- Witnauer, J. E., Castiello, S., Fung, E., Pitliya, R. J., Murphy, R. A., & Miller, R. R. (2023). Determinants of extinction in a streamed trial procedure. *Q J Exp Psychol (Hove)*, *76*(5), 1155-1176. https://doi.org/10.1177/17470218221110827
- Pitliya, R.J., Markovic, D., Folesani, F., Murri, M.B., Castiello, S., Murphy, R.A. (2024). Modelling Agency Perception in a Multi-Agent Context in Depression Using Active Inference. In: Buckley, C.L., *et al.* Active Inference. IWAI 2024. Communications in Computer and Information Science, vol 2193. Springer, Cham. (Preprint available at: https://osf.io/preprints/osf/ckfrt, DOI pending)

Parts of the research have been presented at the following conferences:

- Associative Learning Symposium 2023
- International Workshop on Active Inference 2023
- Associative Learning Symposium 2024
- International Workshop on Active Inference 2024

Note to Readers

To facilitate electronic reading, the Table of Contents in the PDF file allows for easy navigation between chapters and sub-sections. References and supplementary materials are placed at the end of the thesis for easy access. Additionally, hyperlinks are embedded throughout the document, enabling you to click directly on references to tables and figures (e.g., “see Figure X”) to jump to their location. To return to your previous spot in the document, you can use ⌘ +] on Mac or Alt + ← on Windows/Linux. To move forward again, use ⌘ + [on Mac or Alt + → on Windows/Linux, making it simple to toggle between locations.

I hope you enjoy reading this thesis as much as I enjoyed writing it!

Abstract

Understanding agency – the extent to which actions control outcomes – is crucial for comprehending human behaviour and mental health, particularly in major depressive disorder where reduced agency is a key symptom. This thesis investigates how individuals perceive action-outcome contingencies in multi-agent environments, shedding light on individual differences in the acquisition and potential loss of agency perception.

Research shows that individuals tend to perceive positive action-outcome relations when both action and outcome are present, and these relations are weakened when only one occurs. However, people often neglect events where neither action nor outcome occurs, despite their logical relevance in informing positive relations. This thesis examines whether such co-absence events become more salient when other potential causes are present.

Experiments 1-3 examined contingency learning by manipulating the frequency and content of co-absence events. Results revealed that increasing co-absence events reduced perceived contingency, especially when filled with irrelevant stimuli. Participants with higher depressive symptoms demonstrated less sensitivity to these events and reported weaker contingencies overall, notably for negative (inhibitory) relations.

Experiments 4-6 extended these findings to agency perception in multi-agent contexts. Depressed participants perceived lower agency for themselves and others, again, especially in scenarios involving negative relations. Non-depressed participants exhibited a bias where perceiving more control in one agent led to attributing opposing control to another, a bias less evident in depressed participants as they tended to attribute control to the other agent in the same direction.

Behavioural differences emerged: depressed participants engaged in more frequent but aimless actions, while non-depressed participants' actions were fewer but more goal-directed.

When assisted in tracking the number of outcomes produced, depressed participants' actions became more goal-directed and less frequent.

These findings provide valuable insights into agency perception mechanisms, with potential implications for informing interventions. While acknowledging limitations, the methodology and results from these studies establish a foundation for future research in this critical area of mental health and human behaviour.

Chapter 1:

Introduction

Introduction to Agency

An agent is some*one* or some*thing* that controls their actions and influences outcomes in their environment. Agency refers to the agent's capacity to exert control over events in their environment. Perceiving agency involves developing beliefs about one's level of control. Having an ongoing registration of the degree of agency that one has is essential for recognising the causal structure of the environment and facilitating goal-directed behaviours at both individual and group levels (Verschure et al., 2014).

The sense of agency can be understood through two dimensions: implicit and explicit senses of agency (Moore et al., 2012; Synofzik et al., 2008). The implicit sense of agency involves determining whether the cause of an action and its resulting outcome is attributed to oneself (Gallagher, 2000; Pacherie, 2007). This focuses on recognising that the self, rather than another person or external force, is the initiator and controller of the action and thereby the outcome. It has been suggested that this form of agency is based on matching of predicted and actual sensory feedback, dichotomously tagging experiences of action as self-caused or not (Carruthers, 2012; Haggard, 2005; Synofzik et al., 2008).

Experimental psychologists have explored the implicit sense of agency through various methods. One such example is the self-produced tickle paradigm, which demonstrated that sensory feedback from self-generated actions is attenuated, meaning the sensation is weaker when one tries to tickle themselves compared to being tickled by another person (Blakemore et al., 1998). Researchers have also investigated how delays between an action and its outcome affected the implicit sense of agency, finding that longer delays weakened the perception of control (Chambon et al., 2015; Shanks et al., 1989; Wen, 2019). Another widely used method is the intentional binding task, where participants perceived the time between their action and the outcome as shorter when they felt in control, reflecting a stronger sense of agency (Haggard et al., 2002; Moore & Obhi, 2012).

The explicit sense of agency, which this thesis will focus on, concerns the perceived strength or degree to which an agent can influence events in their environment (Bandura, 1982; Friston et al., 2013). From this point on, the term "agency" will refer to the explicit sense of agency – an individual's perception and evaluation of their control over outcomes. Rather than a binary concept, this aspect of agency can be understood as existing on a spectrum, referring to the degree of control the agent has over generating or preventing the occurrence of an outcome (Allan, 1993; Chatlosh et al., 1985; Shanks, 1993). The degree of control can be represented on a scale from -1 to +1, similar to correlations. A positive relation (+1) indicates that an action consistently predicts an outcome (e.g., pressing a button turns on a light). A negative relation (-1) means an action consistently signals the absence of an outcome (e.g., pressing a button turns off a light). A value of zero indicates no relationship between the action and the outcome (e.g., pressing a button does not affect the light). Values at the extremes (-1 or +1) represent deterministic relationships, where actions consistently predict outcomes, and all other values represent probabilistic relationships, where actions predict outcomes only some of the time.

This evaluation of agency extends beyond mere sensory feedback and incorporates a broader understanding of the causal relationships between actions and outcomes. It has been suggested that individuals come to understand the structure of their control over the environment by interacting with it and observing the consequences of their actions. Through this process, they learn how contingent outcomes are on their actions (Baker et al., 2005; Johnson-Laird & Khemlani, 2017). This is where action-outcome contingencies – the relationship between actions and their outcomes – become central to the perception of agency.

An action-outcome contingency refers to the extent to which an outcome depends on both the presence and absence of a specific action. For example, consider pressing a button to turn on a light. If the light consistently turns on only when the button is pressed and does not

turn on when the button is not pressed, there is a strong action-outcome contingency. This consistency reinforces the perception of a causal link between the action (pressing the button) and the outcome (the light turning on). In contrast, probabilities alone do not capture this full relationship. If the light turns on randomly – sometimes after pressing the button and sometimes without pressing it – the probability of the light turning on may be high, but the contingency is weak. The lack of dependence on the action (i.e., the light turning on without pressing the button) undermines the perceived sense of control. In this way, the strength of the contingency, which relies on the outcome being tied to both the action and the lack of action, affects how much control the individual perceives they have over the light's behaviour. Indeed, inaction can play an equally important role in learning these contingencies. As Warren Buffett aptly stated, *"Inactivity strikes us as an intelligent behaviour"* (Buffett, 1997). This perspective highlights that sometimes refraining from action can enhance one's understanding of their environment and foster better decision-making as the absence of action yields valuable insights.

By recognising consistent patterns between actions and outcomes, individuals can develop a clearer sense of control over their environment. This learning process is crucial for goal-directed behaviour, as it allows agents to predict and influence the results of their actions, enabling them to achieve desired outcomes. Stable and predictable contingencies empower agents to refine their actions and make informed decisions, fostering a stronger sense of agency and enhancing their ability to interact effectively with their surroundings (Friston et al., 2013; Verschure et al., 2014). Thus, learning action-outcome contingencies plays a foundational role in both agency and adaptive behaviour, allowing individuals to engage with their environment in a controlled and purposeful manner.

Studying agency in multi-agent contexts is essential because real-world environments are rarely confined to a single agent. In such contexts, individuals must not only assess their own agency but also distinguish it from the agency of others (Declerck et al., 2006). This

differentiation is crucial for understanding one's control, but does not always lead to a more accurate assessment of that control (Wolpert et al., 2003). For example, the presence of others can diffuse responsibility, making it less likely for an individual to act in emergencies, a phenomenon known as the bystander effect (Chekroun & Brauer, 2002; Darley & Latané, 1968). Social coordination and communication depend on this nuanced understanding of agency. Investigating agency perception in multi-agent contexts, therefore, offers valuable insights into social behaviour and decision-making, particularly in collaborative or competitive situations where control is shared or distributed.

Depression and Agency

Depression is characterised by a range of symptoms, many of which relate to a diminished sense of agency. According to the DSM-V, individuals with depression often experience persistent sadness, fatigue, feelings of worthlessness, and helplessness, along with reduced motivation to initiate or sustain goal-directed behaviour (American Psychiatric Association, 2013). These symptoms reflect a broader sense of disempowerment, where individuals feel that they have little control over their environment. This reduced sense of control can lead to passivity and avoidance, which in turn exacerbate depressive symptoms (Abramson et al., 1978; Beck, 1967; Seligman, 1974).

Several key propositions have been suggested to explain why individuals with depression experience a reduced sense of agency. One widely recognised mechanism is learned helplessness, where individuals come to believe that their actions have little or no effect on outcomes. It is proposed that this lack of control is initially learned in specific contexts but then generalises to other areas of life, leading to resignation and a further decline in motivation (Maier & Seligman, 1976; Maier & Seligman, 2016). This may manifest as a cognitive bias, where depressed individuals believe they have less control than they actually do, even when they can influence outcomes (Ackermann & DeRubeis, 1991; Alloy & Abramson, 1979).

Additionally, they may attribute more control to external factors or other agents, which could reduce their perceived sense of personal control (Aiken & Baucom, 1982; Benassi et al., 1988). Another possibility is that lower engagement with the environment, which is common in depression, limits opportunities to experience consistent connections between their actions and outcomes, potentially reinforcing a sense of diminished control (Blanco et al., 2009; Blanco et al., 2012). These propositions are not necessarily mutually exclusive and may interact in complex ways to shape the reduced sense of agency observed in depression.

Exploring action-outcome contingency learning in multi-agent contexts becomes even more pertinent when considering individuals with depression. Do the cognitive biases associated with depression, such as the tendency to underestimate one's control, persist or perhaps intensify in multi-agent scenarios? In situations where only the self has control, do depressed individuals misattribute control to others, perceiving themselves as having reduced agency? Investigating these questions could shed light on the specific mechanisms underlying the reduced sense of agency in depression and how it manifests in social environments. Insights from this investigation could help in developing more targeted and effective interventions to enhance agency perception in both individual and social contexts.

To explore these questions, this thesis examined how individuals across the depression spectrum learn and perceive agency, particularly in multi-agent scenarios. The overview of the thesis, including its structure and the main findings from each chapter, is provided next.

Thesis Overview

This thesis explored how individuals learned and perceived agency in multi-agent situations, and how this process varied with depressive symptoms. It examined how people learned contingencies between their own actions and outcomes, as well as between others' actions and outcomes, to understand perceived control over the environment.

The first part of the thesis (**Chapters 2-4**) investigated the perceptual processes involved in contingency learning, focusing on both cue-outcome and action-outcome relations. Research has indicated that similar cognitive mechanisms govern the learning of contingencies related to both cues and actions (Béghin & Markovits, 2022; Jenkins & Ward, 1965), and therefore understandings from research on cue-outcome contingency learning could provide insight into action-outcome contingency learning.

In **Chapter 2**, key experimental studies and their findings were reviewed, highlighting trends in contingency learning. It was clear that the co-presence of actions/cues with outcomes indicated positive contingencies, while the presence of one and absence of the other suggested negative contingencies (Allan, 1980; Kao & Wasserman, 1993; Murphy et al., 2021). However, the role of the co-absence of actions/cues and outcomes remained ambiguous, as it has been shown to imply positive or negative contingencies or provide no information about contingencies at all, depending on the specific task goal (Arkes & Harkness, 1983; Mata et al., 2015).

Additionally, when a stronger predictor (cue) of an outcome was present, the perceived contingency of a weaker predictor was weakened when both predictors shared the same direction of control (i.e., both positive or both negative) but enhanced when in opposing directions (Baker et al., 1993; Darredeau et al., 2009). However, it remained unclear how contingencies were perceived when the weaker cue had no control at all. Moreover, it was unclear how the presence of another action-/cue-outcome relation (rather than another action/cue and shared target outcome) affected the contingency learning of the target action-/cue-outcome relation. And whether and how the presence of another action-/cue-outcome relation during the co-absence of the target action-/cue and target outcome would influence contingency learning of the target relation. These gaps highlighted the need for further

exploration of how individuals conducted contingency learning, especially in the context of multiple cues or actions.

Chapter 3 examined proposed mechanisms of contingency learning, evaluating different theories in light of the behavioural trends observed. Hypotheses were drawn from these theories to explore the research gaps related to co-absence of agents/cues and outcomes and the influence of multiple agents/cues on contingency learning.

Chapter 4 reported on Experiments 1-3, which explored how individuals across the depression spectrum processed the co-absence of target cues and outcomes and how the presence of non-target cues and outcomes influenced contingency learning of the target relationship. Results showed that the co-absence reduced perceived contingency, and that the presence of a non-target cue and outcome enhanced this effect by leading to a more negative perceived contingency between the target cue and outcome. This effect was dampened with higher depressive symptoms, and, in general, participants with more severe depressive symptoms did not perceive negative contingencies as strongly.

The second part of the thesis (**Chapters 5-7**) shifted focus to assess action-outcome relations, exploring both the perceptual and active processes involved in contingency learning. **Chapter 5** evaluated models of contingency learning in terms of how they accounted for the role of actions. Some models suggest that actions are simply guided by perceived contingencies, where individuals adjust their behaviour based on expected outcomes. Other models, such as the active inference model, propose that actions are used to both gather information about contingencies and achieve outcomes. The active inference model was introduced as a framework that could explain how agency was learned in multi-agent environments.

In **Chapter 6**, Experiments 4-6 were presented, examining how individuals across the depression spectrum learned action-outcome contingencies in multi-agent contexts, including

those with clinical depression. The findings, drawn from behavioural data, and corroborated by fitting an active inference model, suggested that the presence of another agent influenced perceptions of agency, as with non-target cues and outcomes in Experiments 1-3. Even when one agent lacked control, they were perceived as having control in the opposite direction to the agent in control. Non-clinically depressed individuals exhibited a self-serving bias, perceiving themselves as having positive agency while attributing negative agency to the other agent, and vice versa. Clinically depressed participants, however, exhibited a bias for both themselves and others as having positive agency, indicating difficulty processing negative contingencies. They also perceived less agency overall for both themselves and others, despite engaging in more actions. Fitting an active inference model to the behavioural data revealed that non-clinically depressed participants acted less but with more purpose, whereas clinically depressed participants performed more actions, though more aimlessly, potentially learning less from the data generated by their actions. These findings revealed distinct patterns in how clinically and non-clinically depressed individuals perceived and acted in multi-agent contexts to learn how much agency they had.

Finally, **Chapter 7** summarised the key findings of this thesis, presented the main conclusions, and outlined limitations and future research directions. This thesis revealed several key findings about how individuals with varying depressive symptom severity learned and perceived agency in multi-agent contexts. First, it showed that the co-absence of cues and outcomes reduced perceived contingency in this specific task, and that this effect is exacerbated when non-target stimuli are placed during such events. Second, it showed that while individuals with more severe depressive symptoms perceived contingencies as being weaker, they more specifically struggled to process negative contingencies. This was reflected in all perceived contingencies being skewed towards the positive side, even though their overall perception of contingency strength was reduced, reflecting a bias of positive control to both themselves and

others. In contrast, non-clinically depressed participants tended to exhibit a self-serving bias, perceiving themselves as having positive agency while attributing negative agency to others, and vice versa. Third, clinically depressed participants performed more actions, but in a more aimless manner, learning less from the outcomes of their actions, while non-clinically depressed participants acted less but with more purpose, learning more effectively. These findings contributed to both theoretical understanding and potential interventions aimed at improving agency perception in individuals with depression.

Chapter 2:

Trends in Perceiving Contingencies

In this chapter, I review key experimental studies on contingency learning, with a focus on how individuals learn agency through action-outcome contingencies. Insights from cue-outcome contingency learning are also valuable for understanding how these mechanisms function in action-based contexts as it has been shown that both are underpinned by similar cognitive processes (Béghin & Markovits, 2022; Jenkins & Ward, 1965). As a result, both cue-outcome and action-outcome contingency learning are discussed. The chapter begins by outlining a typical experimental task used to examine contingency learning. This is followed by a discussion of trends in perceived contingencies related to single action-/cue-outcome relations and the role of different combinations of the presence and absence of actions/cues and outcomes on perception. Then, I explore trends in perceived contingencies in tasks involving multiple actions/cues and outcomes, before concluding with an examination of how these perceptions vary across individuals with differing levels of depressive symptoms.

Experimental Methods

Contingencies are central for understanding how actions/cues influence outcomes because they capture the relationship between an action/cue and an outcome in, both, the presence and absence of that action/cue. While the probability of an outcome occurring after an action/cue provides some information, it does not fully reveal whether the action/cue itself is responsible for the outcome. For instance, if a light bulb sometimes turns on when a button is pressed, but also turns on when the button is not pressed, the probability of the outcome may seem high, yet the sense of control would be weak. In contrast, a strong contingency exists when the light consistently turns on only after pressing the button and remains off when the button is not pressed. This dual dependence on both action and inaction is crucial for accurately perceiving control.

To investigate how humans learn action-/cue-outcome contingencies, experimental psychologists have studied how individuals process combination of the presence and absence

of actions/cues and outcomes in controlled environments. One strategy to evaluate how individuals learn action-/cue-outcome contingencies through experiments involves manipulating the objective contingency, which refers to the actual statistical relationship between an action/cue and its outcome, independent of the participant's perception.

Table 1.
2x2 Matrix Illustrating Four Types of Information: A, B, C and D.

	Outcome Present	Outcome Absent
Action/Cue Present	**A**	**B**
Action/Cue Absent	**C**	**D**

Notes. Cell **A** represents instances wherein the action/cue and outcome are co-present. Cell **B** represents instances wherein the action/cue is present and the outcome is absent. Cell **C** represents instances wherein the action/cue is absent and the outcome is present. Cell **D** represents instances wherein the action/cue and outcome are co-absent.

A widely used index and formalisation of objective contingency in such experiments is ΔP (delta-P; Allan, 1980). ΔP is defined by a 2x2 matrix (see Table 1) that records the frequencies of four types of information forming the one-way contingency of the binary states (presence vs absence) of two stimuli (an action/cue and an outcome). When referring to any of the four types of information, also known as "events", or as "trials" in an experimental context, I will use bold font. The concurrent presence and absence of two stimuli (**A** and **D**, respectively, in Table 1) serves as confirmatory evidence of a contingency, i.e., it informs a positive contingency. Conversely, the presence of either stimulus with the absence of the other (**B** and **C** in Table 1) is disconfirmatory evidence of a contingency between the two stimuli, i.e., it informs a negative contingency. ΔP is calculated by subtracting the likelihood of the outcome (O) occurring without the action/cue (A) from the likelihood of it occurring with the action/cue present; $\Delta P = P(O|A) - P(O|\sim A) = (\mathbf{A/A+B}) - (\mathbf{C/C+D})$ (Allan, 1980; Shaklee & Tucker, 1980); the index is called ΔP because it represents the difference between these two conditional

probabilities. The value of ΔP varies from -1 to +1, like correlations, quantifying the strength and direction of the relationship.

In the experimental tasks designed to evaluate how individuals perceive cue-outcome contingencies, participants observe a series of trials, with each trial displaying one of the four types of information (**A**, **B**, **C**, or **D** events in Table 1) about the relationship between the two stimuli (De Houwer & Beckers, 2002b; Pineño & Miller, 2007; Shanks, 2007). The cues and outcomes can be either abstract, i.e., not inherently related stimuli, such as random shapes or objects, or they can be concrete stimuli which can be meaningfully related, such as a specific fertiliser and a crop field (Arkes & Harkness, 1983; Baetu & Baker, 2009; Baker et al., 1993; Murphy et al., 2021). The experimenter manipulates the objective contingencies of stimuli (ΔP) by varying the proportion of **A**, **B**, **C**, and **D** trials within the series.

At the end of the trial-series, participants rate the relationship between the cue and outcome they observed in the series on a visual analogue or numeric rating scale. The rating provided by the participants is compared with the objective contingency to assess the accuracy of the participants' contingency learning. Additionally, researchers compare the ratings across different conditions in the experiment to evaluate participants' sensitivity to the variables manipulated by the experimenter, gaining insights into how contingency learning occurs. In some studies, the rating is provided on a unidirectional scale from 0 to 100. However, while nonzero judgements indicate some degree of control, they do not indicate the direction (positive or negative) of the control. Hence, it is preferred that studies use a bidirectional scale, from -100 to 100, allowing expression of both positive and negative judgements (Chatlosh et al., 1985; Msetfi et al., 2007).

A similar task is used to assess action-outcome contingency learning in humans (Castiello et al., 2022; Chatlosh et al., 1985; Jenkins & Ward, 1965). The task involves an action that the participant performs, such as pressing a button or not, and according to the

objective contingency set by the experimenter, an outcome is present or absent, such as a light being on or off. The experimenter sets the likelihood of the outcome occurring given the action and given no action, with the difference between these probabilities providing ΔP.

There are two types of procedures: discrete trial and free operant. In discrete trial procedures, each trial comprises of an action opportunity that is clearly defined by the experimenter, allowing the participant to act only during specific periods, followed by an outcome period. Each trial may be separated from the next with an inter-trial interval, a period where the participant waits for the next trial. In contrast, free operant procedures allow participants to act at any time and however many times during a continuous session, with the outcome presented immediately following the action that produces it, providing a more naturalistic assessment of behaviour patterns. Subsequently, at the end of either type of task, participants report the degree of control they perceive they have. Participants' ratings are measured against the objective contingency to evaluate the accuracy of their contingency learning. Researchers also examine these ratings across various experimental conditions to assess how sensitive participants are to the manipulated variables, providing valuable insights into the process of contingency learning.

Single Action-Outcome Contingency Learning

It has been well-demonstrated that humans are sensitive to single cue-outcome (Arkes & Harkness, 1983; Baetu & Baker, 2009; Baker et al., 1993; Castiello et al., 2022; Murphy et al., 2021; Schustack & Sternberg, 1981) and action-outcome contingencies (Allan & Jenkins, 1980, 1983; Alloy & Abramson, 1979; Chatlosh et al., 1985; Dickinson et al., 1984; Shanks, 1985a; Wasserman, Chatlosh, et al., 1983; Wasserman et al., 1993).

This sensitivity is illustrated by an elegant study conducted by Chatlosh et al. (1985). In this study, participants judged the extent to which pressing a telegraph key (like a button) caused a white light to flash, across all possible pairings of probabilities for the outcome

occurring with and without the action, P(O|A) and P(O|~A), with values of .125, .500, and .875. If a participant pressed the telegraph key during an interval, the light flashed with probability P(O|A); if not, the light flashed with probability P(O|~A). The contingency was positive when P(O|A) was greater than P(O|~A), negative when P(O|A) was less than P(O|~A), and zero when the two were equal. The length of each trial and block of trials differed between participants. At the end of each block of trials, participants rated the response-outcome relation on a scale from -100 to +100. The results indicated a high degree of sensitivity to the actual contingency, with 98.11% of the variance in judgments accounted for by the linear trend of the objective contingency as determined by ΔP. This study demonstrates that individuals are attuned to both positive and negative action-outcome contingencies across different levels, suggesting that contingency learning is fundamental to how people perceive their control over outcomes, or their sense of agency. Moreover, this sensitivity implies that all four types of events (**A**, **B**, **C**, and **D**) are integrated into perceived contingency, mirroring how these events shape ΔP.

Learning Curves

While ratings provided at the end of tasks (as in the studies mentioned thus far) are useful insights, they miss the ongoing process of learning across the trials. To capture this, researchers examined how judgements about contingencies evolved over a series of trials, revealing a learning curve. The concept of a learning curve was first introduced by Hermann Ebbinghaus in the late 19th century and later formalised by Clark Hull in the early 20th century. Hull (1943) proposed that learning involves a gradual process of acquiring knowledge or behaviour over time, and this process could be visualised as a curve reflecting improvements in performance or understanding.

Indeed, studies of contingency learning have consistently demonstrated learning curves. In these studies, participants rated the contingency not only at the end of a trial block but also at various points during the task, allowing researchers to observe changes in perception over

time (Baker et al., 1989; Shanks, 1985a; Wasserman, Chatlosh, et al., 1983). Findings showed that participants gradually adjusted their judgements across trials, with initial ratings starting near zero. When the contingency was positive, judgements started low and increased gradually, stabilising near the objective contingency value. Conversely, with a negative contingency, judgements gradually decreased toward the objective value. This incremental and decremental pattern of judgements across trials clearly illustrated the learning curve.

The gradual adjustments in contingency judgements across trials, captured by learning curves, reflect the dynamic nature of action-outcome learning. These patterns of learning over time are relevant to understanding how individuals develop a sense of agency, as they continuously update their perceptions of control through repeated interactions with their environment.

Inter-Trial Intervals

Agency learning, the process by which individuals understand their ability to influence outcomes through actions, is rooted in the learning and perception of action-outcome contingencies. This relationship is not merely about the probability of actions or outcomes occurring independently, but rather about the reliability with which actions lead to specific outcomes. Contingencies can be formalised as $\Delta P = P(O|A) – P(O|\sim A) = (\mathbf{A}/\mathbf{A}+\mathbf{B}) - (\mathbf{C}/\mathbf{C}+\mathbf{D})$ (Allan, 1980). Statistically, both the co-presence of action and outcome (**A**) and the absence of both (**D**) contribute to a positive contingency, while isolated presences of action or outcome (**B** and **C**) contribute to a negative contingency, and thus, agency.

According to ΔP, while co-presence events readily support the perception of control, co-absence events are crucial for confirming that outcomes follow actions rather than inaction. However, these co-absence events tend to be less salient than co-presence events, as they are characterised by the absence of both action and outcome – essentially, where nothing happens.

This raises the question of whether individuals perceive and utilise information from co-absence (**D**) events, and if so, how does it influence their learning process?

Msetfi and colleagues (2007; 2005) addressed this question by examining the role of inter-trial intervals – periods in which participants wait for the next trial to occur, where neither action nor outcome is present, and thus likened to **D** events – on perceived agency. They found that longer inter-trial intervals were associated with higher perceived agency across negative, zero, and positive contingencies. Their finding suggests that individuals may indeed register co-absence (**D**) events, contributing to the perception of positive contingencies, highlighting the potential relevance of co-absence events in contingency learning.

Additionally, Msetfi and colleagues (2007; 2005) demonstrated that incorporating inter-trial intervals into ΔP calculations as co-absence (**D**) events explains discrepancies in observed outcome density effects across studies. The outcome density effect refers to the tendency for individuals to overestimate their control when outcomes frequently occur independently of their actions (P(O)), relative to the objective contingency (ΔP). Previous research, where the tasks included inter-trial intervals, demonstrated this effect in both passive (cue-related) and active (action-related) scenarios (Alloy & Abramson, 1979; Benassi & Mahler, 1985; Dickinson et al., 1984; Jenkins & Ward, 1965; Msetfi et al., 2005; Musca et al., 2010; Vazquez, 1987). However, some studies, where the tasks did not include inter-trial intervals, did not replicate the outcome density effect (Allan & Jenkins, 1980, 1983; Alloy et al., 1985; Wasserman, Chatlosh, et al., 1983). The inclusion of inter-trial intervals in contingency calculations as **D** events reduces the probability of an outcome occurring without an action (P(O|~A)), thereby increasing the overall contingency (ΔP). This adjustment accounts for the variability in outcome density effects observed across studies, depending on whether inter-trial intervals were included in the experimental design.

Figure 1 illustrates this effect by comparing contingency (ΔP) calculations with and without inter-trial intervals across negative, zero, and positive contingency conditions. The left panels show traditional calculations, while the right panels demonstrate the impact of adding 40 inter-trial interval events as **D** events in the contingency calculations under differing levels of outcome density. Including inter-trial intervals as **D** events decreases the probability of an outcome in the absence of action, resulting in a more positive contingency (ΔP). This effect is more pronounced in scenarios with higher outcome densities.

Other researchers have also found that inter-trial intervals increased the perceived positive contingency (Miller & Escobar, 2002; Urcelay et al., 2009). Interestingly, all these studies found that the influence of inter-trial intervals on perceived contingencies varied depending on the direction of the contingency, with a larger effect in negative contingencies and weaker or non-existent in positive contingencies. This aligns with previous research showing that the outcome density effect is more prominent in negative contingencies and less so in positive contingencies (Lennox et al., 1990; Vazquez, 1987). The calculations presented in Figure 1 reflect this differential impact of adding **D** events to the ΔP calculations across various contingency conditions.

In conclusion, it is suggested that the co-absence of actions and outcomes (**D** events), as represented by inter-trial intervals, contributes to human contingency learning. These events seem to enhance the perception of positive agency (action-outcome contingencies). This finding has important implications for agency learning, as it shows that even when no action is taken, the absence of both action and outcome may reinforce the understanding of control over outcomes.

Figure 1.
How Inter-Trial Intervals Can Increase Perceived Contingency.

Positive contingency / no ITI			Positive contingency / long ITI			Increase in Δp
20	0	P(O|R) = 1.0	20	0	P(O|R) = 1.0	0.33
10	10	P(O|NoR) = .5	10	50	P(O|NoR) = .17	
f(O) = 30		Δp = .5	*f*(O) = 30		Δp = .83	
15	5	P(O|R) = .75	15	5	P(O|R) = .75	.15
5	15	P(O|NoR) = .25	5	55	P(O|NoR) = .10	
f(O) = 20		Δp = .5	*f*(O) = 20		Δp = .65	
10	10	P(O|R) = .50	10	10	P(O|R) = .50	0.0
0	20	P(O|NoR) = 0	0	60	P(O|NoR) = 0	
f(O) = 10		Δp = .5	*f*(O) = 10		Δp = .5	
Zero contingency / no ITI			**Zero contingency / long ITI**			
15	5	P(O|R) = .75	15	5	P(O|R) = .75	0.50
15	5	P(O|NoR) = .75	15	45	P(O|NoR) = .25	
f(O) = 30		Δp = 0	*f*(O) = 30		Δp = .5	
10	10	P(O|R) = .5	10	10	P(O|R) = .50	0.33
10	10	P(O|NoR) = .5	10	50	P(O|NoR) = .17	
f(O) = 20		Δp = 0	*f*(O) = 20		Δp = .33	
5	15	P(O|R) = .25	5	15	P(O|R) = .25	0.15
5	15	P(O|NoR) = .25	5	55	P(O|NoR) = .1	
f(O) = 10		Δp = 0	*f*(O) = 10		Δp = .15	
Negative contingency / no ITI			**Negative contingency / long ITI**			
10	10	P(O|R) = .5	10	10	P(O|R) = .5	0.67
20	0	P(O|NoR) = 1.0	20	40	P(O|NoR) = .33	
f(O) = 30		Δp = –.5	*f*(O) = 30		Δp = .17	
5	15	P(O|R) = .25	5	15	P(O|R) = .25	0.50
15	5	P(O|NoR) = .75	15	45	P(O|NoR) = .25	
f(O) = 20		Δp = –.5	*f*(O) = 20		Δp = 0	
0	20	P(O|R) = 0	0	20	P(O|R) = 0	0.33
10	10	P(O|NoR) = .5	10	50	P(O|NoR) = .17	
f(O) = 10		Δp = –.5	*f*(O) = 10		Δp = –.17	

Notes. Contingency tables showing how changes in ΔP, when inter-trial intervals are integrated into the contingency calculation as cell **D** events, depend on the contingency and the frequency of outcome occurrence (*f*(O) in the figure or P(O) in the main text), across varying levels of outcome density. The left panels show traditional ΔP calculations, while the right panels include 40 additional cell D events in each condition. The increase in ΔP is particularly notable in the higher outcome density scenarios. This figure is Figure 5 from "Depressive realism and the effect of intertrial interval on judgements of zero, positive, and negative contingencies." by R.M. Msetfi, R.A. Murphy, & J. Simpson. (2007). *Quarterly Journal of Experimental Psychology*, *60*(3), 461-481. https://doi.org/10.1080/17470210601002595. Reprinted with permission.

The Role of D Events

The previous section discussed Msetfi and colleagues' (2007; 2005) findings, which demonstrated that inter-trial intervals can increase the perceived contingency between actions and outcomes, thereby enhancing perceived agency. They suggested that inter-trial intervals may function similarly to **D** events, as both involve periods without actions or outcomes. This shows that **D** events in contingency learning may serve to increase the perceived contingency between actions and outcomes. Other researchers have also supported this idea, suggesting that both inter-trial intervals and **D** events contribute to the perception of agency by reinforcing the dependence between actions and their consequences (Baker, Murphy, et al., 2000; Castiello et al., 2022).

While Msetfi and colleagues' (2007; 2005) findings suggest that inter-trial intervals may be processed similarly to **D** events, the two types of events are conceptually and structurally distinct, warranting further investigation into **D** events specifically. Inter-trial intervals, by their nature, occur regularly after each **A**, **B**, or **C** trial, providing participants with predictable breaks during which neither actions nor outcomes are presented. This regularity means that inter-trial intervals are consistent, time-bound gaps following each trial, potentially allowing participants to consolidate information from previous trials (Aust et al., 2019; McGaugh, 1966). In contrast, **D** events would be integrated and occurring irregularly as part of the structure of **A**, **B**, **C**, and **D** trial sequence itself.

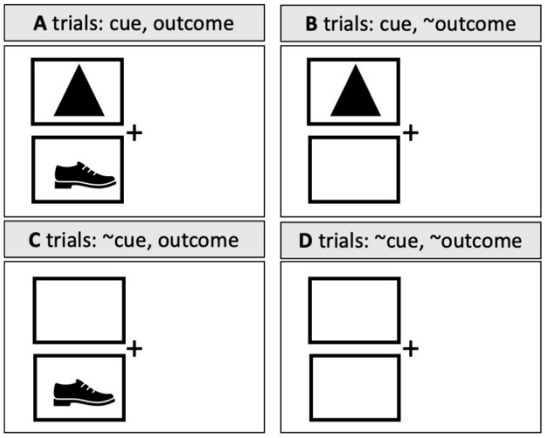
Moreover, Msetfi and colleagues' studies manipulated the duration of the inter-trial intervals while keeping the duration of **A**, **B**, and **C** trials constant, meaning that these trials could vary in length from the inter-trial intervals that are likened to **D** trials. This difference in duration between fixed inter-trial intervals and the irregular nature of **D** events embedded within trials raises the possibility that **D** events, when incorporated more naturally into trial structures, might influence contingency learning differently. Investigating **D** events in isolation

would provide clearer insights into their specific role in learning action-outcome contingencies and their influence on perceived agency, independent of the regularly spaced breaks afforded by inter-trial intervals. Therefore, in two of our previous studies (Castiello et al., 2022; Murphy et al., 2021), we sought to examine the role of **D** events in contingency learning.

Dealing with **D** events present distinct challenges. Defining **D** events within an experiment is difficult as these events are characterised by the lack of stimuli. Furthermore, ensuring that participants perceive these periods as informative as they provide confirmatory information about the contingency, rather than merely as intervals of nothingness, is equally challenging. One way to clearly present **D** events in an experiment is to incorporate local contextual cues within the trial series (composed of **A**, **B**, **C**, and **D** events) presented to participants. Therefore, in our two previous studies (Castiello et al., 2022; Murphy et al., 2021), we employed trial markers in the form of solid-line rectangles to indicate the spatial location of the cue (top trial marker) and the outcome (bottom trial marker); see Figure 2. These trial markers provided a local context that framed the stimuli within each trial, potentially making **D** events more distinguishable due to the presentation of empty trial markers.

Inspired by the work of Msetfi and colleagues (2007; 2005), who showed that longer inter-trial intervals increased perceived contingency, we manipulated both the frequency and duration of **D** trials while keeping the other types of trials constant. Msetfi et al.'s experiments only examined the impact of two inter-trial interval durations, with either 0.5 or 3 seconds for shorter intervals and 15 seconds for longer ones. Hence, it was unclear whether participants parsed longer inter-trial intervals as multiple event instances, thus increasing event frequency, or perceived them as a single extended event, which provided more time for processing. Both mechanisms have been suggested to enhance learning (Bugelski, 1962; Delorme et al., 2018; Melcher, 2001).

Figure 2.
Illustration of A, B, C, and D Trials in a 2x2 Contingency Matrix Framework.



Notes. The four trial types correspond to the 2x2 contingency matrix, illustrating the **A**, **B**, **C** and **D** events as formalised by Allan (1980) and depicted in Table 1. ***2x2 Matrix Illustrating Four Types of Information: A, B, C and D.*** The solid black rectangles serve as trial markers, providing local context and indicating the spatial locations for potential cue and outcome presentations. These markers were designed to enhance the distinguishability of **D** events, particularly in trials where neither a cue nor an outcome was presented. The trial markers and stimuli (cue and outcome) are positioned to the left of the fixation cross in this example, but they alternated between the left and right sides of the fixation cross across trials (i.e., right-left-right-left…), ensuring that each trial presentation was distinct. Note that this is an illustrative example and not the original stimuli or trial structure used in the studies.

To explore this, we carefully manipulated the frequency and duration of **D** trials while controlling for the potential confounding effects of these variables on one another. Our findings indicated that while frequency influenced the perceived associative link between the target stimuli, duration did not have a significant effect. The results regarding **D** events, however, were mixed: Murphy et al. (2021) found no significant effect of increasing the frequency of **D** trials, while (Castiello et al., 2022) found an increase in positive cue-outcome contingency perception.

The disparity in results may be due to differences in complexity of the task design. Murphy et al. (2021) manipulated the frequency of all four trial types independently within the same experiment. This means that between different series of trials, participants were exposed to conditions where the frequency and/or duration of **A** trials was increased while **B**, **C**, and **D** remained constant, and **B** was increased while **A**, **C**, and **D** remained constant, and so on. This

complex design may have masked the effect of **D** trials, potentially leading to an overall weaker impact of **D** trial manipulation on learning the contingency. Castiello et al. (2022), on the other hand, solely manipulated **D** trials while keeping **A**, **B**, and **C** trials consistent across all conditions in an experiment. This simpler, more focused manipulation may have heightened participants' sensitivity to the **D** trials, making any change in its frequency more salient and impactful on the perceived contingency between the target stimuli.

Nonetheless, neither set of results is surprising. Castiello et al. (2022)'s findings are in line with previous studies that examined the effect of inter-trial intervals (likened to **D** events) on perceived contingency, where it was found that inter-trial intervals increased the perceived positive contingency, with a larger effect on negative contingencies than positive (Miller & Escobar, 2002; Msetfi et al., 2007; Urcelay et al., 2009). Murphy et al. (2021)'s findings are also not surprising as researchers have found that under certain conditions, individuals neglect **D** events. This tendency to neglect **D** events, along with its implications for contingency learning, is explored in the paragraphs that follow.

Researchers have suggested that individuals may adopt different judgement strategies and have prior expectations depending on the scenario they are facing, influencing contingency learning (Alloy & Tabachnik, 1984; Shaklee, 1983; Tversky & Kahneman, 1981). Mata et al. (2015) showed that participants' goals and judgement strategy determined whether **D** events are neglected or considered relevant and thereby processed. When **D** events were processed, they informed the positive contingency between stimuli, leading to ratings that aligned more closely with the objective contingency (ΔP). In contrast, when **D** events were neglected, participants provided less accurate ratings. Additionally, the study showed that once participants were freed from a goal that necessitated them to process **D** trials in one task, they tended to (at least in the short-term) process **D** trials in subsequent goal-free tasks with different content, leading to more accurate contingency judgements.

Moreover, while all events logically contribute equally to the overall contingency, researchers have found that individuals tend to differentially weight the four types of information that inform a contingency. They found that the co-occurrence of the cue and outcome (**A** events) is weighted more than **B** and **C** events, and that **D** events are weighted the least (Kao & Wasserman, 1993; Perales & Shanks, 2007; Wasserman et al., 1990; Wasserman et al., 1996). Despite being weighted the least, **D** events still played a role in informing a positive contingency when processed. Researchers have also suggested this unequal weighting of events is amplified under high cognitive demand (Arkes & Harkness, 1983), which may have been the case in Murphy et al. (2021) due to the complex task design, resulting in neglecting of **D** events.

Finally, other attempts to decipher the role of each of the four events could not conclusively determine the role of **D** events. Researchers had participants rate the contingency between the target stimuli after each trial instead of at the end of the series of trials. Schustack and Sternberg (1981) found that some participants gave slightly more positive ratings following **D** events while some gave slightly more negative or unchanged ratings. Similarly, Arkes and Harkness (1983) discovered "more variation in subjects' treatment of Cell **D** data than of the other cells' data. Some subjects considered Cell **D** data to strengthen judged contingency, some considered such data to weaken it, and others simply were undecided." (page 131). However, the rating scales used in both these studies was unidirectional and therefore did not distinguish between positive and negative control properly.

It is clear that people can be sensitive to **D** events depending on the task goal, where these events inform the perceived contingency between the target action/cue and outcome, while under other conditions, **D** events may be neglected. Therefore, the presence of another (non-target) agent, cue, or outcome may influence the processing of **D** events, potentially enhancing their relevance by helping individuals differentiate between the effects of self- and

other-produced actions or multiple cues. In multi-agent contexts, distinguishing between one's own influence on outcomes and that of others becomes crucial for learning accurate contingencies, and **D** events may play an important role in refining one's perception of self-agency. The trends in perceiving contingencies across multiple agents, cues, and outcomes are explored in the next section.

Single and Ready to Mingle: The Role of Others

Understanding how individuals discern control in environments with multiple agents and cues is essential for grasping how people perceive their own agency. In real-world scenarios, where various agents or cues may influence an outcome, identifying who or what has control is key to establishing a clear sense of agency. This process is particularly important in social contexts, where effective coordination and goal-directed behaviour rely on accurately perceiving both self and others' influence on outcomes. The presence of other agents or cues may shape how one perceives their own control, and vice versa. This highlights the importance of examining how multiple predictors impact the learning and perception of agency in complex environments.

Social coordination and communication require a nuanced understanding of agency - not just in terms of one's control over the environment but also in recognising the influence of other agents (Declerck et al., 2006). Individuals must distinguish their influence on outcomes from that of others, requiring them to learn contingencies between self- and other-produced actions and outcomes. While this differentiation is essential for refining one's perception of self-agency, this refinement does not necessarily lead to a more accurate perception of one's control (Wolpert et al., 2003). For example, in social contexts, the presence of others can lead to a diffusion of responsibility, reducing the likelihood that someone will intervene in an emergency, also known as the bystander effect (Chekroun & Brauer, 2002; Darley & Latané,

1968) or decreasing effort in collaborative tasks resulting in social loafing (Karau & Williams, 1993).

This thesis explores how agency is learned and perceived in multi-agent contexts, specifically investigating how the presence of other agents affects the perception of self-agency, as well as how the presence of the self influences the perception of others' agency.

In the previous section, I focused on trends observed from experiments involving a single predictor (action/cue). However, researchers have argued that even in these seemingly straightforward contexts, individuals must distinguish their actions (or cue) from other potential influences, such as contextual cues (Shanks, 1993). For example, in computer-based tasks where participants assess how their actions, like pressing a button, influence outcomes, they must distinguish their actions from other potential influences. When an outcome occurs without their action, participants might attribute the cause to some aspect of the context, such as the computer itself. In this scenario, participants evaluate their control (contingency between self-produced action and outcome) over the outcome relative to the context's influence. Some researchers argue that this competition between the target predictor's and contextual factors' association with the outcome is what results in the reduction of contingency judgements when outcomes occur in the absence of the target predictor (Pineño & Miller, 2007; Shanks & Dickinson, 1988; Urcelay & Miller, 2014).

A more complex scenario arises when another agent is introduced, performing actions alongside the participant. In these multi-agent contexts, participants must learn to navigate the relationships between their own actions, the actions of others, contextual cues, and outcomes to determine their level of control. Whether the additional influence is a contextual stimulus, such as a computer, or another agent, like a fellow participant, individuals are tasked with attributing outcomes to specific causes.

Blocking

When faced with multiple potential predictors, how individuals combine information concerning their own actions, the actions of others, contextual cues, and outcomes to form perceptions of control becomes a central question in contingency learning. Researchers have explored how individuals process information when more than one predictor is present, and one well-known phenomenon that sheds light on this is "blocking". Blocking occurs when individuals, having already learned the contingency between a specific predictor (action/cue) and an outcome, perceive a new predictor as less likely to be associated with the outcome. This illustrates how prior learning shapes the perception of new predictors, even when both could independently influence the outcome.

In experiments demonstrating blocking, participants judged the extent to which firing shells at tanks on a screen using a spacebar (target action) caused the tanks to blow up (outcome) (Dickinson et al., 1984; Kamin, 1968; Shanks, 1985b). An alternative background cause (contextual cue) was present as a minefield that the tanks traversed across during each trial. The objective contingency between the action and outcome was zero, with P(O|A) and P(O|~A) both at 0.75. In the blocking condition, participants observed the outcome occurring in the presence of the context (minefield) alone for a period, with P(O|~A) set at 0.75, before they could act and learn the target action-outcome relationship. In the control condition, participants could act and observe its consequences throughout. Despite identical statistical relationships between action and outcome in both conditions, participants in the blocking condition judged their level of control to be significantly lower. This is an example of "forward blocking".

Researchers have also demonstrated "backward blocking", where similar effects occur using a slightly different task design (Chapman, 1991; De Houwer & Beckers, 2002a; Dickinson, 1996; Miller & Matute, 1996). Backward blocking occurs when an outcome is

initially associated with two potential causes (for example, firing shells at a virtual tank and mines), but later, one of the causes is shown to be sufficient on its own. This "retrospective revaluation" leads individuals to downplay the causal role of the other potential cause. This phenomenon underscores that contingency judgements are dynamic and can be influenced by new information about the relative efficacy of potential causes. Thus, backward blocking reinforces the notion that the perceived predictiveness of a predictor is continually adjusted based on evolving understandings of other causal factors present in the environment.

These results highlight the interplay between potential predictors in contingency learning. When individuals learn that one predictor reliably causes an outcome, they tend to downplay the influence of others. This blocking effect demonstrates that judgements about agency are shaped not only by direct action-outcome relationships but also by the *relative* predictiveness of different factors. In multi-agent contexts, this suggests that people may perceive their own control as diminished when another agent's actions are seen as the primary cause of an outcome. Blocking, therefore, underscores how agency perception is influenced by the comparative strength of different predictors in the environment.

Relative Validity

While blocking studies involve a two-phased approach, where, for example, participants first learn about the alternative predictor before the target predictor is introduced, relative validity studies employ a one-phased design. In these experiments, multiple predictors are presented simultaneously.

Researchers have conducted relative validity experiments in negative and positive contingencies (Baker et al., 1993; Spellman, 1996a, 1996b; Tangen & Allan, 2004; Wagner et al., 1968). They found that the presence of a stronger predictive stimulus diminished the perceived contingency of a weaker predictor with the outcome. Additionally, Murphy and colleagues (Baetu et al., 2005; Murphy et al., 2001) demonstrated that these effects persisted

even when there were two stronger predictors and the two strong predictors did not individually predict all of the outcomes such that the predictors were presented separately in different trials (AX+ and BX+) rather than together in the same trial (ABX+).

Most studies on the relative validity effect have paired a weaker positive predictor with a stronger positive predictor, or, occasionally, a weaker negative predictor with a stronger negative predictor. Therefore, Baker and colleagues (Baker et al., 1993; Baker, Vallée-Tourangeau, et al., 2000; Darredeau et al., 2009; Vallée-Tourangeau et al., 1998) investigated competing predictors of opposite polarity. They found that participants' perceived contingency of a weaker predictor was instead enhanced in the presence of a stronger predictor. This means that in the presence of a stronger positive predictor, the perceived contingency of a weaker negative predictor became more negative. And, in the presence of a stronger negative predictor, the perceived contingency of a weaker positive predictor became more positive.

In situations with multiple predictors, the findings from forward blocking, backward blocking, and relative validity reveal that the presence and relative strength of alternative predictors shape how individuals perceive a target predictor's relationship with an outcome. The order in which these predictors are encountered does not hold much importance, as individuals can retrospectively re-evaluate the predictors based on their relative strengths. Contingency judgments are thus shaped not only by direct experiences but also by the comparative strength of the predictors involved. Directionality adds further complexity, as predictors with opposing directions can amplify the perceived impact of weaker predictors. However, it remains unclear how individuals process situations where the weaker predictor has no control, leaving an open research question about how control is attributed in such cases.

These findings offer important insights into how individuals with depression may misperceive their own control in situations with multiple causes. Depressed individuals may experience diminished agency when they perceive others as having more control, even when

objective contingencies suggest their control is the same or only slightly less. This highlights the importance of studying agency perception, particularly in environments where multiple potential causes are present, across individuals with varying levels of depressive symptoms to understand how they perceive and learn about control.

Summary

Understanding how much agency one has over their environment involves learning the contingency between actions and outcomes. The literature on cue-outcome and action-outcome contingency learning reveals several trends that explain how individuals come to perceive these relationships. In simpler scenarios, where a single action or cue is paired with an outcome, individuals generally develop accurate perceptions of contingencies over time (Chatlosh et al., 1985; Shanks & Dickinson, 1991). This contingency learning process is gradual, and individuals progressively refine their understanding of the control they exert over their environment as they associate actions with outcomes.

A key factor in this learning process is the type of events people process. These events can be classified into four categories: **A** (action and outcome present), **B** (action present, outcome absent), **C** (action absent, outcome present), and **D** (action and outcome absent). While **D** events (the co-absence of both action and outcome) are statistically significant in determining whether an action is necessary for an outcome, it remains unclear whether and how they are processed during human contingency learning. Research has shown that when **D** events are processed, they inform positive contingencies (agency), reinforcing the necessity of the action for the outcome to occur (Castiello et al., 2022; Msetfi et al., 2007). However, **D** events are often overlooked, especially when cognitive demands are high or depending on the goals of the task (Arkes & Harkness, 1983; Mata et al., 2015). Understanding when and how **D** events contribute to perceptions of control remains an important area of inquiry.

In more complex environments, where multiple potential predictors are present, individuals must navigate competing sources of information, like their own actions, the actions of others, contextual cues, and outcomes. Research shows that the relative strength of each predictor plays a key role in shaping perceived contingencies. When the contingencies of the predictors are in the same direction (e.g., both positive), the stronger predictor tends to reduce the perceived contingency of the weaker one (Baker et al., 1993; Baker, Vallée-Tourangeau, et al., 2000). However, when the contingencies are in opposing directions (e.g., one positive and one negative), the weaker predictor's contingency tend to be enhanced in the opposite direction to the stronger predictor (Baker et al., 1993; Darredeau et al., 2009).

These findings underscore the complexity of learning contingencies in situations with multiple potential predictors. Yet, much of the research focuses on cases where predictors share a common outcome, leaving some questions unanswered. For instance, does the presence of a separate (non-target, i.e., potentially unrelated) predictor-outcome relationship affect how individuals perceive the target predictor-outcome relationship in the same way as when there is a shared outcome between two competing predictors? Additionally, when a non-target predictor and outcome are present during the co-absence of the target predictor and outcome (**D** event), does this enhance or interfere with the processing of **D** events and how does it influence the perception of the target contingency?

These questions are explored in Experiments 1-3, discussed in Chapter 4, which investigate the role of **D** events in contingency learning and explore how non-target predictors and outcomes influence perceptions of the target predictor-outcome relationship. The remainder of this chapter will discuss trends in contingency learning associated with depressive symptoms, while Chapter 3 evaluates proposed mechanisms of contingency learning, focusing on the role of **D** events and non-target predictors.

Depression and Agency

Agency, the sense of control one has over their actions and their outcomes, is a fundamental element of human cognition, shaping our understanding of the causal structure of the environment and our ability to navigate within it. A strong sense of agency plays a critical role in how individuals perceive and interact with their surroundings. The clarity in knowing what they have control over, to what extent, and whether that control is positive (excitatory) or negative (inhibitory) enables them to navigate challenges more effectively.

This perception of control strengthens motivation, as individuals come to recognise that their efforts will have meaningful consequences. This belief in one's ability to control outcomes directly influences behaviour. Individuals with a strong sense of agency are more likely to engage in exploratory behaviour, driven by the expectation that their actions will yield valuable information or outcomes (Declerck et al., 2006; Kayser et al., 2015; Penton et al., 2018). Their confidence in their capacity to influence their environment and exploration resulting in trying out different strategies fosters flexibility, allowing them to adapt their strategies when faced with obstacles or changing circumstances. This adaptability enables more purposeful, goal-directed action, as individuals continuously refine their actions to align with desired outcomes. The sense of control they possess drives persistence and perseverance, facilitating the achievement of short- and long-term goals.

Moreover, a strong sense of agency is linked to psychological well-being. People who perceive control over their environment report higher levels of life satisfaction and resilience (DeNeve & Cooper, 1998). This sense of control also enhances stress management. Individuals with a strong sense of agency are better equipped to handle stress, as they believe they can influence their circumstances, and this is evident in the finding that such individuals tend to have lower cortisol responses in stressful situations (Bollini et al., 2004).

In contrast, a diminished sense of agency can lead to negative psychological and behavioural consequences. Individuals who perceive less control over their actions and outcomes are more likely to experience feelings of helplessness and frustration (Mehta et al., 2023; Yoshie & Haggard, 2013). A reduced sense of agency is closely associated with learned helplessness, a state in which individuals, after experiencing uncontrollable outcomes, generalise this lack of control across various situations, leading to passivity and withdrawal (Alloy & Abramson, 1982; Maier & Seligman, 1976; Maier & Seligman, 2016). Such experiences can also increase the risk of developing symptoms of apathy and anhedonia, which are key features of depression (Thompson et al., 2013).

Depression is often characterised by a marked reduction in the sense of agency (Beck, 1967; Seligman, 1974). Individuals experiencing depression frequently report feeling as though their actions have little impact on their environment, exacerbating feelings of helplessness and reinforcing a cycle of inaction. This diminished sense of control is supported by psychometric studies, where individuals with depression consistently scored lower on measures of perceived agency, as shown by studies using psychometric scales (Abdolmanafi et al., 2011; Fresco et al., 2006; Hovenkamp-Hermelink et al., 2019). Behavioural studies similarly show that individuals with depression perceived action-outcome contingencies as reduced (Castiello et al., 2020).

Understanding how individuals with depression learn action-outcome contingencies can provide valuable insights into the broader cognitive mechanisms of agency in the general population. By examining how this learning process differs in individuals with varying levels of depressive symptoms, we may better understand why those with depressive symptoms experience reduced agency and struggle to regain a sense of control. This knowledge could inform the development of interventions designed to restore agency by helping individuals re-

establish the relationship between their actions and outcomes, ultimately enhancing motivation and psychological well-being.

Depressive Realism

An intriguing phenomenon found by researchers examining action-outcome contingency learning in individuals with and without depression is known as depressive realism (Alloy & Abramson, 1979). Depressive realism suggests that individuals with depression perceive their level of control more accurately than non-depressed individuals, who harbour a self-enhancing, optimistic perception about their control and influence. This concept challenges the notion that good mental health is synonymous with accurate perceptions of reality.

Alloy and Abramson (1979) conducted the formative study that led to the formulation of the depressive realism hypothesis. Two groups of participants, depressed and non-depressed, were asked to judge their level of control in a task where they could press a button to turn on a light. In conditions where the task was programmed such that neither group had control (zero contingencies), depressed participants were more realistic in assessing their lack of control, while non-depressed individuals tended to overestimate their influence. These findings have been replicated by other researchers (Benassi & Mahler, 1985; Martin et al., 1984; Vazquez, 1987).

However, the depressive realism effect where depressed individuals more accurately perceive their level of control (or more like the lack thereof) is not universally observed and appears to be dependent on specific circumstances. It seems to emerge primarily in scenarios where participants have no actual control: no differences between depressed and non-depressed individuals have been reported in scenarios where the participant does have control over the outcome, specifically when their actions produce outcomes and inactions do not, i.e., positive agency (Alloy et al., 1985; Lennox et al., 1990; Msetfi et al., 2007; Vazquez, 1987).

Researchers posited that zero contingencies may fit with depressed individuals' a priori expectations that they do not have control over their environment, due to their lower self-esteem, resulting in more accurate judgements in zero contingencies (Ackermann & DeRubeis, 1991; Alloy & Abramson, 1979). It is then argued that non-depressed people expect that they have control, driven by the motivation to maintain self-esteem, resulting in higher judgements. These explanations for depressive realism concur with the finding of no difference in perceived control between depressed and non-depressed individuals when positive contingencies are tested.

However, this does not account for why the depressive realism effect only occurs in high outcome density conditions, where outcomes frequently occur regardless of the individual's actions (Allan et al., 2007; Alloy & Abramson, 1979; Msetfi et al., 2007; Msetfi et al., 2005). When outcome density is low, both groups tend to accurately judge their lack of control. This suggests that individuals with depressive symptoms may be less sensitive to outcome density variations.

As mentioned previously, research has indicated that outcome density effects only emerged when inter-trial intervals (likened to **D** events) were included in the task, with longer inter-trial intervals enhancing perceptions of contingencies in the positive direction under high outcome density (Msetfi et al., 2007). This effect is particularly pronounced in negative contingencies and is still present in zero contingencies, although the effects are minimal in positive contingencies; this is in line with the depressive realism effects not being found in positive contingencies. It is therefore not surprising that studies that demonstrated the depressive realism effect used procedures containing long inter-trial intervals (14-15 seconds) (Allan et al., 2007; Alloy & Abramson, 1979; Msetfi et al., 2007; Msetfi et al., 2005). In contrast, studies that did not replicate the depressive realism effect, nor find an effect of outcome density on perceived contingencies, used tasks without inter-trial intervals or with

very short inter-trial intervals (.5-5 seconds) (Dobson & Pusch, 1995; Kapçi & Cramer, 1999; Msetfi et al., 2007).

These findings suggest that while non-depressed individuals may integrate inter-trial intervals into their perception of contingency, increasing perceived control, depressed individuals may do so to a lesser extent. This implies that depressive realism may stem from an impairment in processing periods of time without the target cue or outcome (such as inter-trial intervals or **D** events) rather than simply reflecting accurate, non-optimistic expectations.

Depression is often linked to attentional impairments (Keller et al., 2019), which could hinder the processing of **D** events where stimuli are absent and therefore less attention-grabbing. This might be further compounded by attentional biases in depression, where individuals focus more on negative stimuli and overlook neutral or non-salient events (Gotlib & Joormann, 2010). Additionally, the reduced motivation often seen in depression could lead to less thorough processing of these absence-based events, as they may seem less relevant or not worth the effort. Given that **D** events are more likely to be neglected under high cognitive load (Arkes & Harkness, 1983), the cognitive deficits associated with depression (Bowie et al., 2017) might make situations more cognitively demanding, causing depressed individuals to neglect **D** events more easily than non-depressed individuals, which further diminishes their perceived sense of control.

Moreover, researchers found that the depressive realism effect is not replicated when there is a passive observer present (Benassi & Mahler, 1985) and it is replicated when there is an active observer present (Martin et al., 1984). Depressed participants tended to overestimate how much control the active confederate had over the shared outcome, suggesting that individuals attributed more control to others than the self. Cognitive biases are thought to drive this reduced sense of agency in depressed individuals, who often believe that other agents and/or external factors have more control over their environment (Aiken & Baucom, 1982;

Benassi et al., 1988). These biases may result in depressed individuals processing other potential predictors differently from less depressed individuals, leading to a distorted perception of the target relationship and an underestimation of contingencies of outcomes to their own actions, and thereby more accurate ratings regarding their lack of control.

Furthermore, researchers propose that the depressive realism effect may arise because individuals with higher levels of depressive symptoms exhibit lower levels of behaviour, sampling their environment less (Blanco et al., 2012; Ferster, 1973; Lewinsohn et al., 1980). Since behaviour has the potential to fill empty temporal spaces with information that strongly confirms control if it exists (e.g., action-outcome pairings, or **A** events) (Matute, 1996), lower levels of behaviour would produce fewer **A** events (Blanco et al., 2009). This reduced interaction with the environment can result in less data about the outcomes of their actions, further diminishing their perceived control. How people with varying levels of depression sample their environments to produce data to perceive contingencies will be explored in the second half of the thesis.

To conclude, it is worth examining whether the processing of **D** events differs for individuals with more depressive symptoms. This could help explain why the depressive realism effect only emerges in specific contexts, such as under high outcome density and in zero contingencies, but not in positive contingencies. Moreover, it is crucial to explore whether the presence of a separate (non-target) predictor-outcome relationship affects how individuals, particularly those with higher depressive symptoms, perceive the target predictor-outcome relationship. As depressive realism is often linked to biases that over-attribute control to external factors, answering this question could shed more light on how these cognitive distortions influence contingency learning, especially in situations where the outcome is not being directly competed over. Finally, understanding how individuals with more depressive symptoms respond to the co-absence of target cues and outcomes (**D** events) when a non-target

predictor-outcome relation is present may reveal whether their perception of **D** events can be altered. This could offer insight into whether depressive realism and reduced agency can be mitigated through changes in their learning environment.

I explore these questions in Experiments 1-3, reported in Chapter 4. In the next chapter, I review the theoretical frameworks that inform these research questions.

Chapter 3:

Mechanisms of Contingency Learning

In this chapter, I examine the key mechanisms proposed for contingency learning, as action-outcome contingency learning is relevant for perceiving how much agency one has over their environment, and I evaluate these frameworks in light of the trends discussed in **Chapter 2**. The primary focus is on how statistical, associative, and inferential models explain the processing of **D** events and predict the influence of non-target actions/cues and outcomes on the learning of target action-/cue-outcome contingencies. This is because the co-absence of an action and outcome (**D** events) may be informative in deciphering whether one has (positive) agency over an outcome in a situation where they are the only candidate agent, and it may be even more informative in a multi-agent situation to distinguish who has control.

Rather than providing an exhaustive review of each model within these classes, I aim to explain the general principles underlying each approach and how they propose to explain **D** events and scenarios involving multiple potential predictors. I begin by discussing the most well-known model within each class of models before introducing other relevant models.

Statistical Accounts

Statistical theories of learning were originally developed to account for how individuals understand the contingencies between actions/cues and outcomes. I begin with this class of models, as I have already introduced the key model from this group: ΔP. In this section, I will review how various statistical models, including ΔP, propose the role of **D** events and how contingency learning unfolds in scenarios involving multiple predictors. However, I will demonstrate that these models fall short in explaining certain observed trends in contingency learning, and that research indicates that individuals do not represent statistical information exactly as these models suggest, even though their judgments often align with the contingencies generated by them. Therefore, ΔP remains a useful descriptor of the objective unidirectional relationship between predictors and outcomes, but it does not offer a comprehensive mechanistic account of how contingency learning occurs.

This class of models emphasises the role of statistical computations in understanding relationships between predictors and outcomes. It suggests that learners track frequencies of events and perform rule-based calculations to form expectations about the future. Previous research has shown that under certain conditions, participants' judgements of contingency strongly correlate with ΔP, a normative measure of the objective one-way contingency between an action/cue and an outcome introduced earlier (Allan, 1980; Allan & Jenkins, 1980; Chatlosh et al., 1985; Dickinson et al., 1984; Wasserman et al., 1993). This finding provides fundamental support for statistical theories of contingency learning.

To recap, ΔP is calculated by subtracting the probability of the outcome occurring without the action/cue from the probability of it occurring with the action/cue present: $\Delta P = P(O|A) - P(O|\sim A)$, with values ranging from -1 to +1. A positive ΔP indicates the action/cue predicts the outcome, while a negative ΔP signals that the action/cue inhibits the outcome. ΔP is based on four event types: **A**, where both action/cue and outcome are present, **B**, where the action/cue is present and outcome is absent, **C**, where the action/cue is absent and the outcome is present, and **D**, where both the action/cue and outcome are absent. **A** and **D** events inform a positive contingency while **B** and **C** events inform a negative contingency.

However, some studies found that participants' judgements of contingencies also aligned with other rule-based measures, and that different participants' judgements were correlated with different rules, suggesting that people may adopt different strategies (Allan, 1980; Allan & Jenkins, 1980, 1983; Wasserman, Chatlosh, et al., 1983). Unlike ΔP, which is the difference in the probability of the outcome occurring with the predictor versus without the predictor ($\Delta P = (\mathbf{A}/\mathbf{A}+\mathbf{B}) - (\mathbf{C}/\mathbf{C}+\mathbf{D})$), these other rules used different approaches. They might compare the frequency of outcomes with and without the predictor (**A** - **C**) or count the number of co-occurrences of the cue and outcome (**A**) only.

The only other rule that has come close to being as good of a model of how individuals perceive contingencies as ΔP is the ΔD rule, where ΔD = (**A** + **D**) - (**B** + **C**) (Allan & Jenkins, 1983; Inhelder & Piaget, 1958; Wasserman, Chatlosh, et al., 1983). While both rules require tracking all four types of events, they differ in what they compute. ΔP focuses on the difference in probabilities, calculating how likely the outcome is to occur with the action/cue present compared to when it is absent. This approach captures not just the presence of confirming or disconfirming evidence but how consistently the action/cue predicts the outcome based on conditional probabilities. In contrast, ΔD focuses on the raw difference between the number of confirming events (**A** and **D**) and disconfirming events (**B** and **C**), simply tallying the frequencies without considering proportionality. By centring on raw event counts, ΔD gives a broader sense of the balance between evidence for and against a contingency but does not account for how reliably the action/cue predicts the outcome. This is why ΔP, which accounts for the actual likelihoods, has consistently proven to be a better predictor of how individuals perceive contingencies across studies.

To evaluate whether people use the ΔP rule to perceive contingencies between actions/cues and outcomes, Wasserman and colleagues (1990; 1993) investigated whether participants' estimates of P(O|A) and P(O|~A) matched the programmed values of these probabilities. They found discrepancies between participants' estimates and the actual conditional probabilities. However, if people truly base their contingency judgements on those probabilities, as suggested by statistical models of contingency learning, then the ΔP computed using their inaccurate estimates should still better predict their ratings than the objective contingency. Contrary to this expectation, the researchers found that the objective contingencies explained participants' ratings more accurately than the ΔP calculated using their estimates of the probabilities. This challenges the idea that contingency judgements are fully

driven by internal calculations of probabilities, while still indicating sensitivity to objective contingencies.

Additionally, while statistical models treat the probabilities as independent, it was found that judgements of the two probabilities inversely influenced each other. The researchers attempted to improve the calculations by adding weights to the subjective estimates of P(O|A) and P(O|~A), reasoning that the former is more memorable than the latter. However, according to this weighted rule, when the objective contingency is zero, the participants' judgements should consistently be more positive, which the researchers did not find. And this weighted calculation did not fit the data better than the objective contingency. This implies that judgements are not based on internal representations of P(O|A) and P(O|~A) as posited by ΔP, or at least not using the rule that ΔP proposes, while still arriving at the ΔP value.

Nevertheless, the ΔP model, along with other statistical models of contingency learning, propose that **D** events play a critical role in determining the degree to which the outcome depends on the action or cue. **D** events, where neither the action nor the outcome occurs, are essential for clarifying whether the absence of action reliably leads to the absence of the outcome, helping to confirm that the outcome is indeed contingent on the action. Taking the example of pressing a button to turn on a lightbulb, while **A** events (button press and lightbulb turning on) demonstrate the action-outcome relationship, **D** events (no button press and no lightbulb illumination) further verify that the lightbulb does not turn on spontaneously without the action. This reinforces the perception that the action is necessary to achieve the outcome. Thus, **D** events serve a critical function by confirming that inaction leads to no outcome, reinforcing the notion that the action is directly responsible for producing the outcome.

Moreover, ΔP accounts for outcome density (P(O)) effects. As you may recall, researchers found that participants perceived more control over outcomes (higher perceived

action-outcome contingency) as the outcome density increased (Alloy et al., 1985; Msetfi et al., 2007; Vazquez, 1987). Msetfi et al. (2007) noted that the outcome density effects were mediated by the inclusion of inter-trial intervals and modulated by the direction of the contingency, with negative contingencies exhibiting a significantly larger outcome density effect than positive ones. They explained for outcome density effects by regarding inter-trial intervals as **D** events in the ΔP calculations (see Figure 1).

Another explanation of outcome density effects involves an adjustment made to the models by having an increased weighting on the co-occurrence of predictor and outcome (**A** events) during contingency learning (Arkes & Harkness, 1983; Snyder & Swann, 1978; Wason & Johnson-Laird, 1972; Wasserman et al., 1990). In zero contingency conditions, it is proposed that a high outcome density results in more predictor-outcome pairings (see Figure 1, top row of the middle panel), which when given more weight leads to an inflated perception of a positive relationship according to ΔP and probabilistic contrast models. Conversely, in low outcome density, the infrequent outcome leads to fewer predictor-outcome pairings, resulting in the more accurate perception of no actual relationship between the predictor and the outcome. The overemphasis of **A** events may be due to perceptual salience (visual stimulation of the presence of the predictor and outcome), attentional salience (increased focus on present stimuli rather than those that are absent), and/or logical salience (the mental emphasis on the association between stimuli when they occur together) (Arkes & Harkness, 1983; Wason & Johnson-Laird, 1972; Wasserman et al., 1993).

Blanco et al. (2013) expanded on this by investigating the effects of cue density in addition to outcome density. In their study, they independently manipulated the probability of both in a factorial design, while maintaining the non-contingent (zero contingency) relation between the cue and outcome. Their findings revealed that high cue density and high outcome density together amplified the overestimation of contingency, more so than when either factor

was manipulated alone. This result suggests that the biases observed cannot be fully explained by the overweighting of **A** events alone. If **A** events were solely responsible, an additive effect would be expected; however, Blanco et al. found interactive, asymmetrical effects on contingency judgments. Specifically, higher outcome density had a stronger impact on overestimation than higher cue density, even when the number of **A** events was identical. This suggests that other types of events also contribute to contingency learning but are weighted differently. Indeed, prior research supports the notion that individuals weigh different events unequally when judging contingencies, with a typical ranking of event weights as follows: **A** > **B** > **C** > **D** (Kao & Wasserman, 1993; Murphy et al., 2021; Perales & Shanks, 2007; Wasserman et al., 1990; Wasserman et al., 1996), aligning with Blanco et al. (2013)'s results.

However, ΔP alone does not explain the trends regarding multiple predictors. To account for these trends, researchers have developed models built on ΔP, aiming to better explain how learners perceive relationships between predictors and outcomes. One such model is the probabilistic contrast model (Cheng, 1997). The probabilistic contrast model considers other potential predictors in the environment, where the perceived effectiveness of a predictor depends on the presence and effectiveness of the other predictors. The probabilistic contrast model proposes that individuals evaluate the effect of the target predictor while controlling for the presence of other predictors. This means that the model considers how other potential predictors might influence the outcome and adjusts the assessment of the target predictor's effectiveness accordingly. In practice, this involves calculating ΔP for each predictor both overall and conditionally, based on whether other predictors are present or absent. This is then normalised by the frequency of the outcome irrespective of the presence or absence of the predictors.

This approach helps explain how people judge the relative validity of different predictors, wherein one learned predictor-outcome relation interferes with the learning of

another predictor-outcome association, lowering the perceived contingency of the weaker predictor (Baker et al., 1996; Spellman, 1996a). However, the probabilistic contrast model does not account for all behavioural trends observed in situations of multiple potential predictors. For example, it does not explain the enhancement of a weaker predictor's relation with the outcome when the stronger predictor is of the opposite polarity (Darredeau et al., 2009).

Another significant limitation of statistical models, such as ΔP, is their inability to fully capture the shape of typical learning curves. In the original models, the learning is linear, and when adapted to include unequal weighting of different event types, these models predict a non-linear but consistent progression in learning. This does not align with the observed pattern in empirical data, where individuals tend to learn quickly at the beginning, followed by a plateau or gradual slowing down as learning progresses and the contingency becomes more established (Baker et al., 1989; Shanks, 1985a; Wasserman, Chatlosh, et al., 1983).

In conclusion, statistical models, such as ΔP, fail to account for certain key trends observed in contingency learning. While the probabilistic contrast model, an extension of ΔP, explains how the presence of a stronger predictor reduces the perceived contingency of a weaker predictor, it does not account for the enhancement of the weaker predictor's contingency when it is in the opposing direction of the stronger predictor. Additionally, these models fall short in explaining learning curves, where individuals typically show rapid initial learning followed by a plateau, a pattern not predicted by the linear or weighted-event approaches in these models.

Moreover, statistical models do not provide strong mechanistic explanations for contingency learning, as research has shown that individuals' judgments are not based on the internal probability calculations assumed by these models. However, individuals' judgments still tend to align with ΔP, particularly in simpler situations involving a single predictor and outcome. Therefore, ΔP remains a reliable descriptor of the unidirectional relationship between

predictor and outcome, offering a logical foundation for understanding the contributions of the four event types (**A**, **B**, **C**, and **D**) in contingency learning. Despite its limitations, ΔP continues to be a useful tool for understanding the objective structure of contingency relations and therefore provides an important foundation for investigating how agency is learned and perceived.

Associative Accounts

Around the same time as statistical models of contingency learning were proposed, a different perspective on contingency judgement emerged. Researchers observed parallels between human contingency judgement and Pavlovian and instrumental conditioning in nonhuman animals, which used associative learning models to explain behaviour (Alloy & Abramson, 1979; Alloy & Tabachnik, 1984). In this view, the action or cue in human contingency judgement is likened to the conditioned stimulus (CS) in conditioning – typically a neutral stimulus that, through repeated pairing with a significant outcome, becomes predictive of that outcome. The outcome in human contingency judgement is likened to the unconditioned stimulus (US) – an inherently meaningful event, such as food or a shock, which elicits a natural response. The strength of an animal's conditioned response, which is the learned reaction to the CS, reflected its judgement of the contingency between the CS and the US. These parallels suggest that both humans and nonhuman animals share a common mechanism for detecting causal or predictive relationships.

Three main patterns were identified as being shared between human contingency learning and nonhuman animal conditioning. First, when P(O|A) remained constant, human contingency judgements decreased as P(O|~A) increased (Allan & Jenkins, 1980; Alloy & Abramson, 1979; Dickinson et al., 1984). Similarly, when the probability of the US occurring in the presence of the CS (P(US|CS)) was kept constant, excitatory conditioning decreased as the probability of the US occurring in the absence of the CS (P(US|~CS)) increased (Rescorla,

1968). Second, when predictors and outcomes were noncontingent, judgements tended to be influenced by the frequency of the outcome (Alloy & Abramson, 1979; Dickinson et al., 1984; Msetfi et al., 2005). In nonhuman animals, excitatory conditioning often increased with the frequency of the US during exposure to a non-contingent CS-US relationship (Rescorla, 1972). Third, humans could detect negative predictor-outcome contingencies where the predictor predicted the absence of the outcome (Dickinson et al., 1984; Msetfi et al., 2007; Murphy et al., 2021). In Pavlovian conditioning, a stimulus became an inhibitor if it was negatively correlated with the US, such that the US was more likely to occur in the absence of the CS (Rescorla, 1969; Witcher & Ayres, 1980). These similarities make a case for associative learning theories to account for human contingency judgement.

Associative learning accounts propose that contingency judgements are based on associations between the mental representations of the predictor and outcome, with this association being incremented or decremented trial-by-trial according to a learning rule. Associative learning theory posits that a predictor-outcome association is formed and strengthened by **A** and weakened by **B** events. However, explaining how a predictor's association with an outcome is influenced by the outcome alone (when the predictor is absent; **C** events), or by trials with neither predictor nor outcome present (**D** events), requires assuming that the static contextual stimuli also acquire associative strength (Shanks, 1993).

According to the Rescorla-Wagner model (Rescorla & Wagner, 1972), the predictor and context compete for association with the outcome. This means that a strong predictor can diminish the associative value of another (Shanks & Dickinson, 1988; Urcelay & Miller, 2014). When the predictor and outcome are presented together (**A** events), the predictor competes with the context for associative strength. Consequently, during **C** events, the context acquires strength not experienced by the target predictor, resulting in decreased judgements of contingency between the predictor and outcome. Shanks (1986; 1989) demonstrated that

adding a non-target cue during the context-outcome (**C**) trials instead increased contingency judgements regarding the target cue and outcome compared to conditions where only the context was present with the outcome. They suggested that this occurred because the presence of the non-target cue reduced the context's associative strength, reducing its ability to compete with the target cue.

During **D** events, then, where only the context is present without the predictor or outcome, the context's association with the outcome is weakened or extinguished. Additional **D** events therefore reduce the context's ability to compete with the predictor for association with the outcome, thereby strengthening the target predictor-outcome association and increasing the perceived contingency between the stimuli, as observed in Castiello et al. (2022). Based on Shanks' (1986; 1989) studies, it is hypothesised that adding non-target stimuli to context-only trials may further extinguish the context's association with the outcome (i.e., reducing the context's associative strength), reducing its ability to compete with the target predictor, and thereby increasing contingency judgements of the target predictor and outcome, especially when participants have been directed to attend to the target predictor and outcome.

To contextualise this in agency learning, associative learning models explain action-outcome contingency learning by illustrating how individuals discern whether their actions, such as pressing a button, lead to specific outcomes, like turning on a lightbulb. According to models like Rescorla-Wagner, the perception of agency strengthens when a button press turns on the light (**A** events), as the association between the action and outcome is reinforced. If the light turns on without a button press (**C** events), the context, which could be another person's action, begins to compete for the perceived control over the outcome, weakening the individual's sense of control over this lightbulb. On the other hand, when neither the button is pressed nor the light turns on (**D** events), the context's association with the outcome weakens, enhancing the perception that the button press alone controls the light. These models suggest

that the cause of an outcome is always attributed to something, and the context (like other potential predictors) plays a pivotal role in shaping how individuals learn the contingency between their actions and outcomes.

Researchers have assessed how well some associative learning models explain the perceived contingencies in experiments (Dickinson et al., 1984; Kao & Wasserman, 1993). Simulations of associative learning theories, particularly Pearce and Hall (1980) and Rescorla and Wagner (1972), closely matched the experimental results and align with the three main patterns described earlier.

Additionally, associative learning models, like the Rescorla-Wagner model, explain learning curves as they incorporate a prediction error term. This prediction error represents the difference between the expected outcome of an action/cue and the actual outcome. At the beginning of learning, prediction errors are large, leading to rapid adjustments in perceived contingency and steep changes in the learning curve. As individuals learn to predict outcomes more accurately, the prediction errors become smaller, resulting in slower adjustments and a gradual flattening of the learning curve

Moreover, the Rescorla-Wagner model accounts for the outcome density effects. Msetfi et al. (2007) noted that the outcome density effects were mediated by the inclusion of inter-trial intervals and modulated by the direction of the contingency, with negative contingencies exhibiting a significantly larger outcome density effect than positive ones. Similar to their explanation for how statistical models can explain these effects by regarding inter-trial intervals as **D** events, they proposed that the Rescorla-Wagner model can consider inter-trial intervals as context-only events. The context's association with the outcome would be extinguished in such trails, strengthening the association between the predictor and outcome.

For individuals with depression, it is hypothesised that they may process context-only events, such as inter-trial intervals, less effectively. This diminished processing could result in

a weaker or non-existent extinction of the context-outcome association during inter-trial intervals. Consequently, the association between the predictor and the outcome might not be strengthened as it typically would be in non-depressed individuals. As a result, depressed individuals might show no change in perceived association between the predictor and the outcome, even as more inter-trial intervals (context-only) events are presented, explaining the depressive realism effect.

While the Rescorla-Wagner model successfully accounts for many aspects of contingency learning, it encounters challenges when explaining more complex phenomena such as retrospective revaluation. Because values get updated with each trial according to associative learning models, like the Rescorla-Wagner model, explaining phenomena such as retrospective revaluation requires additional assumptions. Retrospective revaluation effects are where an individual learns that certain predictors such as both their own action and another agent's action together leads to an outcome (AB+), but later experiences suggest that one of the other agent's actions are responsible for that outcome (B+), and so the individual retrospectively reassesses and downgrades the importance of the initial predictor (A), reducing their own perceived agency over the outcome (Chapman, 1991; De Houwer & Beckers, 2002a; Dickinson, 1996; Miller & Matute, 1996).

Associative learning models assume that only the associative strengths of predictors present during a trial can change regarding their ability to elicit an outcome (Dickinson, 2001; Van Hamme & Wasserman, 1994; Wasserman & Castro, 2005), which does not explain how the effectiveness of a predictor is altered due to subsequent experiences when the predictor itself is not directly involved in those later experiences. Consequently, researchers revised associative models to account for retrospective revaluation effects (Dickinson, 1996; Van Hamme & Wasserman, 1994). These revised models propose that associations are not only formed between individual predictors and outcomes but also within compounds of predictors

that have appeared together. This allows the associative strength of a predictor to be adjusted even when it is absent during a trial, as long as it has previously been associated with a predictor that is present. For example, if A and B have been associated as a compound (AB+), and later B is confirmed to predict the outcome alone (B+), the associative strength of A can be diminished retrospectively.

However, research has shown that higher order retrospective revaluation can occur, wherein the relationship with the outcome is also influenced by the relationship between stimuli that have never occurred together but are linked indirectly through intermediary associations (De Houwer & Beckers, 2002a; Denniston et al., 2003; Denniston et al., 2000; Macho & Burkart, 2002). For instance, Denniston et al. (2003) showed that when AX+ trials are followed by XY+ trials, rats that subsequently received A- extinction trials responded less strongly to the Y predictor than those that did not receive such extinction trials, even though the A and Y stimuli never occurred together.

Denniston et al. (2000) thereby proposed the extended comparator hypothesis where a predictor's effectiveness results from a comparison between the representation of the outcome directly activated by the target predictor and the representation of the outcome indirectly activated by other predictors with which the target predictor has been associated in the past. Unlike traditional associative learning theories, this hypothesis suggests that predictors do not compete for associative strength during learning; rather, they compete for control over the outcome when the learned associations are being expressed or assessed. The hypothesis can therefore explain higher order retrospective revaluation effects as it suggests a performance-focused, structured information-processing capability, demonstrating that associative learning principles can be revised to account for a web of possible causal connections.

Darredeau et al. (2009) explained how the extended comparator hypothesis predicted their finding that in the presence of a stronger predictor of the opposite polarity, the weaker

predictor's judgements are enhanced. This suggests that adding non-target stimuli to context-only events (**D** trials) could increase perceived contingency between the target stimuli by comparing and contrasting between the contingencies of competing predictors with the target outcome. Non-target predictors and outcomes in the **D** trials could be regarded as being **A** trials for another association, and thus, the non-target stimuli in the **D** trials would be perfectly negatively correlated with the target outcome. A strongly negative predictor (the non-target stimuli) would therefore enhance judgements of a weaker positive predictor of the outcome (the target cue) in the opposing direction of the stronger predictor, i.e., toward the positive side, away from zero (Baker et al., 1993; Baker, Vallée-Tourangeau, et al., 2000; Darredeau et al., 2009; Vallée-Tourangeau et al., 1998). And a strongly negative predictor would shift judgements of a weaker negative predictor to be less negative, i.e., also toward the positive side. According to this hypothesis, across all baseline contingencies of the target relationship, the target relationship would be perceived as a more positive contingency if **D** events included a non-target cue and outcome that strongly negatively correlated with the target outcome.

Associative learning theories, such as the Rescorla-Wagner model and its extensions, seem to account for all the trends of human contingency learning discussed so far. Associative learning accounts suggest that in a simple situation where an individual is learning whether the outcome is contingent on their actions, **D** events (co-absence of action and outcome) would enhance perceived contingency between the action and outcome by extinguishing the associative strength of the context with the outcome. Then, in a more complex situation where an individual is learning whether they or another agent has control over the outcome, the associative strengths of each agent's actions, and the context's, with the outcome would compete to determine who (or what) has control over the outcome. And, so, if an individual is learning whether they have control over the outcome with another agent beside them controlling another outcome, then the other agent's actions and outcomes would be negatively

contingent on the individual's target outcome, and that may enhance the perceived contingency of the individual's agency over the target outcome in the positive direction.

Inferential Accounts

Inferential models of contingency learning emerged in the late 20th century as an alternative or extension to statistical and associative models (Penn & Povinelli, 2007; Pineño & Miller, 2007). These models emphasise the role of higher-order cognitive processes, such as causal reasoning, hypothesis testing, and logical deduction, in how individuals form judgements about relationships between actions, cues, and outcomes (Johnson-Laird & Khemlani, 2017; Pearl, 2000; Waldmann, 1996; Wason & Johnson-Laird, 1972). Unlike statistical and associative models, inferential accounts suggest that learners actively test hypotheses about causal structures rather than passively accumulating associations or performing rule-based calculations (De Houwer, 2009; Waldmann, 2000; Waldmann & Holyoak, 1992). The "active" aspect of this process is covered in detail in Chapter 5, introducing active inference, whereas this section emphasises the perceptual aspect of inferential reasoning.

Central to inferential models is the idea that learners construct causal models of their environment based on observations and prior knowledge. These models help learners hypothesise about the causal relationships between multiple potential causes (e.g., actions of different agents) and outcomes, which are then tested as new evidence arises (Waldmann & Holyoak, 1992). Beliefs about these causal structures guide the learner's interpretation of events. Beliefs refer to an internal representation that individuals hold about how different elements in their environment are related (Albarracin & Pitliya, 2022). For example, an individual may believe that their button press controls a lightbulb, and they would then test this hypothesis by observing whether the lightbulb illuminates when they press the button and whether it remains off when they do not.

Bayesian models offer a computational framework for this inferential process (Glymour, 2003; Waldmann & Martignon, 1998). Learners hold prior beliefs about causal structures, which are updated with new observations using Bayes' rule. This allows individuals to assess the likelihood of various causal hypotheses and revise their beliefs accordingly, resulting in a set of posterior probabilities that reflect the learner's perceived causal relationships (Hagmayer & Waldmann, 2007; Tenenbaum & Griffiths, 2002; Tenenbaum et al., 2006). For instance, if pressing the button turns on the lightbulb, the learner's belief that they control the light is strengthened. If the light does not turn on immediately after pressing the button, the learner's belief that they control the light is weakened.

Because causal structures are represented as probability distributions rather than single-point estimates, inferential models allow learners to entertain multiple competing hypotheses and weigh them against the evidence. This flexibility enables these models to account for more complex phenomena, such as forward and backward blocking and relative validity (Gopnik & Wellman, 2012; Kruschke, 2008). I will now describe findings from studies that used cues as predictors (Beckers et al., 2005; De Houwer et al., 2002; Peter et al., 2003), but for clarity, I will use illustrative examples involving people trying to accomplish a goal to demonstrate the key concepts they investigated.

Imagine you observe one person partially opening a heavy door. Later, you see that person and another person pushing the same door, but still only partially opening it. You would infer that the second person is not contributing significantly (blocking). However, if you saw that one person could fully open the door and that the one person and another person together could also fully open it, you would find it difficult to determine whether the second person is contributing (no blocking). In contrast, if you observed two people, separately and together, lighting a candle, you would assume that both have equal control, as neither prevents the other from achieving the outcome (no blocking). These examples illustrate how prior beliefs about

the causal relationships, knowledge about outcome properties (total strength of outcome available), and the relative contributions of different predictors (e.g., actions) shape how individuals learn and refine their understanding of which factors are responsible for achieving the desired result, especially when the total effect seems unchanged despite additional efforts. This finding is difficult to reconcile with purely associative or statistical accounts but aligns well with an inferential approach where learners actively reason about the causal structure of the task (Pineño & Miller, 2007).

The causal model itself determines what information the learner considers relevant. (Alloy & Tabachnik, 1984; Arkes & Harkness, 1983; Shaklee, 1983; Tversky & Kahneman, 1981). Inferential models also account for how learners interpret **D** events (co-absence of action and outcome) in contingency learning. If an individual is learning whether their button press controls a lightbulb, **D** events (no press and no illumination) support their hypothesis by showing that inaction does not produce the outcome. Furthermore, top-down causal knowledge can refine the learner's interpretation of events. For example, if the learner knows that the button activates a circuit with a delay, they will not treat the gap between action and outcome as separate **B** (action-only) or **C** (outcome-only) events, which would weaken their perception of the contingency. Instead, this causal understanding allows them to integrate the delay into their model and strengthen their perceived control over the lightbulb.

In multi-agent scenarios, inferential accounts propose that learners reason about the causal roles of different agents. For instance, if both an individual's action and another agent's action precede an outcome, they might hypothesise various causal structures, such as both actions being necessary or only one being sufficient. In this context, **D** events can help distinguish between agents by confirming which actions are essential to producing or preventing the outcome. Observations are then used to test and refine these hypotheses.

While inferential models are adept at explaining complex phenomena such as blocking and retrospective revaluation, some researchers argue that these models may be unnecessarily complex for simpler forms of learning (Penn & Povinelli, 2007; Pineño & Miller, 2007). In such cases, automatic associative learning may still occur at a more basic level, where associations between actions and outcomes are formed without conscious reasoning. For example, in scenarios where an individual repeatedly experiences a direct pairing of an action and an outcome, like pressing a button to turn on a light, learning may occur through automatic, associative processes rather than through inferential reasoning. This suggests that both forms of learning – associative and inferential – can coexist, with automatic processes operating at the bottom level to handle simpler contingencies, while more complex scenarios may require inferential reasoning. To account for this distinction, dual-process models have been proposed, integrating both associative and inferential processes. These models suggest that automatic, associative learning operates in parallel with controlled, inferential reasoning, with the relative influence of each varying depending on the context (McLaren et al., 2018; Penn & Povinelli, 2007; Pineño & Miller, 2007).

Additionally, Kalman filtering, a Bayesian approach, has been proposed as an extension of the Rescorla-Wagner model to offer a more flexible, computational framework (Kalman, 1960; Kang et al., 2024; Kruschke, 2008). Whereas the Rescorla-Wagner model updates associations incrementally based on prediction errors, Kalman filtering continuously adjusts predictions by integrating prior beliefs with new evidence. This Bayesian updating allows for dynamic learning, accounting for both automatic, associative learning and more controlled, inferential reasoning, depending on the complexity of the task and the learner's expectations.

In conclusion, inferential reasoning offers a compelling account of contingency learning that emphasises the role of higher-order cognitive processes and causal reasoning in interpreting and refining causal relationships. These models excel in explaining complex

phenomena, the role of prior knowledge, and the influence of causal structures on learning. While they face challenges in accounting for more automatic forms of learning, integrating inferential reasoning with insights from associative and statistical models may offer a more comprehensive understanding of contingency learning across diverse contexts.

Chapter 4:

Modifying Contingency Learning through D Event Manipulation (Experiments 1-3)

The first three chapters of this thesis have highlighted trends and proposed mechanisms by which individuals learn the extent of their control – i.e., agency – over outcomes in their environment. **Chapter 1** introduced the concept of agency, defining it as the degree of control an individual perceives over their actions and the resulting outcomes, which maps onto implicit and explicit senses of agency respectively (Moore et al., 2012; Synofzik et al., 2008). This thesis focuses on the explicit sense of agency, which involves learning whether outcomes are contingent on one's actions (i.e., learning action-outcome contingencies) (Bandura, 1982; Friston et al., 2013), and if so, determining whether that control is positive (excitatory) or negative (inhibitory), and to what extent. Additionally, it was noted that depressive symptoms are associated with a diminished explicit sense of agency (Abramson et al., 1978; Beck, 1967; Seligman, 1974).

Chapter 2 reviewed key experiments that explored how individuals learn and perceive both cue-outcome and action-outcome contingencies, as these two types of learning share similar underlying cognitive processes (Béghin & Markovits, 2022; Jenkins & Ward, 1965). Research has demonstrated that individuals typically perceive these contingencies with high accuracy in straightforward scenarios involving a single action or cue and a corresponding outcome, across varying directions of contingencies – negative, zero, and positive (Allan & Jenkins, 1980; Arkes & Harkness, 1983; Chatlosh et al., 1985; Shanks, 1985a; Wasserman et al., 1993).

In typical contingency learning experiments, participants experience a series of trials that present different proportions of four types of events: **A**, **B**, **C**, and **D**. These events represent combinations of the presence and absence of an action or cue and its corresponding outcome. Specifically, **A** events occur when both the action or cue and the outcome are present, **B** events occur when the action or cue is present but the outcome is absent, **C** events occur when the action or cue is absent but the outcome is present, and **D** events occur when both the action or

cue and the outcome are absent. After observing these trials, participants are asked to rate the perceived relationship between the action or cue and the outcome, typically on a scale ranging from -1 to +1, where negative values indicate inhibitory relationships and positive values indicate excitatory ones. These ratings are then compared to the objective contingency, which is defined by $\Delta P = P(O|A) - P(O|\sim A) = (\mathbf{A}/\mathbf{A}+\mathbf{B}) - (\mathbf{C}/\mathbf{C}+\mathbf{D})$, a normative index reflecting the strength of the one-way contingency between an action or cue and an outcome (Allan, 1980).

In terms of how individuals process these events, **A** events are typically informative of positive contingencies, while **B** and **C** events inform negative contingencies (Murphy et al., 2021; Schustack & Sternberg, 1981). The role of **D** events, however, remains unclear, despite their logical relevance in indicating that the action or cue is necessary for the outcome. Research has shown that **D** events are often neglected, particularly under conditions of cognitive load or when deemed less relevant to the task at hand (Arkes & Harkness, 1983; Kao & Wasserman, 1993; Mata et al., 2015). Nonetheless, some studies suggest that when **D** events are processed, they can enhance the perception of positive contingencies (Castiello et al., 2022; Msetfi et al., 2007).

Another key finding is that the presence of other (non-target) actions or cues can influence the perceived relationship between the target action or cue and the outcome. In cases where the directions of contingencies of both predictors (target and non-target) are aligned (either both positive or both negative), the stronger predictor reduces the perceived contingency of the weaker predictor (Baker et al., 1993; Spellman, 1996a; Tangen & Allan, 2004). When the predictors are in opposing directions, the weaker predictor's perceived contingency is enhanced in the opposite direction of the stronger predictor (Baker et al., 1993; Baker, Vallée-Tourangeau, et al., 2000; Darredeau et al., 2009). However, these studies only examined cases where both predictors shared the same outcome, leaving it unclear how a separate non-target action/cue-outcome relationship might influence the target relationship.

Therefore, the experiments in this chapter examine whether the presence of a non-target cue and outcome affects the processing of **D** events and their impact on learning the contingency between the target cue and outcome. We focus on the perceptual aspects of contingency learning, opting to use cues rather than actions, even though the broader aim of this thesis is to explore agency perception (action-outcome contingencies). This choice allows us to examine the effects of these manipulations in a simpler, more controlled setting, where the experimenter can more precisely manipulate the events. This approach provides a clear understanding of the underlying mechanisms before shifting the focus to the more complex perceptual and active processes involved in agency, which will be explored in **Chapter 6** through Experiments 4 to 6. It also allows us to assess whether depression is related to contingency learning in general or if it is specific to action-outcome contingency learning, which pertains to agency.

In the experiments, we varied the frequency of **D** trials while keeping **A**, **B**, and **C** trial frequencies constant, allowing us to isolate the effect of **D** events on perceived contingency. The experiments were conducted across three baseline contingencies: negative, zero, and positive. Additionally, we manipulated the content of **D** trials by either keeping them empty or adding non-target cues and outcomes and compared the effects of these manipulations on participants' contingency ratings. Participants rated the perceived contingency between the target cue and outcome on a scale from -10 to 10.

Chapter 3 explored the theoretical models underpinning contingency learning, including statistical, associative, and inferential models. Statistical models offer limited insight into the cognitive mechanisms underlying contingency learning, however, ΔP remains a valuable tool as a measure of the relationship between actions or cues and outcomes, effectively quantifying the strength of this contingency and describing the logical information that the four event types provide. Associative and inferential models provide stronger mechanistic

explanations, and some researchers suggest that a combined approach, integrating top-down inferential reasoning with bottom-up associative learning, may offer the most comprehensive understanding. In our experiments, we used cues and outcomes with no inherent causal significance to minimise the influence of participants' pre-existing knowledge, focusing on bottom-up learning processes.

Our hypotheses are as follows: participants will distinguish between the different types of contingencies, assigning higher ratings to positive contingencies, followed by zero, and then negative contingencies. However, individuals with higher depressive symptom levels may show a weakened ability to make these distinctions, as previous research indicates that depression is associated with weakened action-outcome contingency learning (Ackermann & DeRubeis, 1991; Alloy & Abramson, 1982; Castiello et al., 2020; Msetfi et al., 2007).

We also hypothesise that the perceived contingency will be influenced by the frequency of **D** trials. Specifically, we hypothesise that increasing frequencies of **D** trials would increase perceived contingency, as associative learning models suggest that context-only events (**D** trials) reduce the context's association with the outcome, thus enhancing the target cue-outcome association. We further expect that participants with higher levels of depressive symptoms process the **D** trials less effectively and therefore would provide lower ratings of contingency. This is based on the results from Msetfi et al. (2007), where the inter-trial intervals, which were likened to **D** trials as they are context-only periods between trials, were processed less by participants with higher depressive symptoms.

In terms of the hypotheses regarding non-target stimuli, we predict that **D** trials containing non-target stimuli would enhance the effect of empty **D** trials. As such, we hypothesise that the perceived contingency (ratings) between the target cue and outcome would increase more as the frequency of **D** trials with non-target stimuli increases compared to empty **D** trials. This prediction builds on Shanks' (1986; 1989) findings, which demonstrated that the

presence of a non-target cue during context-outcome (**C**) trials reduced the context's associative strength, enhancing contingency judgments regarding the target cue and outcome. Shanks proposed that this effect occurs because the non-target cue competes with the context for association with the outcome, reducing the context's ability to interfere with the target cue's association. In our experiments, we hypothesise that a similar associative competition occurs when non-target stimuli are added to **D** trials, extinguishing the context's association with the target outcome, thereby increasing the perceived contingency between the target cue and outcome. For participants with higher depressive symptoms, while the ratings are expected to increase with respect to the conditions with empty **D** trials, we hypothesise that these ratings will remain lower than those of participants with lower depressive symptoms, as the enhanced processing of **D** trials due to the non-target stimuli should be proportional to the baseline processing of empty **D** trials.

Experiment 1

Participants

We recruited 64 participants from Amazon Mechanical Turk (MTurk) (Buhrmester et al., 2011). Participants were between the ages of 18 and 50, were free from photosensitive epilepsy and could complete the task using a desktop computer or laptop, but not a tablet or phone. We recruited 64 participants to ensure a robust final sample size after applying the exclusion criteria. Our previous studies (Castiello et al., 2022; Murphy et al., 2021) used sample sizes ranging from 40 to 51 participants. By recruiting slightly more participants, we aimed to account for potential exclusions, enhance statistical power, and mitigate the impact of unforeseen data quality issues.

To ensure compliance with attentional demands of the task, those who gave the same response 90% or more of the time across all conditions were excluded; no participants were excluded based on this criterion. The decision to set the threshold at 90% was informed by

unpublished experiments conducted by the same research groups using a similar task. We examined the distribution of response variability across participants, excluding those whose responses fell more than two standard deviations away from the mean. Based on these findings, a 90% threshold was then established across these experiments to ensure attentional engagement with the task. Additional exclusion criteria included participants who failed to identify a particular shape from an array of five shapes, and those who responded "yes" when asked if they were distracted by their smartphones or other environmental factors while completing the task. Seven participants were excluded based on these additional criteria in this experiment. The final sample comprised of 57 participants, with 36 males and 21 females, and a mean age of 36.91 (SD = 6.40). The ethics committee at the University of Oxford approved this research (reference code: R74631/RE001).

Participants completed the Beck Depression Inventory-II (BDI-II; Beck et al. (1996)) as a measure of depressive symptoms. The mean BDI-II score among participants was 13.368 (SD = 14.115), with scores ranging from 0 to 43. According to BDI-II categorisation (Beck et al., 1996), around 60% of participants fell into the minimal depression range (scores of 0–13; N=34), 17.5% into the mild-to-moderate range (scores of 14–28; N=10), and 22.5% into the severe range (scores of 29–63; N=13).

Materials and Procedure

Participants were presented with a briefing sheet detailing the task, data usage and storage, contact information for queries, and their right to withdraw. Participants provided informed consent and demographic information (age and gender) and confirmed they did not have a history of seizures before completing the computer task.

Participants were told that they will be watching a series of rapidly presented shapes and objects, and after each series, a question screen will appear, asking them to rate the degree of relatedness between the two. They were informed that the ratings were to be provided on a

scale from -10 to 10, wherein 10 indicated a strong positive rating that the shape and object were always presented together and absent together, -10 indicated a strong negative rating that the shape was always present without the object and the object was always presented without the shape, and 0 indicated no relationship between the shape and object.

They were instructed to keep their eyes on the fixation cross in the centre of the screen, which was present on all trials. To prevent the mouse cursor from visually interfering with the contingency learning task by acting as a non-target stimulus or causing distraction during the contingency learning trials, we removed it from the screen and informed participants of this change beforehand. Participants were allowed a maximum of two five-minute breaks between streams, and they were informed of that. Participants then completed the task. Participants' attentiveness was encouraged by asking them twice at the end of each trial stream whether they recalled a specific stimulus; note that this was for maintaining engagement and was not used as a criterion for elimination.

After completing the task, participants were asked to fill out BDI-II, a self-report questionnaire designed to assess the presence and severity of depressive symptoms. The questionnaire comprises 21 items, each assessing whether participants have experienced specific symptoms, with responses scored on a 4-point scale ranging from 0 (indicating the symptom is not present) to 3 (indicating the symptom is intensely experienced). The total score ranges from 0 to 63, with higher scores indicating greater levels of depression. BDI-II has demonstrated strong reliability and validity across various populations (Wang & Gorenstein, 2013).

After completing the depression questionnaire, participants were asked to identify a particular shape from an array of five shapes presented to them, and they were asked if they gave their full attention to the experimental task or were distracted by their smartphones or other things in their environment. Participants were then presented with a debrief sheet,

outlining the objective of the experiment. The whole experiment took approximately 20 minutes to complete. The exact instructions and procedure are detailed in Supplementary Material 1.

We used the Gorilla Experiment Builder (www.gorilla.sc) to create and host our experiment (Anwyl-Irvine et al., 2020). The stimuli were created using Microsoft PowerPoint for Mac Version 16.45. The stimuli were shapes for target cues, selected from the collection available on PowerPoint, and illustrations of objects for target outcomes, selected from the pool of Icons in PowerPoint. Mathematical Greek letters were selected for the non-target stimuli from the category 'Math Symbols' in PowerPoint. The symbols were manually selected by the researchers to ensure the symbols used were not commonly used and recognised symbols. Stimuli with no inherent causal significance were used to minimise the influence of participants' pre-existing knowledge. For each symbol, the font was 'Calibri (body)', with font size 600pts. All slides from the PowerPoint were exported as .png files to be used as stimuli. The dimensions for each image were 720x720 (WxH) pixels.

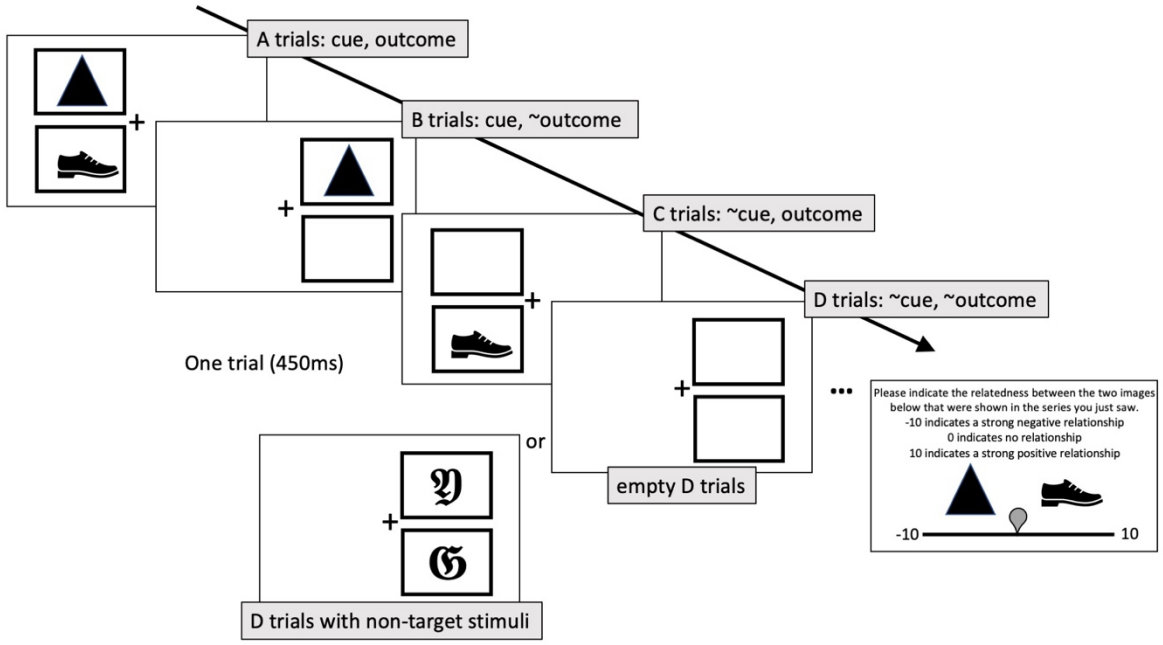
Task Design

Each experimental condition consisted of a stream of trials, wherein each trial was one of the four trial types (**A**, **B**, **C**, or **D**; see Figure 3). The rapid streamed-trial procedure was employed (Castiello et al., 2022; Crump et al., 2007; Hannah et al., 2009; Murphy et al., 2021), wherein trials were presented in quick succession with different proportions of the four trial types.

In this experiment, we established two sets of **A**, **B**, **C**, and **D** trial combinations to create "baseline contingency" conditions, producing both negative and positive baseline contingency conditions. In the negative baseline contingency condition, there were 4 **A** trials and 44 **B** trials, while the positive baseline contingency condition comprised 44 **A** trials and 4 **B** trials. Both conditions included 16 **C** trials and 16 **D** trials.

To examine the effect of **D** trials, we then manipulated the frequency of **D** trials across three levels (0, 16, 72) while keeping the number of **A**, **B**, and **C** trials constant within each baseline contingency condition. The "objective contingency" refers to the ΔP value of the trial-stream, which changes as a function of the frequency of **D** trials within each baseline contingency condition. For a detailed breakdown of the number of **A**, **B**, **C**, and **D** trials in each condition and the respective ΔP values, please refer to Supplementary Material 2A.

Figure 3.
The Task for Experiments 1-3.



Notes. The experimental task presented a series of trials that varied in the proportion of **A**, **B**, **C**, and **D** trials were presented to participants; this is called a trial-stream. The content of **D** trials could change, wherein they could either be empty or contain non-target stimuli. Each trial was presented for 450ms, and there was a fixation cross in the middle of each trial. Every trial was displayed in a sequentially alternating position on the screen respective to the fixation cross (i.e., right-left-right-left…). At the end of the trial-stream, participants were to provide a rating, on a scale from -10 to 10, regarding the association between the shape and object presented in the trial-stream. Note that the rating scale illustrated in this figure is not an accurate representation of what participants saw – please see Supplementary Material 1 for the screenshot of the rating scale participants saw.

We used trial markers in the form of solid-line rectangles to indicate the spatial location of the cue (top trial marker) and outcome (bottom trial marker) that would be present or absent according to the trial type; see Figure 3. The trial markers served as a local context in which the stimuli were presented. We manipulated the content of **D** trials by either leaving them empty or adding non-target stimuli (see Figure 3). In the empty condition, **D** trials consisted of just the fixation cross and trial markers, without any non-target stimuli. In the non-target condition, a single non-target stimulus appeared in both the cue and outcome trial markers. For each trial stream, one non-target cue and one non-target outcome were randomly selected from a pool of stimuli, without replacement.

The stimuli were shapes and illustrations of objects randomly selected, without replacement, from their respective pools of stimuli (10 stimuli in each pool) to be displayed in the **A**, **B**, and **C** trials in the trial-stream. The target shape and object pair remained the same within each trial-stream, with the pair changing across trial-streams. Trials were displayed sequentially in an alternating position respective to the fixation cross (i.e., right-left-right-left…; see Figure 3). This was employed to ensure that the sequential trials were distinguishable from one another and avoid ambiguities regarding the duration of a trial when two consecutive trials of the same trial type were presented since there were no programmed inter-trial intervals.

A within-subjects design was adopted, with 10 experimental conditions that mapped onto a combination of the three experimental manipulations conducted: [2 levels of contingencies (negative and positive) x 2 levels of **D** trial content (empty and non-target) x 2 levels of **D** trial frequency (16 and 72)] + 2 **D** trial frequency 0, one for each baseline contingency manipulation as we cannot conduct the content manipulation when there are no **D** trials presented in that condition; see Table 2. The experimental conditions were blocked based on the **D** trial content manipulations. The sequence of these blocks was fully counterbalanced

across participants. The second level of blocking was by baseline contingency. Moreover, within each trial-stream, we further blocked the trials into four sub-blocks, wherein each sub-block of trials had the same contingency as the overall trial-stream's objective contingency by containing the ratio of trials relevant for the contingency. The presentation of these sub-blocks was also counterbalanced to avoid order effects.

Table 2.
Task Design for Experiment 1.

Baseline Contingencies	**D Trial Content**	
	Empty	Non-Target
	0	
Negative	16	16
	72	72
	0	
Positive	16	16
	72	72

Notes. The numbers in the cells (0, 16, 72) refer to the levels of **D** trial frequency manipulation. When there are no (0) **D** trials, there is no content manipulation as there are no **D** trials to manipulate the content of, hence its placement in the middle of the two columns in the table.

Data Analysis

All data pre-processing, statistical analyses, and visualisations were conducted in R (R Core Team, 2021). The *dplyr* (Wickham et al., 2019) and *ggplot2* packages (Wickham et al., 2016) were used to facilitate data wrangling and visualisations. For all analyses, the significance threshold was set at an α level of 0.05.

A data-driven analysis approach was employed to identify trends and relationships directly from the data. The analysis began with fitting a full linear mixed-effects model (LMM) using the *lmer()* function from the *lme4* package in R (Bates et al., 2015; Bates et al., 2009).

This model was chosen to assess the influence of both fixed effects, experienced by all participants, and random effects, which varied between participants.

The full LMM included all relevant predictors: **D** trial frequency, **D** trial content, baseline contingency, block order, depressive symptom severity (BDI-II scores), and their interactions. Frequency and depression scores were treated as continuous variables. The levels of frequency manipulation (0, 16, and 72) were standardised by subtracting the mean from each value and dividing by the standard deviation. This transformation centred the variables around zero and scaled them to have unit variance, to improve model convergence, enhance numerical stability, and facilitate proportional contrast comparisons (Gelman & Hill, 2007; Harrison et al., 2018; Schielzeth, 2010). Baseline contingencies, content, and block order were treated as discrete variables with specified factor levels. Block order referred to whether participants first experienced the block of **D** trials that were empty, contained non-target stimuli, or had no **D** trials. The random intercept was set as the participants to account for variability in baseline ratings across participants.

To refine the model, a backward stepwise algorithm was applied using the *step()* function in the *lmertest* R package (Kuznetsova et al., 2015). This algorithm iteratively removed the least significant fixed effects based on the Akaike Information Criterion (AIC) (Venables & Ripley, 2013), aiming to simplify the model while maintaining the complexity necessary to account for random effects and covariances that LMMs are designed to handle (West et al., 2022).

After this backward selection process, a series of nested models were constructed and compared in a hierarchical manner to ensure that the resulting model was indeed the best-fitting model. Model assumptions, including normality of residuals, homoscedasticity, and linearity, were checked to ensure the validity of the final model. The linearity of the relationship between the independent and dependent variables was evaluated by plotting the observed ratings against

the predicted (fitted) values. The normality of residuals was assessed using both a histogram and a Q-Q (quantile-quantile) plot. The assumption of homoscedasticity (constant variance of residuals) was checked by plotting the residuals against the fitted values to ensure that the spread of residuals remained constant across the range of predicted values. If model assumptions were met, the resulting model was analysed using a standard LMM. If model assumptions were violated, robust standard errors (SEs) were applied to the LMM. Specifically, a cluster-robust variance-covariance estimator was used to adjust the SEs. This adjustment ensures that the *p*-values are more reliable, making the conclusions drawn from the model robust to assumption violations.

The task design was not fully factorial, meaning that not all possible combinations of factors were present in the data. Specifically, when the frequency of **D** trials was zero, the content manipulation (empty or non-target) could not be applied as there were no **D** trials, resulting in the absence of data for these combinations. In some contexts, this might be seen as a limitation because it restricts the ability to estimate the effects of certain combinations directly. However, in this study, the primary focus is on understanding how the presence and content of **D** trials affect perceived contingency. Since the condition where **D** trials are absent inherently does not involve content manipulation, the lack of this combination in the design is not a significant limitation and does not impede the ability to test the key research questions. Furthermore, a sensitivity analysis was conducted to assess the impact of these missing combinations, and the results remained consistent, confirming that the missing data did not significantly influence the overall findings.

For post-hoc analyses, estimated marginal means were computed using the *emmeans* package in R (Lenth, 2023) for the factor levels in the significant predictors of the resulting model. All variables were treated as discrete for these analyses. Participants were grouped into "low depressive symptoms" (N=34, 60%) and "high depressive symptoms" (N=23, 40%) based

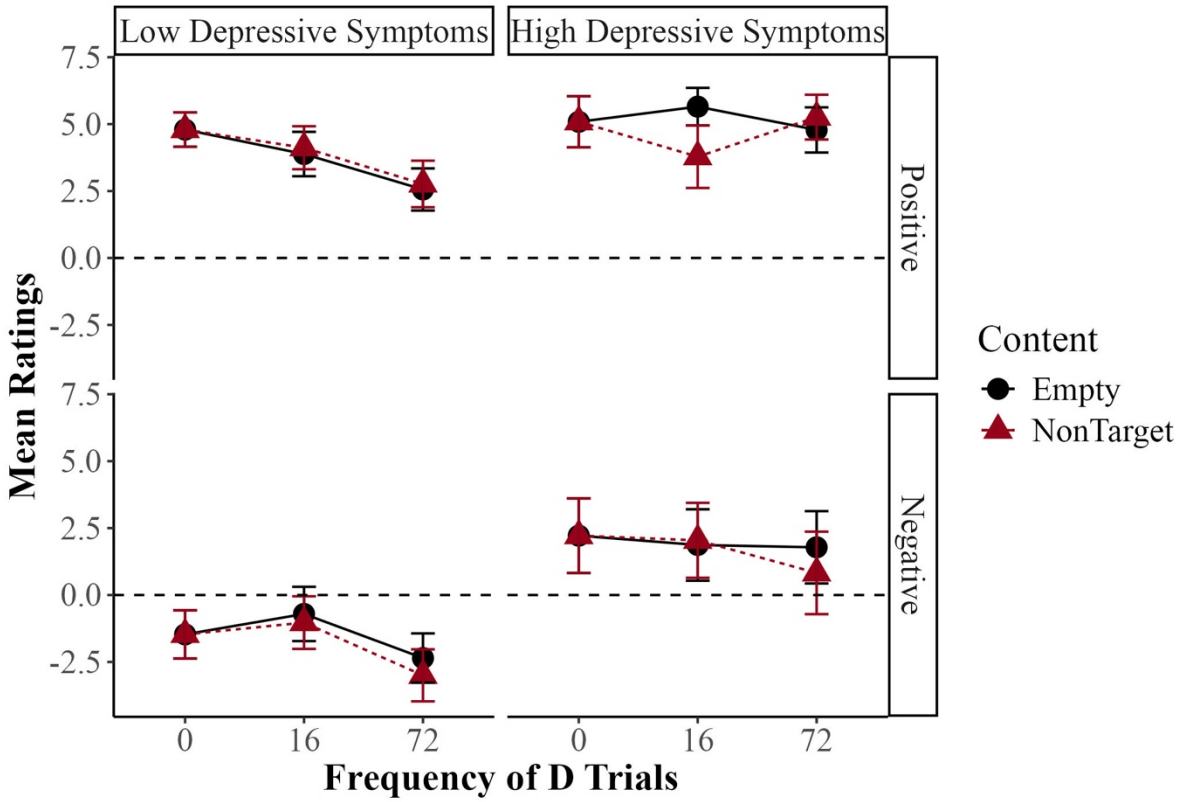
on established cut-offs in non-clinical populations, where a score of 13 or below indicates minimal to no depressive symptoms, and a score above 13 suggests the presence of higher depressive symptoms (Beck et al., 1996; von Glischinski et al., 2019). This grouping ensured that marginal means could be estimated for pairwise comparisons, as the small sample sizes in the four separate groups as specified in BDI-II (Beck et al., 1996) resulted in insufficient data for some conditions, making the calculation of averages across factors impossible. We selected the terms "low" and "high" to acknowledge that depression exists on a continuous spectrum (American Psychiatric Association, 2013). Pairwise comparisons of these estimated marginal means were then conducted to determine which specific levels of manipulation differed significantly. The *p*-values were adjusted for multiple comparisons using the Tukey method.

Results

The findings from this experiment revealed three key results: First, participants generally distinguished between positive and negative contingencies, providing higher ratings for positive contingencies and lower ratings for negative ones, supporting our hypothesis. However, participants with higher levels of depressive symptoms found this discrimination more difficult, reflected in their higher ratings for negative contingencies compared to participants with lower depressive symptom levels. Second, contrary to our hypothesis, the presence of **D** trials reduced perceived contingencies, with a higher frequency of **D** trials leading to lower ratings. Although, participants with higher levels of depressive symptoms processed **D** trials less, such that the frequency of **D** trials had no effect of ratings. Third, while the content of **D** trials did not significantly influence perceived contingencies, participants who experienced blocks with empty **D** trials first showed further reductions in perceived contingency when later exposed to **D** trials with non-target stimuli; this effect did not vary with depressive symptom levels. Overall, perceived contingencies were influenced by **D** trial frequency and content but not in the way we expected.

To arrive at these findings, a backward stepwise algorithm was employed to refine the full model. It identified three significant interaction effects that were retained in the final model: baseline contingency with block order and BDI-II score ($F(2, 513) = 4.738$, $p = 0.009$), **D** trial frequency with BDI-II score ($F(1, 513) = 4.565$, $p = 0.033$), and **D** trial content with block order ($F(4, 513) = 2.822$, $p = 0.025$). The final model, confirmed through comparison with nested models, was adjusted using robust SEs to ensure that the conclusions were reliable despite violations of model assumptions (see Supplementary Material 3A). I will now elaborate on these three sets of key findings and present the detailed statistical analyses in the following subsections.

Figure 4.
Results from Experiment 1: Ratings as a Function of D Trial Frequency, D Trial Content, Baseline Contingencies, and Depressive Symptom Levels.



Notes. The figure shows mean ratings (y-axis), reflecting perceived contingency between two target stimuli, as a function of the frequency of **D** trials (x-axis) under negative and positive baseline contingencies and low and high depressive symptom levels as measured by Beck's Depression Inventory-II (BDI-II). The data are further split by the content of the **D** trials, which were either empty (round black points) or contained non-target stimuli (red triangle points). Error bars represent the standard error of the mean (SEM).

The Effect of Baseline Contingency.

Participants successfully distinguished between positive and negative baseline contingencies, with higher ratings for positive and lower ratings for negative conditions, as shown in Figure 4 (β = 7.797, SE = 1.688, t(10.697) = 4.620, $p < 0.001$). However, this distinction was less pronounced in participants with higher depressive symptoms. As illustrated in the right panel of Figure 4, participants with higher depressive symptoms provided higher ratings for negative contingencies compared to those with lower symptoms, often rating them above zero (grey dotted line), suggesting difficulty in processing negative contingencies (β = - 0.245, SE = 0.06, t(8.500) = -3.864, p = 0.004). Despite this, they still showed some ability to differentiate between the two types of contingencies.

Post-hoc analyses confirmed that both groups could still discriminate between the positive and negative contingencies. As shown in Figure 4, participants with low depressive symptoms (β = 5.38, SE = 0.506, t(528) = 10.634, $p < 0.001$) and those with higher depression levels (β = 3.19, SE = 0.620, t(528) = 5.146, $p < 0.001$) both provided significantly higher ratings for positive contingencies and lower ratings for negative contingencies. However, the smaller difference in ratings in the higher depressive symptoms group, as indicated by lower effect β estimates and t-values, suggests that while both groups could distinguish between contingencies, this ability diminished with higher depressive symptom severity.

Moreover, while the backward stepwise algorithm initially identified an interaction between block order, baseline contingency, and depressive symptom severity (F(2, 513) = 4.738, p = 0.009), this interaction was not statistically significant when robust standard errors were applied to the LMM (β = 0.184, SE = 0.106, t(17.926) = 1.739, p = 0.099). As a result, no further post-hoc analyses were conducted based on block order, as these interactions did not reach statistical significance.

To summarise, participants generally distinguished between positive and negative contingencies, but this ability weakened in those with higher depressive symptoms, particularly when rating negative contingencies. This pattern highlights that depressive symptoms may influence the perception of negative contingencies, although the overall distinction between contingency types was still present across groups.

The Effect of D Trial Frequency.

Surprisingly, increasing the frequency of **D** trials led to more negative ratings overall (β = -1.154, SE = 0.359, t(513) = -3.214, p = 0.001). However, participants with higher levels of depressive symptoms were less influenced by the frequency of **D** trials, showing less reduced ratings as the frequency of **D** trials increased (β = 0.035, SE = 0.017, t(513) = 2.137, p = 0.033). Pairwise comparisons revealed that participants with lower depressive symptoms rated contingencies significantly lower as the frequency of **D** trials increased (β = -1.574, SE = 0.359, t(513) = -3.214, p = 0.001), while participants with higher depressive symptoms did not show a significant change in their ratings across different frequencies of **D** trials (β = -0.174, SE = 0.683, t(528) = -0.255, p = 0.800).

These trends are clearly illustrated in Figure 4. The left panel, displaying data from participants with lower depressive symptoms, shows a clear downward trend in ratings as the frequency of **D** trials increases. In contrast, the right panel, representing participants with higher depressive symptoms, depicts flatter lines, indicating that their ratings did not decrease significantly as **D** trials increased. This suggests that individuals with higher levels of depressive symptoms processed **D** trials less.

The Effect of D Trial Content.

The main effect of **D** trial content on perceived contingency between the two target stimuli was not significant. However, there was an interaction effect with block order wherein experiencing the block of conditions with empty **D** trials first resulted in reduced perceived

contingency when the **D** trials contained non-target stimuli (β = -2.404, SE = 0.928, t(38.902) = -2.589, p = 0.013). Pairwise comparisons confirmed that only when participants first experienced conditions with empty **D** trials was there a significant difference in ratings between empty **D** trials and those containing non-target stimuli, with the latter leading to lower ratings (β = 1.667, SE = 0.715, t(528) = 2.331, p = 0.020). This suggests a priming effect where initial exposure influences subsequent processing.

Discussion

The findings from Experiment 1 provide important insights into how individuals perceive contingencies between target stimuli and the effects of depressive symptoms on this learning process. Overall, the results partially support our hypotheses, while also highlighting unexpected trends that merit further investigation.

First, as hypothesised, participants were generally able to differentiate between positive and negative contingencies, with higher ratings assigned to positive contingencies and lower ratings to negative contingencies. This sensitivity to the programmed relationships indicates that the participants successfully learned the contingency structures across varying baselines, which aligns with previous research (Allan & Jenkins, 1980; Chatlosh et al., 1985; Wasserman et al., 1993).

And indeed, participants with higher levels of depressive symptoms demonstrated poorer discrimination between these contingencies. This is consistent with findings that depressed individuals perceive reduced action-outcome contingencies (Ackermann & DeRubeis, 1991; Alloy & Abramson, 1982; Castiello et al., 2020; Msetfi et al., 2007). The current results extend this understanding, suggesting that reduced contingency learning in depression may apply more broadly beyond action-outcome relations. Moreover, we found that this tendency of reduced perceived contingency was particularly weaker regarding negative

contingencies, which corroborates Msetfi et al. (2007), the only other study to show this effect (but in action-outcome contingencies).

The second key finding, concerning the frequency of **D** trials, challenged our initial hypothesis. We predicted that increasing **D** trial frequency would enhance contingency ratings, in line with associative learning models that suggest context-only events should strengthen target stimulus-outcome associations by reducing the context's interference. However, the results indicated the opposite effect: as **D** trial frequency increased, perceived contingencies decreased. This finding contradicts previous studies where **D** trials were found to have either no impact or a positive effect on perceived contingencies (Castiello et al., 2022; Murphy et al., 2021). Our result may be due to increased cognitive load, but even in studies where the cognitive load could be argued to be higher, such as Murphy et al. (2021) there were no effects of **D** trials rather than them decreasing perceived contingency. This warrants further investigation. Furthermore, the lack of response to **D** trial frequency among participants with higher depressive symptoms reinforces previous findings that individuals with depression process inter-trial intervals (likened to **D** events) less effectively (Msetfi et al., 2007).

The third key finding concerns the content of **D** trials. In line with our hypothesis, non-target stimuli in **D** trials did enhance the effect of **D** trials, but since the effect of **D** trials was to reduce perceived contingency, it further reduced perceived contingency. But this effect only occurred when participants first experienced conditions with empty **D** trials and then conditions with **D** trials with non-target stimuli in them. This priming effect suggests that initial exposure to empty trials heightened sensitivity to subsequent variations, possibly because the absence of information in the empty trials made participants more attuned to any additional stimuli in later trials. This finding underscores the importance of temporal order in learning, as early experiences can shape subsequent perceptions and processing strategies.

In the next experiment, given the unexpected reduction in ratings as the number of **D** trials increased, we increased the maximum frequency of **D** trials to 216 to further explore the impact of **D** trial frequency on perceived contingencies, following the highest level of frequency manipulation conducted by Castiello et al. (2022) in their second experiment, where a higher **D** trial frequency led to an increase in contingency ratings. Additionally, we examined the baseline contingency of zero, a condition often examined in studies of depressive realism, to assess whether the findings from this experiment generalise to non-contingent scenarios.

Experiment 2

Participants

57 participants were recruited from MTurk. Adhering to the same inclusion and exclusion criteria as in Experiment 1, two participants were excluded. The final sample comprised of 55 participants in total, with 32 males, 21 females, and 2 participants who preferred not to reveal their gender, with a mean age of 37.63 (SD = 6.63).

Participants completed the BDI-II as a measure of depressive symptoms. The mean BDI-II score among participants was 12.055 (SD = 14.113), with scores ranging from 0 to 47. According to BDI-II categorisation (Beck et al., 1996), 65.5% of participants fell into the minimal depression range (scores of 0–13; N=36), 16.5% into the mild-to-moderate range (scores of 14–28; N=8), and 20% into the severe range (scores of 29–63; N=11).

Task Design

A within-subjects design was adopted, with 5 experimental conditions that mapped onto a combination of the two experimental manipulations (see Table 3): 2 levels of **D** trial content (empty and non-target) x 2 levels of **D** trial frequency (16 and 216)] + 1 **D** trial frequency 0. There was only one baseline contingency, which was zero, formed by 16 **A**, **B**, **C**, and **D** trials.

Details of the number of **A**, **B**, **C**, and **D** trials in each condition and the respective ΔP values are provided in Supplementary Material 2B.

Table 3.
*Task Design for Experiment 2***.**

Baseline Contingencies	**D Trial Content**	
	Empty	Non-Target
	0	
Zero	16	16
	216	216

Notes. The numbers in the cells (0, 16, 216) refer to the levels of **D** trial frequency manipulation. When there are no (0) **D** trials, there is no content manipulation as there are no **D** trials to manipulate the content of, hence its placement in the middle of the two columns in the table.

Data Analysis

Data was analysed using the same approach as in Experiment 1. A backward stepwise algorithm was applied onto the full LMM, and post-hoc analyses were conducted on the resulting model to further investigate significant effects. For the purposes of post-hoc analyses, participants were grouped into "low depressive symptoms" (N=36; 65.5%) and "high depressive symptoms" (N=19; 34.5%) based on established cut-offs in non-clinical populations, where a score of 13 or below indicates minimal to no depressive symptoms, and a score above 13 suggests the presence of higher depressive symptoms (Beck et al., 1996; von Glischinski et al., 2019).

Results

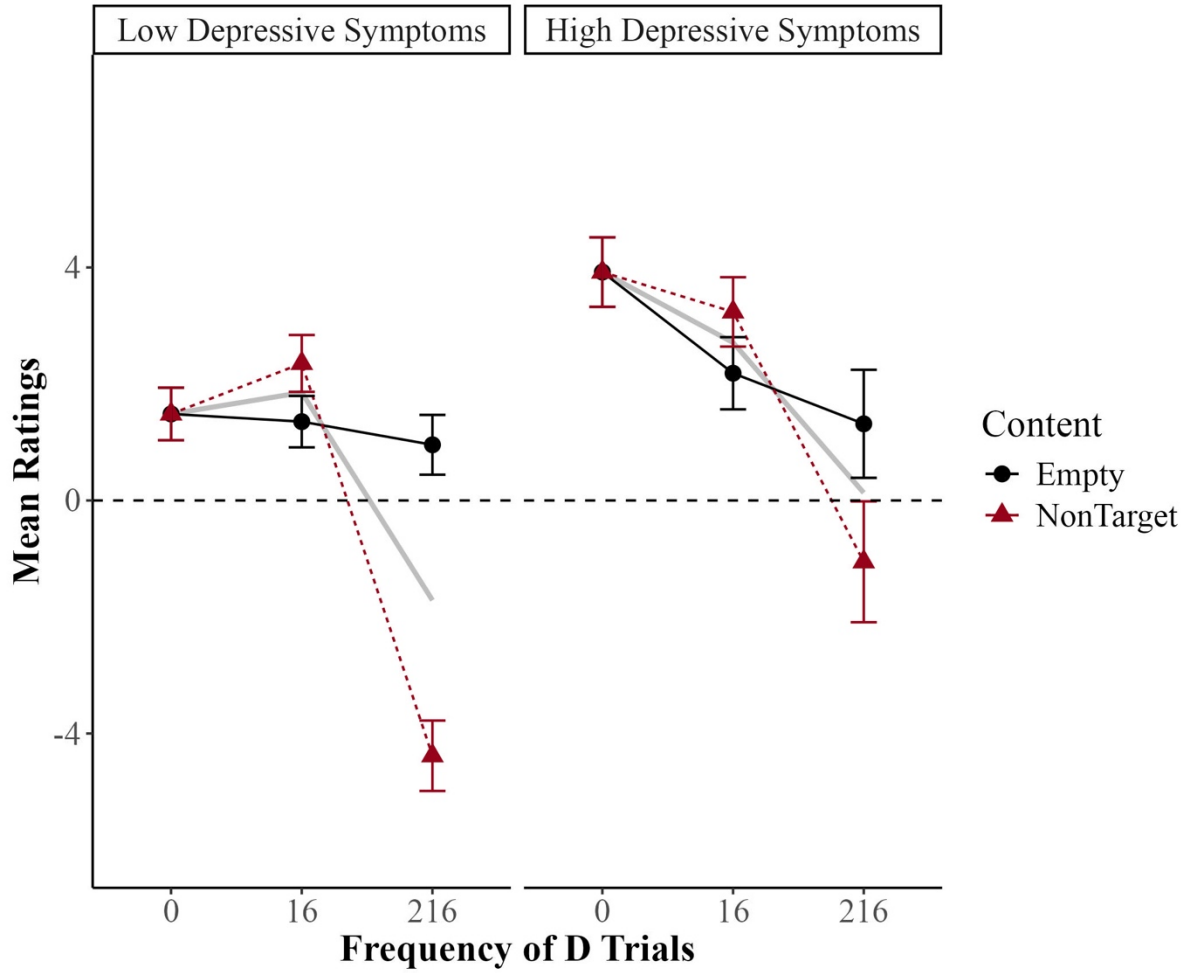
The findings from Experiment 2 align closely with those of Experiment 1, particularly in demonstrating that increasing the frequency of **D** trials reduces perceived contingency between target stimuli. This effect was enhanced when **D** trials contained non-target stimuli, further reducing perceived contingency. As in Experiment 1, participants with higher depressive symptoms processed **D** trials less effectively, which led to ratings regressing towards zero across conditions.

The final model, obtained using a backward stepwise algorithm, retained the main effect of block order ($F(2, 53) = 5.586$, $p = 0.006$) and a three-way interaction between D trial frequency, content, and BDI-II scores ($F(1, 477) = 4.969$, $p = 0.026$). The final model, confirmed through comparison with nested models, was adjusted using robust SEs to ensure that the conclusions were reliable despite violations of model assumptions (see Supplementary Material 3B). However, the three-way interaction did not remain significant after adjusting for robust SEs. Nevertheless, two-way interactions between these variables persisted and are examined in detail below.

The Effect of D Trial Frequency

As observed in Experiment 1, increasing the frequency of **D** trials led to more negative ratings overall ($\beta = -4.263$, SE = 0.569, $t(34.733) = -7.491$, $p < 0.001$). However, participants with higher depressive symptom levels exhibited a less pronounced reduction in ratings, with marginal significance ($\beta = 0.059$, SE = 0.031, $t(21.496) = 1.912$, $p = 0.060$). This trend suggests that while participants with higher depressive symptoms did show reduced processing of **D** trials, this reduction was less evident compared to Experiment 1.

Figure 5.
Results from Experiment 2: Ratings as a Function of D Trial Frequency, D Trial Content, and Depressive Symptom Levels.



Notes. The figure shows mean ratings (y-axis), reflecting perceived contingency between two target stimuli, as a function of the frequency of **D** trials (x-axis) across low and high depressive symptom levels as measured by Beck's Depression Inventory-II (BDI-II). The baseline contingency is zero. The data are further split by the content of the **D** trials, which were either empty (black round points) or contained non-target stimuli (red triangle points). The grey lines depict the ratings as a function of **D** trial frequency, collapsed over the content. Error bars represent the standard error of the mean (SEM).

Pairwise comparisons further revealed significant decreases in perceived contingency ratings with increasing **D** trial frequency for both low depressive symptoms (β = -3.57, SE = 0.469, $t(483) = -7.602$, $p < 0.001$) and high depressive symptoms groups (β = -2.58, SE = 0.683, $t(483) = -4.109$, $p < 0.001$). However, the lower β estimates and t-values for the high depressive

symptoms group indicated a diminished effect on perceived contingency ratings in response to increased **D** trial frequency.

These patterns are illustrated in Figure 5, where the left panel, representing participants with low depressive symptoms, shows the grey solid line having a sharper decline in ratings as **D** trial frequency increases. In contrast, the right panel, representing high depressive symptoms, shows a similar but less steep decline.

The Effect of D Trial Content

D trial content had a significant main effect on perceived contingency. Ratings were lower when **D** trials contained non-target stimuli (β = 1.455, SE = 0.678, t(34.733) = 2.147, p = 0.039) compared to when **D** trials were empty (β = 2.242, SE = 0.717, t(34.733) = 3.125, p = 0.004).

This effect interacted with depressive symptom severity: participants with lower depressive symptoms provided lower ratings for both empty **D** trials (β = -0.069, SE = 0.031, t(21.496) = -2.262, p = 0.034) and **D** trials with non-target stimuli (β = -0.054, SE = 0.026, t(21.496) = -2.127, p = 0.045). Pairwise comparisons revealed there was a significant difference in these ratings (β = 2.169, SE = 0.469, t(483) = 4.624, p < 0.001). In contrast, ratings from participants with higher depressive symptoms did not exhibit a significant difference between **D** trials that were empty and with non-target stimuli conditions (p = 0.295).

In Figure 5, the left panel shows a larger difference between the ratings for empty **D** trials (black solid line) and **D** trials with non-target stimuli (red dotted line) among participants with low depressive symptoms. This difference is less pronounced in the right panel, indicating that participants with higher depressive symptoms were less sensitive to variations in **D** trial content.

The Effect of Block Order.

Block order had a significant main effect, such that participants who first experienced the block with empty **D** trials provided lower ratings overall (β = -2.661, SE = 0.819, t(29.222) = -3.249, p = 0.003). This finding aligns with the findings from Experiment 1, where participants exposed to empty **D** trials first were more likely to give lower ratings when later encountering **D** with non-target stimuli trials. The exposure to empty **D** trials may have primed participants to lower their ratings as they subsequently processed the non-target stimuli.

Discussion

The results of Experiment 2 reinforce key findings from Experiment 1. As in Experiment 1 (and contrasting with our hypothesis), increasing the frequency of **D** trials reduced perceived contingency between the two target stimuli, with this effect being less pronounced in participants with high depressive symptoms. Moreover, **D** trials containing non-target stimuli further amplified the reduction in perceived contingency, supporting the hypothesis that **D** trials with non-target stimuli enhance the **D** trial frequency effect. This effect was attenuated in participants with higher depressive symptoms, suggesting they process these **D** trials less effectively.

Additionally, the block order effect observed in Experiment 2, where participants who first encountered empty **D** trials gave lower ratings overall, mirrors the findings from Experiment 1. This supports the idea that initial exposure to empty trials primes participants to lower their ratings. This block order effect highlights the importance of temporal order in learning processes, suggesting that the sequence in which participants encounter different information influences their perceptions of contingency. It is important to consider this effect when designing future studies to ensure that trial orders are carefully balanced.

However, investigating the effects of **D** trials in zero contingencies alone may have been limiting because without the context of positive and negative contingencies, participants'

internal scales for rating contingencies might have been skewed or misrepresented. In other words, the lack of comparison with positive and negative contingencies may have led to an inaccurate internal calibration of what zero contingency truly represents. This could have resulted in ratings that did not accurately reflect participants' perceptions of contingency but rather reflect a biased or misaligned interpretation due to the absence of a broader context. For example, as seen in Figure 5, all participants generally provide ratings above zero across all conditions, except when there were 216 **D** trials containing non-target stimuli when the ratings were below zero – i.e., the ratings were not at zero.

Additionally, the findings from the current experiment, concerning zero contingencies, are currently being interpreted independently of the positive and negative contingencies examined in Experiment 1. This between-subjects approach could be a limitation, as it assumes that participants' perceptions of zero, positive, and negative contingencies are directly comparable, despite the fact that these contingencies were not experienced within the same experimental context by the same participants. For instance, in Experiment 1, the effect of **D** trials with non-target stimuli was only stronger when participants first experienced empty **D** trials. However, in this experiment, the stronger effects of non-target stimuli were observed regardless of the order in which conditions were experienced. This difference suggests that the results may not be fully comparable across experiments, limiting our ability to draw clear conclusions about how **D** trials affect contingency perception in different contexts.

In conclusion, Experiment 2 extended the findings of Experiment 1 by demonstrating that the impact of **D** trials, both in terms of frequency and content, is modulated by depressive symptoms. The block order effect, particularly the priming influence of empty **D** trials, further reinforces the role of temporal factors in contingency learning. However, the between-subjects design limits the comparability of the results across experiments, necessitating further investigation. Experiment 3 will address this limitation by employing a within-subjects design

with consistent **D** trial manipulations across all baseline contingencies, allowing for a more robust analysis of how **D** trials influence contingency learning, particularly in relation to depressive symptoms.

Experiment 3

Participants

103 participants were recruited from MTurk. The increase in sample size was to increase power. Based on the exclusion criteria established in the first experiment, three participants were excluded. The final sample comprised of 100 participants in total, with 62 males, 38 females, with a mean age of 39.05 (SD = 5.79).

Participants completed the BDI-II as a measure of depressive symptoms. The mean BDI-II score among participants was 10.340 (SD = 13.156), with scores ranging from 0 to 54. According to BDI-II categorisation (Beck et al., 1996), 69.0% of participants fell into the minimal depression range (scores of 0–13; N=71), 17.5% into the mild-to-moderate range (scores of 14–28; N=18), and 13.5% into the severe range (scores of 29–63; N=14).

Task Design

A within-subjects design was adopted, with 15 experimental conditions that mapped onto a combination of the three experimental manipulations (see Table 4): [3 levels of Contingencies (negative, zero, and positive) x 2 levels of **D** trial content (empty and non-target) x 2 levels of **D** trial frequency (16 and 216)] + 3 **D** trial frequency 0, one for each baseline contingency manipulation as we could not manipulate the content when there are no **D** trials presented. Details of the number of **A**, **B**, **C**, and **D** trials in each condition and the respective objective contingency (ΔP) values are provided in Supplementary Material 2C.

Table 4.
Task Design for Experiment 3.

Baseline Contingencies	D Trial Content	
	Empty	Non-Target
Negative	0	
	16	16
	216	216
Zero	0	
	16	16
	216	216
Positive	0	
	16	16
	216	216

Notes. The numbers in the cells (0, 16, 216) refer to the levels of **D** trial frequency manipulation. When there are no (0) **D** trials, there is no content manipulation as there are no **D** trials to manipulate the content of, hence its placement in the middle of the two columns in the table.

Data Analysis

Data was analysed using the same approach as in Experiment 1 and 2. A backward stepwise algorithm was applied onto the full LMM, and post-hoc analyses were conducted on the resulting model to further investigate significant effects. For the purposes of post-hoc analyses, participants were grouped into "low depressive symptoms" (N=71; 69.0%) and "high depressive symptoms" (N=32; 31.0%) based on established cut-offs in non-clinical populations, where a score of 13 or below indicates minimal to no depressive symptoms, and a score above 13 suggests the presence of higher depressive symptoms (Beck et al., 1996; von Glischinski et al., 2019).

Results

The findings from Experiment 3 are consistent with the results of the previous two experiments. Participants successfully distinguished between the three baseline contingencies

(positive, zero, and negative), with stronger differentiation between negative and zero contingencies than between zero and positive contingencies. This distinction was particularly enhanced in participants with lower depression levels when **D** trials contained non-target stimuli. As in the previous experiments, increasing the frequency of **D** trials reduced perceived contingency between the target stimuli, an effect that was more pronounced in participants with lower depression levels. Unique to this experiment, this frequency effect was further amplified when participants experienced the block of **D** trials containing non-target stimuli first. Conversely, participants with higher levels of depression showed a diminished frequency effect.

The backward stepwise algorithm identified three significant three-way interactions with BDI-II scores in the final model: baseline contingency, content, and BDI-II score ($F(2, 1700) = 6.429$, $p = 0.002$); baseline contingency, block order, and BDI-II score ($F(4, 1700) = 4.108$, $p = 0.003$); and D trial frequency, block order, and BDI-II score ($F(2, 1700) = 4.747$, $p = 0.009$). The final model, confirmed through comparison with nested models, was adjusted using robust SEs to ensure that the conclusions were reliable despite violations of model assumptions (see Supplementary Material 3C). While the interaction between baseline contingency, block order, and BDI-II score did not remain significant after adjusting for robust SEs, the other three-way interactions persisted and are discussed below.

The Effect of Baseline Contingency.

In line with Experiment 1, participants successfully differentiated between the three baseline contingencies: ratings were highest for positive contingencies, lower for zero contingencies ($\beta = -4.008$, SE = 0.788, $t(30.788) = -5.089$, $p < 0.001$), and lowest for negative contingencies ($\beta = -7.702$, SE = 1.139, $t(30.788) = -6.765$, $p < 0.001$). This pattern is shown by the clear downward slopes in Figure 6. Notably, the distinction between negative and zero contingencies ($\beta = 3.60$, SE = 0.247, $t(1731) = 14.546$, $p < 0.001$) was stronger than the

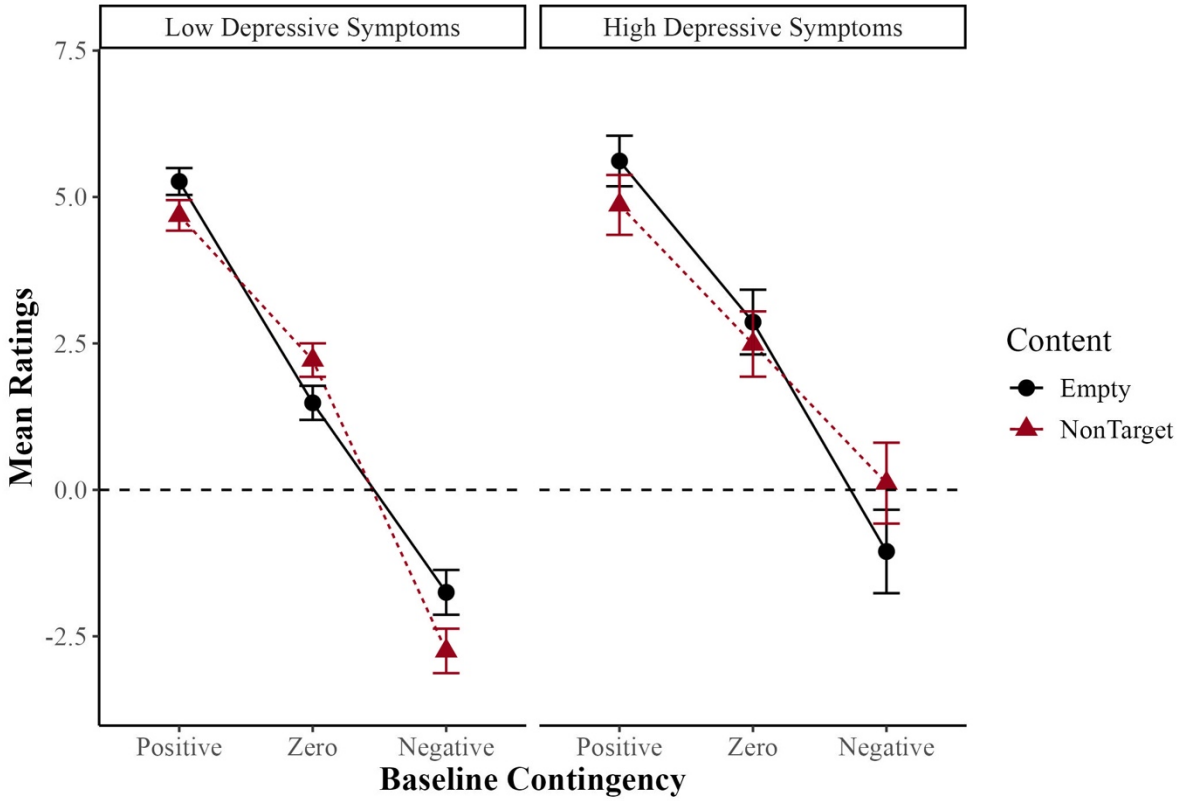
distinction between positive and zero contingencies ($\beta = 2.91$, SE = 0.247, $t(1731) = 11.754$, $p < 0.001$) as indicated by the β estimates and t-test values.

However, the effect of baseline contingency was moderated by both **D** trial content and depressive symptom levels. Although participants across both depressive symptom groups could distinguish between the three contingencies, those with higher depressive symptoms processed negative contingencies less effectively. This was evident in the larger difference in ratings between negative and zero contingencies for participants with lower depressive symptoms ($\beta = 4.06$, SE = 0.274, $t(1731) = 14.830$, $p < 0.001$) compared to those with higher depression levels ($\beta = 3.14$, SE = 0.412, $t(1731) = 7.617$, $p < 0.001$); see the β estimates and t-test values. The distinction between zero and positive contingencies, however, was similar across both groups (lower depression group: $\beta = 3.13$, SE = 0.274, $t(1731) = 11.430$, $p < 0.001$; higher depression group: $\beta = 2.69$, SE = 0. 412, $t(1731) = 6.522$, $p < 0.001$). These findings are reflected in Figure 6, where the ratings are similar for positive baseline contingency between the low (left panel) and high depressive symptom groups (right panel), but higher for negative baseline contingency for participants with high depressive symptom severity. These findings indicate reduced processing of negative contingencies for participants with higher depression levels.

Furthermore, the ability to distinguish baseline contingencies was weaker in participants with higher depression levels when **D** trials contained non-target stimuli ($\beta = 0.093$, SE = 0.030, $t(26.011) = 3.135$, $p = 0.004$). As shown in Figure 6, in the right panel (representing participants with higher depressive symptoms), the difference in ratings across contingencies is less pronounced when **D** trials contained non-target stimuli (less steep red dotted line) compared to when D trials were empty (steeper black solid line). This suggests that non-target stimuli in D trials may have enhanced participants' ability to differentiate contingencies, but this effect was diminished for those with higher depressive symptoms.

Overall, participants were generally able to distinguish between negative, zero, and positive contingencies, but this ability was enhanced in participants with lower depression levels, particularly when **D** trials contained non-target stimuli. Conversely, participants with higher depression levels exhibited weaker differentiation, especially when **D** trials contained non-target stimuli.

Figure 6.
Results from Experiment 3: Ratings as a Function of Baseline Contingencies, D Trial Content, and Depressive Symptom Levels.



Notes. The figure shows mean ratings (y-axis), reflecting perceived contingency between two target stimuli, as a function of the three baseline contingencies: positive, zero, and negative (x-axis) across low and high depressive symptom levels as measured by Beck's Depression Inventory-II (BDI-II). The data are further split by the content of the **D** trials, which were either empty (black round points) or contained non-target stimuli (red triangle points). Error bars represent the standard error of the mean (SEM).

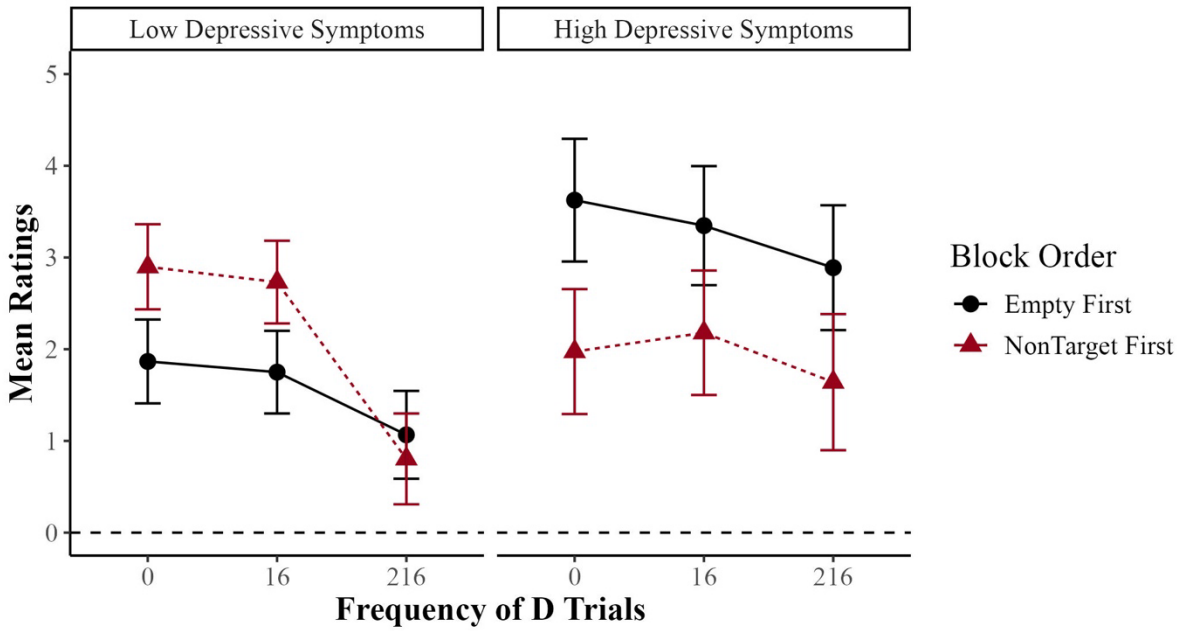
The Effect of D Trial Frequency.

In line with the findings from Experiments 1 and 2, increasing the frequency of **D** trials led to a reduction in perceived contingency (β = -0.478, SE = 0.224, t(21.489) = -2.129, p = 0.045). This downward trend is clearly illustrated in Figure 7.

A unique finding in Experiment 3 was the significant interaction between **D** trial frequency and block order (β = -0.775, SE = 0.330, t(43.940) = -2.349, p = 0.023). Pairwise comparisons revealed that the frequency effect was only significant when the block of **D** trials containing non-target stimuli was experienced first (β = -1.213, SE = 0.382, t(1178) = -3.181, p = 0.004). When the block of empty **D** trials was experienced first, the frequency effect was not significant (p = 0.135).

This two-way interaction further interacted with depressive symptom severity (β = 0.074, SE = 0.032, t(9.265) = 2.349, p = 0.043). Pairwise comparisons revealed that the **D** trial frequency effect, significant when **D** trials with non-target stimuli were experienced first, was particularly pronounced in participants with lower depressive symptom severity (β = -2.094, SE = 0.459, t(1178) = -4.567, $p < 0.001$). As shown in Figure 7, the right panel (representing participants with high depressive symptoms) shows flat lines, indicating little change in ratings with increasing **D** trial frequency. In contrast, the left panel (representing participants with low depressive symptoms) shows a sharp decrease in ratings when participants experienced **D** trials with non-target stimuli first (red dotted line) compared to empty **D** trials first (black solid line, which is relatively flat). This suggests that experiencing **D** trials with non-target stimuli first enhances the impact of **D** trials on reducing perceived contingencies in participants with lower depression levels.

Figure 7.
Results from Experiment 3: Ratings as a Function of D Trial Frequency, Block Order, and Depressive Symptom Levels.



Notes. The figure shows mean ratings (y-axis), reflecting perceived contingency between two target stimuli, as a function of the frequency of **D** trials (x-axis), and whether the first experienced block of conditions was with empty **D** trials (black round points) or **D** trials containing non-target stimuli (red triangle points). The data is split by participants with lower and higher depression levels as measured by Beck's Depression Inventory-II (BDI-II). Error bars represent the standard error of the mean (SEM).

General Discussion

The three experiments presented in this chapter provide valuable insights into how **D** events (i.e., the co-absence of target cues and outcomes) influence contingency learning and how non-target cues and outcomes impact perceived contingencies across individuals with varying depression symptoms. Across all experiments, increasing the number of **D** trials consistently decreased perceived contingency between the target cue and outcome, regardless of the baseline contingency. This effect was most pronounced when **D** trials contained non-target stimuli. Notably, individuals with higher depression symptoms demonstrated reduced sensitivity to **D** trials. However, when **D** trials involved non-target stimuli, participants with

higher depression symptoms exhibited some processing, albeit to a lesser extent, as indicated by a slight reduction in their perceived contingencies.

A consistent finding across all experiments was that participants rated contingencies in alignment with the programmed contingencies, supporting our hypotheses. Participants gave higher ratings to positive contingencies, lower ratings to zero contingencies, and the lowest ratings to negative contingencies. Individuals with higher depression symptoms, while still able to distinguish between contingencies, showed a diminished capacity to do so compared to those with lower symptom levels. This observation aligns with previous research indicating that depression is associated with a reduced perception of action-outcome contingencies (Ackermann & DeRubeis, 1991; Alloy & Abramson, 1982; Castiello et al., 2020; Msetfi et al., 2007). Our findings extend this understanding by suggesting that reduced contingency learning in depression applies beyond action-outcome relations to more general cue-outcome learning.

Moreover, participants were generally more sensitive to negative contingencies than positive ones, which aligns with previous findings (Castiello et al., 2022; Msetfi et al., 2007). However, a novel contribution of this study lies in the observation that participants with higher depression symptoms demonstrated weaker processing of negative contingencies specifically, suggesting a specific difficulty in dealing with inhibitory relationships. This reduced processing likely explains their diminished ability to distinguish between contingencies: while they perceived positive contingencies similarly to those with lower symptoms, they struggled more with zero contingencies and showed even greater difficulty in accurately perceiving negative contingencies. The reduced learning of negative contingencies observed here corroborates the findings of Msetfi et al. (2007), which showed similar results but in the context of action-outcome contingencies. This is the first study to extend this finding to cue-outcome relationships, suggesting a broader deficit in processing negative contingencies.

Unexpectedly, across all three experiments, increasing **D** trials consistently reduced perceived contingencies across all baseline contingencies, contradicting our hypothesis that **D** trials would enhance perceived contingency. Associative learning models predict that context-only events (such as **D** trials) should strengthen target stimulus-outcome associations by reducing contextual interference. However, the opposite effect occurred: as the frequency of **D** trials increased, perceived contingencies decreased, and particularly when non-target stimuli were present. This finding diverges from earlier studies that suggested **D** events either increased perceived positive contingency or had no significant effect (Castiello et al., 2022; Murphy et al., 2021).

However, this finding is partially in line with Arkes and Harkness (1983) as they observed that "some subjects considered Cell **D** data to strengthen judged contingency, some considered such data to weaken it, and others simply were undecided." (page 131). Schustack and Sternberg (1981) reported similar findings. Although, the rating scales used in these studies were unidirectional, making it difficult to distinguish between positive and negative control. In contrast, our experiments used a bidirectional scale, consistently revealing that **D** events informed negative contingencies. This consistency may be attributed to the bidirectional scale, which allowed participants to differentiate between positive and negative control more effectively.

Statistical and associative learning models do not account for the observed reduction in perceived contingencies due to **D** events, leading us to consider inferential reasoning as a possible explanation. The presence of **C** events, where the target outcome occurred without the target cue, may have led participants to attribute the outcome to a third, external cause (i.e., not the non-target cue). Consequently, during **D** events – where neither the target cue nor the target outcome was present – participants might have inferred that the external cause was also absent, associating the **D** events with the absence of that external cue-outcome relation. This inference

could have diminished their perceived contingency between the target cue and the outcome, as the presence of an alternative explanation (the external cause) may have reduced the perceived strength of the target cue-outcome relationship due to the principle of relative validity.

Another potential explanation for this reduction in perceived contingency as the number of **D** trials increased is the phenomenon of attentional disengagement, akin to the attentional blink effect (Nieuwenstein et al., 2005; Nieuwenstein et al., 2009; Ophir et al., 2020). In attentional blink paradigms, rapid re-engagement with task-relevant stimuli is challenging following a period of attentional disengagement. In our experiments, the empty **D** trials may have triggered attentional disengagement due to the lack of target stimuli, leading to reduced processing of subsequent target stimuli and a subsequent weakening of perceived contingencies. This corroborates with other attentional blink studies which showed that attention blinks when target stimuli are separated by a blank interval (Di Lollo et al., 2005; Nieuwenstein et al., 2009). When non-target stimuli were added to **D** trials, they may have been attended to be perceived as stimuli distinct to the target stimuli (Caputo & Guerra, 1998; Geng & DiQuattro, 2010), further exacerbating attentional disengagement. Attention to the non-target stimuli was also encouraged by the task design as we asked participants at the end of each trial stream whether they recalled a specific stimulus. This effect was more pronounced in individuals with higher depression symptoms, who may already have difficulties sustaining attention (American Psychiatric Association, 2013; Keller et al., 2019).

Under the attentional disengagement explanation, we would expect perceived contingencies in the baseline positive contingency conditions to regress towards zero, and for them to continue decreasing below zero in the baseline zero contingency conditions, which is consistent with our findings. However, this explanation would also predict that perceived contingencies in the baseline negative contingency conditions would regress towards zero (i.e., increase perceived contingencies), but this was not observed. Instead, perceived contingencies

in the negative condition lowered. This suggests a potentially complex interaction between attention and the processing of **D** events, particularly when different types of stimuli are involved. Future research could explore how various types of stimuli within **D** events influence contingency learning. For instance, would adding multiple non-target stimuli within **D** events further decrease perceived contingencies, or might these stimuli become so irrelevant to the task that participants can retain attention on target stimuli, thereby increasing perceived contingencies? These questions highlight the need for further investigation into the attentional processes underlying contingency learning in contexts involving non-target cues.

Another possible explanation for the observed decrease in contingency judgements as the number of **D** trials increased lies in the relative proportion of **A** trials. In the experiments, the number of **A**, **B**, and **C** trials remained constant while only the number of **D** trials was altered. As the number of **D** trials increased, the proportion of **A** trials—those in which the target cue was paired with the outcome—decreased relative to the total number of trials. **A** trials may have a greater influence on perceived contingencies due to their perceptual and logical salience: they involve observable events and clearly establish a predictive relationship between the target cue and the outcome. Previous research supports the heightened impact of such trial types on perceived contingencies (Kao & Wasserman, 1993; Perales & Shanks, 2007; Wasserman et al., 1990; Wasserman et al., 1996). Consequently, the reduced relative frequency of **A** trials as the number of **D** trials increase may weaken the overall perceived association between the target cue and the outcome. Thus, even though the absolute number of **A** trials remained unchanged, their diminished prominence within the trial block likely contributed to the observed reduction in perceived contingency.

Alternatively, the finding of the presence of non-target stimuli in **D** trials reducing perceived contingencies may be due to the associative competition between target and non-target stimuli. The relationship between the non-target cue and outcome may have been

perceived to be a strong positive relationship due to the exclusive presentation of co-presence of non-target cue and non-target outcome (**A** events). The presence of a strong positive relationship between non-target cues and outcomes may have thereby weakened the perceived contingency of the target cue-outcome pair. This aligns with previous findings showing that when multiple predictors are present, the stronger predictor can reduce the perceived contingency of the weaker one if their contingencies are in the same direction (i.e., both positive) (Baker et al., 1993; Spellman, 1996a; Tangen & Allan, 2004). In contrast, when predictors have opposing contingencies, the weaker predictor's contingency is enhanced in the opposite direction (Baker et al., 1993; Baker, Vallée-Tourangeau, et al., 2000; Darredeau et al., 2009). This associative competition might explain the reduction in perceived contingency across all baseline conditions when **D** trials contained non-target stimuli.

While associative learning models, such as the extended comparator hypothesis, have been used to account for these trends (Denniston et al., 2000), they do not fully explain the current study's results. This is because the extended comparator hypothesis proposes that a predictor's effectiveness results from a comparison between the representation of the outcome directly activated by the target predictor and the representation of the outcome indirectly activated by other predictors with which the target predictor has been associated in the past. However, in these experiments, that non-target stimuli were never paired with either of the target stimuli. Perhaps inferential reasoning models may offer a more comprehensive explanation, suggesting that participants compare relations between cues and outcomes even when those relations involve separate outcomes.

A potential avenue for future research involves the exploration of non-target stimuli that are not only strong positives but also strong negatives. In the current experiments, non-target stimuli generally represented positive relations. Investigating whether strong negative

non-target cues (i.e., inhibitory relationships) would enhance the perceived contingency of target stimuli could reveal interesting dynamics between competing predictors.

Regarding depressive symptom severity, participants with higher symptoms consistently showed diminished processing of **D** events, particularly in trials without non-target stimuli. When non-target stimuli were present, depressive participants exhibited some sensitivity to **D** events, but their perceived contingencies were still lower than those of non-depressed participants. This finding builds on previous research showing that individuals with depression process inter-trial intervals, akin to **D** events, less effectively (Msetfi et al., 2007). However, inter-trial intervals differ from **D** trials in that they occur regularly after each trial containing target stimuli and may appear for durations that vary from target trials, while **D** trials in our study were explicitly manipulated and interspersed with other trial types. Nevertheless, our study is the first to show this reduced processing in a paradigm that directly manipulates **D** events, suggesting a more generalised difficulty in processing absent cue-outcome relations among those with higher depressive symptoms.

Although, it is worth noting that individuals with higher depressive symptoms may process all four event types (**A**, **B**, **C**, and **D**) to a lesser extent, not just **D** events, as these participants exhibited reduced perceived contingencies across all conditions and contingencies. This suggests a more pervasive deficit in contingency learning that extends beyond **D** events. Future research should investigate this further by independently manipulating the frequency of each event type while holding the other three constant, to better understand how depressive symptom severity influences the processing of each event type.

Another important avenue for future research involves the potential priming effects of trial order. Our results suggest that the content of **D** events can have a priming effect on subsequent trials, influencing perceived contingencies. Exploring this effect could reveal individual differences in how trial sequences shape contingency learning.

One limitation of the current experiments is the uniform use of the same trial markers across all conditions. This consistency may have led participants to dissociate the context provided by trial markers from the actual stimuli, weakening the effect of the **D** event manipulation. Future research should explore varying the trial markers between conditions, which may reinforce the contextual relevance of **D** events and improve participants' contingency ratings.

Another limitation is in the sample used for these experiments. The distribution of depressive symptom severity (as measured by BDI-II scores) was skewed towards the lower end, with more participants exhibiting lower depressive symptoms than higher symptoms. Future studies would benefit from a more balanced sample, with participants evenly distributed across the four categories of depressive severity as defined by Beck's Depression Inventory (Beck et al., 1996). This would allow for more robust comparisons across the spectrum of depressive symptoms, offering clearer insights into how varying levels of depression influence contingency learning. Ensuring balanced groups could help disentangle whether the effects observed here are unique to participants with higher depression symptoms or reflect more general trends.

With these limitations and future directions in mind, this study provides a foundational understanding of how non-target stimuli and **D** events influence contingency learning, particularly in individuals with varying depressive symptoms. Future research can build on these findings by expanding the scope of causal models tested, balancing participant samples, and examining how different forms of non-target stimuli might enhance or reduce the perception of target contingencies.

The results from these experiments, particularly the influence of non-target stimuli on perceived contingencies for all participants, and the weaker learning of negative contingencies and **D** events in individuals with higher depression symptoms, provide a foundation for

examining how action-outcome contingencies are learned and how they influence perceptions of agency across varying levels of depressive symptoms. These findings directly inform the next set of experiments in **Chapter 6** (Experiments 4-6), which will investigate how individuals with different levels of depressive symptoms, including clinically depressed individuals, learn action-outcome contingencies in a multi-agent context.

Given the relevance of non-target stimuli reducing perceived target cue contingencies in these cue-outcome experiments, the upcoming experiments will specifically examine how the presence of other agents affects perceptions of action-outcome contingencies. We aim to explore whether the presence of another agent influences perceptions of one's own control over outcomes and vice versa, reflecting the findings from Experiments 1-3, where non-target cues led to a reduction in the perception of target contingencies.

In this next phase of the thesis, we will focus not only on the perceptual processes but also on the active processes involved in action-outcome contingency learning. Understanding how individuals interact with their environment to generate the data required for contingency learning will be key in the upcoming experiments. This will allow us to examine whether individuals with varying levels of depressive symptoms produce the four different event types related to action-outcome contingencies differently, and whether this influences their perceptions of control.

The broader aim is to determine how individuals perceive their agency and the accuracy of this perception, particularly in those with higher depression symptoms. The results of these forthcoming experiments will offer critical insights into how strategies can be developed to foster more accurate action-outcome contingency learning, which could ultimately improve agency judgements in individuals, especially those with depression.

Chapter 5:

Explaining Actions

In the preceding chapters, we examined the perceptual processes underlying agency perception, focusing on how statistical and associative learning theories explain the perception of action-outcome contingencies. These models largely treat the learner as a passive recipient of information, with actions conducted to attain specific outcomes. For example, the Rescorla-Wagner model suggests that the associative strengths between actions and outcome determine which action is selected – the one with the highest associative strength with the outcome.

Empirical evidence supports these models, showing that participants tend to act more frequently when the contingency between their action and the outcome is positive, less when it is zero, and least when it is negative (Chatlosh et al., 1985; Shanks & Dickinson, 1991; Wasserman, 1990; Wasserman, Neunaber, et al., 1983). A study by Karsh and Eitam (2015) further illustrates this point. Participants were prompted to press one of four keys as randomly as possible, and each key had a different probability of producing an outcome. Despite the instruction to press randomly, participants were more likely to press the keys with a higher probability of an outcome and did so more quickly, suggesting that they were motivated by the likelihood of achieving an outcome. This finding supports the view that action selection is influenced by perceived control, in line with statistical and associative learning models, which emphasise that actions are carried out to achieve outcomes.

Further exploration of this relationship between actions taken and the perceived action-outcome contingency has revealed the role of actions as a mediating factor in the learning process. Researchers found that directly manipulating participants' probability of responding, through experimental instructions (such as "press the button approximately half the time") or trial-by-trial command to act or not, influenced their perceived contingencies (Blanco et al., 2011; Hannah & Beneteau, 2009; Matute, 1996). Even when the outcome was not contingent on the participants' actions (zero contingency), higher probability of actions was aligned with higher perceived contingencies.

Moreover, Blanco et al. (2012) investigated whether the depressive realism effect – where depressed individuals more accurately assess lack of control – was mediated by the probability of actions rather than being a direct result of mood. By manipulating participants' probability of actions through instructions, they observed that this intervention cancelled the effect of mood on control judgements – the probability of actions aligned with perceived contingencies. This finding suggests that the more accurate perception of their lack of control observed in depressed participants is mediated by behavioural factors. This suggests that therapists might focus on modifying behavioural patterns to influence patients' perceptions of control. This approach could offer a more direct treatment method than solely targeting mood, and may be a way to address the potential bias that the self has no control.

The fact that experimentally manipulated action frequencies influence perceived contingencies, even when participants have no control over their action rates, suggests a bidirectional and mutually influential relationship between action and perception in contingency learning: actions shape perceived contingencies, and these perceptions, in turn, influence future actions. Actions produce observations from which individuals can learn and perceive.

So, is the perspective that actions are conducted solely to obtain goals or rewards (exploitation) sufficient to fully explain how individuals act when learning action-outcome contingencies? While goal-directed actions are undoubtedly important, as evidenced by the alignment of action probability with perceived outcome contingency, this view fails to account for crucial aspects of human behaviour in the context of learning. We must consider the role of exploration in addition to exploitation (where actions are solely conducted to achieve immediate goals) to explain these aspects.

Exploration refers to actions taken to gather information about the environment, which may sometimes be at the expense of immediate goals (Kruschke, 2008; Sutton & Barto, 2018).

In novel situations where the true agentic structure is initially unknown, relying solely on exploitation to explain actions would fail to account for how individuals discover new, potentially more effective actions to produce outcomes. One might argue that in such situations, random actions could suffice to generate observations from which action-outcome contingencies could be learned, after which exploitation of the learned contingencies would occur (Daw et al., 2006; Sutton & Barto, 2018). However, this explanation falls short in several ways.

The idea of random action followed by learning does not account for the ongoing balance between exploration and exploitation observed in human behaviour. Even as individuals gain knowledge about their environment and the contingencies within it, they continue to engage in exploratory actions, indicated by choices that appear to be suboptimal for immediate reward maximisation (Kidd & Hayden, 2015; Lee et al., 2011; Wilson et al., 2014). This persistent exploration suggests a more complex decision-making process than simply switching from random actions to exploitation once sufficient information is gathered.

Moreover, studies indicate that exploration is often systematic rather than random. While some theories suggest that actions not sampled recently might be prioritised (Kakade & Dayan, 2002; Wittmann et al., 2008), more recent studies indicate that people tend to be even more strategic. Rather than acting randomly when faced with uncertainty, individuals often select actions they expect to be most informative (Gibson, 1988; Gureckis & Markant, 2012; Wilson et al., 2014).

This information-driven exploration is based on the idea that certain actions provide more valuable knowledge than others, and that information gain can be quantified using concepts from information theory (Carter, 2007). One approach is to calculate the expected reduction in uncertainty, commonly measured by entropy, that follows from performing a

particular action. Actions that are likely to reduce uncertainty the most are prioritised, as they provide more useful information.

From an information theory perspective, action-outcome contingency learning can be understood in terms of how much an action reduces uncertainty about an outcome (Gallistel & Balsam, 2014; Gallistel et al., 2014). For instance, entropy measures the uncertainty surrounding outcomes before an action is taken. When an action conveys information about the outcome, it reduces that uncertainty. This reduction is quantified as mutual information – the difference between the uncertainty of the outcome alone and the uncertainty once the action has been considered. Thus, forming an association between actions and outcomes relies on this mutual information, ultimately reducing the unpredictability of the outcome.

For example, imagine a scenario where you learn whether pressing a button turns on a lightbulb. A probability distribution would represent your internal (prior) belief about whether pressing the button (action) will turn the lightbulb on or whether doing nothing (no action) will have the same effect: p(x) = [p(action); p(no action)]. Entropy is calculated as: $H\big(p(x)\big) = -\sum_x p(x) \ln p(x)$ (Carter, 2007). If your belief is strongly biased, say p(x) = [0.99; 0.01], meaning you are almost certain that pressing the button will turn the light on, the entropy is low (≈ 0.05), reflecting high certainty. However, if you have no knowledge about the situation and your belief is p(x) = [0.5; 0.5], the entropy is much higher (≈ 0.69), indicating greater uncertainty. Note the shift from using simple value estimates, like those seen in statistical and associative learning models, to probability distributions that reflect degrees of uncertainty across beliefs. Instead of a definitive outcome prediction, these distributions account for varying levels of uncertainty in the world, allowing for more nuanced representation of how actions or inactions affect outcomes like the light turning on, and allowing for more complex set of beliefs and causal structures.

Moreover, researchers have argued that exploration plays a crucial role in hypothesis testing, which is key in understanding whether a correlative relationship is causal (Hohwy, 2016; Mill & Nagel, 2008; Pearl, 2000). Here, it is posited that when people explore their environment, they are often testing ideas about how things work. They do this by actively manipulating variables in the environment to gather evidence that supports or refutes their ideas about cause and effect (Knox et al., 2012; Nelson, 2005; Sobel & Kushnir, 2006). Through this systematic exploration, learners can go beyond simply noticing that two events often occur together (correlation) to understanding whether one event actually causes the other (causation). This process of hypothesis testing through exploration is more powerful than passive observation alone, as shown by studies that yoke the actions of one group of participants, who actively explore, with a second group that passively observes the same sequence of actions (Gureckis & Markant, 2012; MacDonald & Frank, 2016; Markant & Gureckis, 2014). It allows individuals to build a deeper understanding of the causal structure underlying action-outcome relationships.

Furthermore, in volatile situations where action-outcome contingencies may change over time, periodic exploratory actions are necessary to detect changes that might not be apparent from exploiting known contingencies. Researchers have shown that participants who completed such a decision-making task engaged in strategic exploration to track changes in causal structures, occasionally choosing options they believed to be suboptimal to check if the environment had changed (Eckstein et al., 2019; Engel, 1963; Schulz et al., 2019). This behaviour demonstrates the importance of exploration in maintaining an accurate understanding of one's control over outcomes in volatile environments, prioritising long-term understanding over short-term goal attainment.

These findings collectively suggest that to fully explain how people act when learning about the causal structure of their environment, we must consider both exploitation and

exploration. This exploration-exploitation trade-off is not a simple dichotomy but a continuum that individuals navigate dynamically. Hybrid models that incorporate both random and information-driven exploration may best capture human learning behaviour (Friston et al., 2011; Gershman, 2018; Parr et al., 2022; Wilson et al., 2014). These models recognise that while strategic actions often aim to reduce uncertainty, there remains an element of randomness, serving to balance exploitation with discovering unexpected opportunities.

Given the limitations of purely associative or statistical models in accounting for exploratory behaviour, Bayesian frameworks offer a more comprehensive approach. Unlike earlier models that provide point estimates of causal strength (e.g., ΔP, Rescorla-Wagner), Bayesian models maintain and update entire probability distributions over causal hypotheses, representing uncertainty explicitly and driving exploratory behaviour.

Now, I will introduce active inference, a Bayesian framework grounded in the Free Energy Principle, which integrates exploration and exploitation within a single imperative: minimising expected free energy (Friston, 2010; Friston et al., 2017; Parr et al., 2022). This unified account of perception, action, and learning offers a compelling explanation of adaptive behaviour. While other Bayesian models, such as Kalman filters and Bayesian reinforcement learning, also consider the exploration-exploitation trade-off (Ath et al., 2021; Kruschke, 2008), this thesis will focus on using active inference to explain the results of Experiments 4-6 in **Chapter 6**. These studies are the first to apply active inference to model fitting in this context, leaving model comparisons with other Bayesian models for future research.

Active Inference

Active inference, rooted in the Bayesian framework, operationalises the free energy principle (FEP), which suggests that agents perceive, learn, and act by minimising free energy – a quantity related to surprise, uncertainty, or entropy. Here, an agent forms beliefs about the world (perception), updates those beliefs when encountering new information (learning), and

chooses actions to make its internal model more accurate and achieve its goals. When learning how much agency they have over their environment, active inference agents continually refine their internal model by observing their actions and outcomes, while also acting to generate new observations that further inform their sense of agency.

The internal model is called the generative model (because it generates predictions), and includes the following main components: (i) the prior probability distribution, representing the agent's initial beliefs about aspects (called states) of the world (e.g., I tend to have positive control over this light bulb); (ii) the likelihood distribution, which reflects the agent's belief about the probability of observing specific outcomes, given certain states (e.g., if I press this button and have positive control, the light bulb will turn on); (iii) beliefs about how these states change over time (e.g., my ability to control the light bulb will remain consistent); and (iv) preferences for specific outcomes (e.g., wanting to see the light bulb turned on). The generative model is shaped by past experiences and represents the agent's beliefs about the causal structure and dynamics of the environment, including how their actions influence outcomes (Mathys et al., 2011; Piray & Daw, 2021).

The generative process refers to the way states and outcomes unfold in the true environment, independent of the agent's internal beliefs (the generative model). It reflects the actual causal structure of the environment the agent interacts with and generates the observations the agent encounters. By comparing these real-world observations from the generative process with the agent's predictions, it refines its generative model, enabling it to make better predictions and improve its understanding of how its actions influence the environment.

Perception involves forming beliefs about states of the environment, which are all hidden, i.e., they cannot be directly observed and must be inferred from sensory observations. For example, a hidden state might be whether the agent has control (agency) over the light

bulb. These states are not directly visible, so the agent must rely on its observations – such as seeing the light bulb turn on after pressing a button – to infer them. By using its generative model, the agent can compare its prior beliefs (e.g., I tend to have positive control over this light bulb) with the likelihood of its observations (e.g., the light bulb will turn on after I press the button if I have positive control) to infer the hidden state of agency (e.g., I have positive control).

Learning occurs when an agent updates its generative model, adjusting beliefs about hidden states and environmental dynamics as it gathers new observations. The gap between expected and observed outcomes leads to an updated understanding, known as the posterior distribution. For example, if an agent expects to control the light bulb but fails to turn it on after pressing the button, this reduces its certainty of agency (a "flatter" posterior distribution).

In active inference, agents minimise variational free energy (VFE) to refine their beliefs about the world. It is important to note that agents do not directly apply Bayes' rule in its exact form, as calculating the true posterior distribution would require determining the precise probability of every possible hidden state – a task that quickly becomes computationally intractable, especially in complex and noisy environments. Instead, variational methods from information theory are employed, where agents approximate the true posterior by fitting simpler, more manageable distributions that capture the key features of the true distribution. This approximation allows the agent to handle the complexity of the environment while minimising the computational cost.

VFE is composed of two main terms: Kullback-Leibler (KL) divergence and log evidence. KL divergence quantifies how much the agent's approximate beliefs (the variational posterior) differ from the true posterior. In other words, it captures how much an agent's beliefs have shifted after encountering new evidence compared to its prior beliefs. The log evidence measures how well the agent's generative model predicts the actual observations. By

minimising VFE, the agent reduces this divergence (increasing accuracy), refining its model and adjusting its understanding of the environment while being computationally efficient.

As we can see, perception and learning in active inference are intertwined, similar to traditional Bayesian methods, as the agent continually compares its predictions to observations, updating its beliefs but through the minimisation of VFE. This process would allow an agent to refine its sense of agency and maintain an adaptive understanding of its control over the environment.

Actions in active inference are cast as part of the inference process. Actions are inferred as part of the agent's goal to minimise free energy – more specifically, expected free energy (EFE). The agent evaluates potential action sequences (policies) and their expected consequences within the generative model, aiming to balance improving the model's predictive accuracy (exploration) and achieving preferred outcomes (exploitation).

Exploration involves seeking new information to reduce uncertainty about the environment. For example, if the agent is uncertain about its control over a light bulb, it might press and not press the button several times to gather evidence and refine its internal model. This process of evaluating actions can be understood as hypothesis-testing behaviour, where the agent entertains multiple hypotheses about its control and tests them through action and observation. These actions serve to reduce uncertainty about its agency and improve the accuracy of its generative model. In contrast, exploitation refers to actions that align with the agent's current beliefs and preferences. Once the agent infers that pressing the button reliably turns on the light, it will exploit this knowledge to achieve the preferred state of having the light on.

In summary, actions in active inference are framed as an ongoing process of refining the agent's generative model. EFE guides this process by helping the agent evaluate and select

actions that minimise uncertainty and achieve desired outcomes, ensuring a continuous adaptation of behaviour.

To contextualise this under agency learning, under the active inference framework, individuals learning whether and how much control they have over turning on a light bulb would observe their actions, such as pressing or not pressing a button, and infer that they have carried out these actions. They would also observe the outcome – whether the light turns on or off – and use these observations to infer their sense of agency. In a novel context, the individual would explore by testing hypotheses about their control. They might press the button multiple times to observe the effects, and refrain from pressing it at other times to evaluate what happens in the absence of action. By gathering these observations, they would update their generative model, refining their understanding of the relationship between their actions and the outcomes. Once a reliable pattern is inferred, they would exploit this knowledge to achieve their preferred outcome – having the light turn on when desired (Pitliya & Murphy, 2023).

Individuals with higher levels of depression, often associated with a reduced sense of agency, may behave differently under the active inference framework. Their internal generative model might be biased towards expecting limited or no control over outcomes, leading to reduced exploration or slower learning of agency. For example, even when pressing the button, they might not update their beliefs sufficiently, potentially perceiving their actions as ineffective or attributing the outcome to external factors. Additionally, they may perceive that others have more control over the environment, reinforcing their belief in their own lack of agency. As a result, they may be less likely to test hypotheses about their control (i.e., reduced exploration) and more likely to assume that their actions do not influence the environment, reinforcing a sense of helplessness or passivity. This might manifest as fewer attempts to engage with their surroundings or a reluctance to act, reflecting the reduced expectation of successful outcomes, characteristic of learned helplessness in depression.

Chapter 6:

Agency Learning in a Multi-Agent Context

(Experiments 4-6)

The aim of this thesis is to investigate how individuals learn the extent of agency they and others have over their environment – to understand what they control – as this understanding enables them to navigate their surroundings and engage in goal-directed, adaptive behaviour (Verschure et al., 2014). This investigation is conducted among individuals with varying levels of depressive symptoms, as these symptoms are associated with reduced perceptions of self-agency and an increased attribution of agency to others (Benassi et al., 1988; Castiello et al., 2020; Maier & Seligman, 1976). A key aspect of agency is learning how contingent outcomes are on one's actions (Moore et al., 2012).

Learning action-outcome contingencies involves recognising patterns of how often outcomes occur with and without actions (Allan, 1980; Marien et al., 2015). In theory, four types of information are available to guide this process: the co-occurrence of actions and outcomes (**A** events), the presence of either the action or outcome and the absence of the other (**B** and **C** events), and the co-absence of actions and outcomes (**D** events). Logically and behaviourally, **A** events clearly suggest a positive contingency, while **B** and **C** events indicate a negative one (Murphy et al., 2021). However, **D** events may or may not be processed as while they are logically informative for judging whether an action is necessary for the outcome, the absence of stimuli is less salient than its presence or enough information may be elicited by the other three event types (Castiello et al., 2022; Murphy et al., 2021). The processing of **D** events may be dependent on the goal (Mata et al., 2015), and so they may be helpful in distinguishing multiple predictor-outcome relationships, particularly when non-target stimuli are present, as the absence of the target stimuli might highlight the necessity of the predictor for the outcome.

Additionally, previous studies have shown that the presence of other predictors influences the perceived relationship between the target predictor and the outcome (Baker et al., 1993; Darredeau et al., 2009). However, these studies primarily examined shared outcomes,

leaving it uncertain whether a completely separate, non-target predictor-outcome relationship would also affect the learning and perception of the target contingency.

In the first set of experiments (Experiments 1-3), we explored how non-target stimuli and **D** events influence contingency learning among individuals with varying levels of depressive symptoms. We found that **D** events reduced perceived contingency, and the presence of non-target predictor and outcome (which had a strong positive contingency) further reduced the perceived contingency between the target predictor and the outcome. Participants with higher depressive symptoms showed less pronounced perceptions of contingencies, with weaker processing of negative contingencies and **D** events.

In this chapter, I report on Experiments 4-6, which investigate how individuals with different levels of depressive symptoms, including clinically depressed individuals, learn action-outcome contingencies in a multi-agent context. We aim to explore whether the presence of another agent influences perceptions of one's own control over outcomes, and vice versa, echoing the findings from Experiments 1-3, where the presence of a non-target relation diminished the perception of contingencies of the target relation. Note that the Discussion sections in this chapter are brief and the results are discussed in more detail in the next chapter.

These experiments will not only focus on perceptual processes but also on the active processes involved in action-outcome contingency learning. Understanding how individuals interact with their environment to generate the data necessary for learning will be key in this series of studies. This will allow us to examine whether individuals with varying levels of depressive symptoms produce the four types of events related to action-outcome contingencies differently, and how this influences their perceived sense of control.

We will fit the behavioural data to an active inference model to compare the model's predictions with actual human behaviour. This will enable us to examine how model parameters vary among individuals with and without clinical depression, offering insights into

how different people perceive agency within the same environment. These results could provide crucial insights into strategies for improving the accuracy of action-outcome contingency learning, which could ultimately enhance agency judgements in individuals, particularly those with depression.

Briefly, the task designed for these experiments involves participants exploring their control over an on-screen outcome through button presses across multiple trials, while simultaneously observing a simulated agent's button presses. After each block of trials, participants rate each agent's perceived control over the outcomes. The experimental conditions vary which agent has control, as well as the type and direction of that control (positive/excitatory, negative/inhibitory, or no control).

We hypothesise that all participants will correctly identify which agent has control, assigning higher ratings to the agent with control in the direction of their control. In the same experimental condition, the agent without control is expected to receive ratings suggesting some level of control in the direction opposite to that of the controlling agent, though these ratings would not match the magnitude of those given to the agent that genuinely has control. This pattern aligns with previous findings, including our Experiments 1-3 and other studies (Baker et al., 1993; Darredeau et al., 2009), which suggest that perceived control can partially shift towards a non-controlling agent when another agent exerts influence in the opposing direction; this will be termed the "opposing control" pattern.

We further hypothesise that participants with higher depressive symptoms will provide lower overall ratings of perceived control, though their rating patterns (i.e., opposing control pattern) will remain similar to those of participants with lower depressive symptoms. This expectation aligns with our previous findings (Experiments 1-3), which indicated that individuals with higher depressive symptoms tend to perceive weaker action-outcome contingencies, even when they are similarly influenced by non-target stimuli as participants

with lower symptoms. Additionally, we anticipate that they will show particularly weaker processing of negative contingencies, as observed in Experiments 1-3.

We also hypothesise that participants with lower depressive symptoms will engage more actively with their environment, with their actions becoming progressively more goal-oriented as they learn the causal structure (i.e., they will perform actions that consistently produce desired outcomes). In contrast, participants with higher depressive symptoms may show reduced interaction with their environment, resulting in less goal-directed behaviour. This difference is supported by research indicating that depressed individuals often exhibit diminished reward processing (Pitliya et al., 2022; Pizzagalli et al., 2009; Vrieze et al., 2013) and lower engagement with their environment (Blanco et al., 2009; Blanco et al., 2012). sampling their environment less (Blanco et al., 2012; Ferster, 1973; Lewinsohn et al., 1980). This reduced engagement means they sample their environment less, which could limit their production of action-outcome pairings (**A** events) that are crucial for establishing a sense of control (Matute, 1996). As a result, their lower levels of interaction may yield fewer opportunities to confirm control over outcomes, further diminishing their perception of agency.

Experiment 4

Participants

A total of 70 undergraduates were recruited from The University of Oxford, UK. Participants were required to use a desktop computer or laptop to complete the task, with tablets or phones not permitted. To ensure adequate attention to the task, participants who provided uniform responses across 90% or more of the task conditions were excluded from analysis, resulting in three exclusions. The 90% threshold was chosen based on its use in prior experiments (Experiments 1-3) reported in this thesis and was further supported by a visual inspection of the response distributions, which indicated that the excluded responses appeared as anomalies compared to the remaining data.

The final analysed sample consisted of 67 participants, including 23 males, 41 females, 2 non-binary individuals, and 1 participant who preferred not to disclose their gender. The overall mean age of the final sample was 20.4 years (SD of 2.7). The ethics committee at the University of Oxford approved this research (reference code: R74631/RE001).

Participants also completed the Beck Depression Inventory-II (BDI-II; Beck et al. (1996)), assessing the severity of their depressive symptoms. The mean BDI-II score among participants was 17.642 (SD = 12.130), with scores ranging from 2 to 47. According to BDI-II categorisation (Beck et al., 1996), approximately 41.8% of participants fell into the minimal range (scores of 0–13; N=28), 37.3% into the mild-to-moderate range (scores of 14–28; N=25), and 20.9% into the severe range (scores of 29–63; N=14).

Materials, Procedure, and Task Design

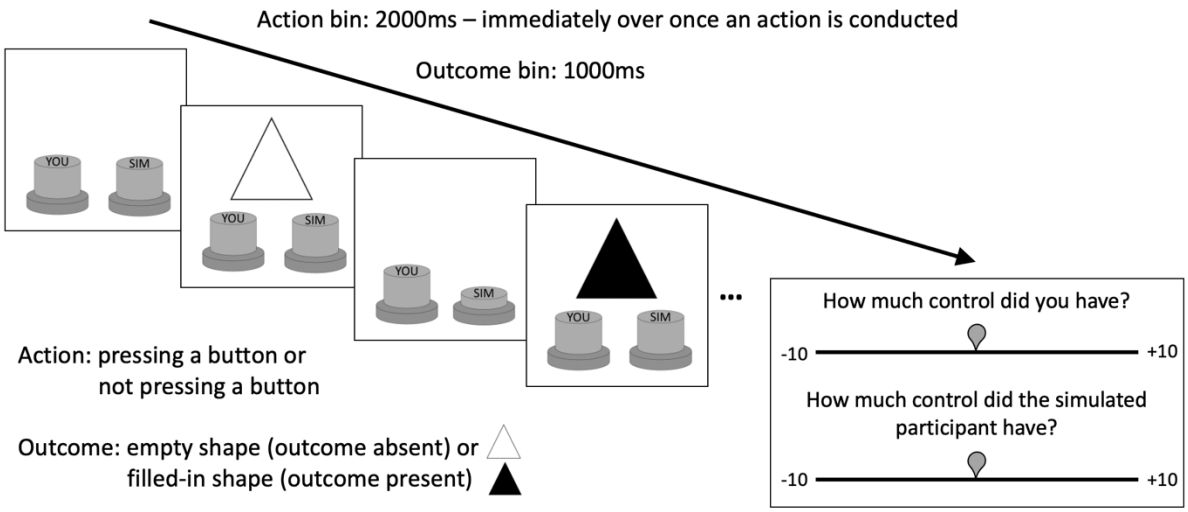
At the beginning of the experiment, participants were presented with a briefing sheet detailing the task, data usage and storage, contact information for queries, and their right to withdraw. They provided informed consent and demographic information (age and gender) and confirmed they did not have a history of seizures or diagnosed motor or cognitive impairment before completing the computer task.

Overall, in each block of trials, participants freely produced actions, observed another (simulated) participant's actions, integrated this information with the absence or presence of outcomes, and provided ratings about how much control both agents had, reflecting their learning of action-outcome contingencies (Figure 8). I will now go into the details.

We employed a free-operant task, where each trial was divided into two distinct phases: an action bin and an outcome bin. In the action bin, participants were presented with two buttons on the screen: one they could control (labelled "YOU") and one the other (simulated) participant could control (labelled "SIM"). The placement of the buttons (left or right) was counterbalanced between participants. When an agent pressed their respective button, it would

visually appear pressed to represent the recording of the action. Unknown to participants, the other participant randomly pressed the button 50% of the time across all experimental conditions. The participants themselves were free to press or not press the button as they wished. The action bin had a maximum duration of 2000 milliseconds but would end earlier if the participant pressed their respective button. This design choice was implemented to create a more seamless and responsive experience, enhancing the sense of real-time control for the participant.

Figure 8.
The Task for Experiments 4-6.



Notes. An illustration of the free-operant two-agent agency learning task completed by participants. Each trial consisted of a 2000ms maximum action bin and a 1000ms outcome bin. Participants explored their control over the outcome by pressing or not pressing the button labelled 'YOU' across 24 trials, while observing the simulated agent's button presses (labelled 'SIM'). The outcome, displayed in the outcome bin, was either an empty shape (no outcome) or a filled shape (outcome present). After the block of trials, participants rated their perceived control over the outcome for both themselves and the simulated agent on a scale from -10 to 10.

Following the action bin, the outcome bin lasted for a fixed duration of 1000 milliseconds. During this phase, the outcome was visually represented as a shape that was either empty (outline only) or filled (solid black), indicating the absence or presence of the outcome, respectively. As a result of this structure, the total duration of a trial could vary between 1 to 3 seconds, depending on when or if a button was pressed during the action bin. A shape was chosen as the outcome stimulus to provide a neutral, novel cue and minimise potential biases or prior beliefs from driving the learning.

Participants were instructed to aim to produce the filled shape as much as possible throughout the task and that they would be asked how much control the self and other agent had at the end of the block of trials. This goal-oriented instruction was designed to encourage active engagement with the task and promote learning about the action-outcome contingencies. Participants' actions were recorded throughout to evaluate how they sampled their environment and whether it aligned with perceived control.

After 24 such trials, which formed one block, participants rated their perceived level of control over the outcome for both themselves and the other participant. The rating was provided on a scale from -10 to 10, where 10 indicated a strong positive rating (pressing the button resulted in outcome presence, not pressing resulted in absence), -10 indicated a strong negative rating (pressing the button resulted in outcome absence, and vice versa), and 0 indicated no relationship between button pressing and outcome. The rating scale presented to participants visually displayed these values and their corresponding meanings to ensure clarity.

The independent variables were: (i) which agent had control (self or other, referring to the real participant or the simulated agent), (ii) the programmed contingency between actions and outcomes (positive, zero, or negative), and (iii) the degree of control (probabilistic at 75% or deterministic at 100%). This produced nine conditions: one control condition (Zero), where neither agent had control, and eight experimental conditions. These experimental conditions

included Self Positive, Self Negative, Other Positive, and Other Negative, each with two variations – probabilistic and deterministic. When one agent had control, the other had none. This setup was designed to elicit any interaction effects between the two agents - for example, do people tend to believe that when the self has positive agency, the other has negative agency when in fact the other has none?

Participants completed a total of ten blocks of trials, one practice block of Deterministic Self Positive, and then the nine experimental conditions. Participants were allowed a maximum of two five-minute breaks between blocks of trials. The whole experiment took around 20 minutes to complete. To prevent visual interference, the mouse cursor was removed from the screen during the contingency learning trials, and participants were informed of this change beforehand. The exact instructions participants were given for the task are detailed in Supplementary Material 4.

We used the Gorilla Experiment Builder (www.gorilla.sc) to create and host our experiment (Anwyl-Irvine et al., 2020). The stimuli were created using Microsoft PowerPoint for Mac Version 16.45. The outcomes were shapes selected from the collection available on PowerPoint. The buttons were created using the shapes on PowerPoint, with text overlaid to label which agent the button belonged to. All slides from the PowerPoint were exported as .png files to be used as stimuli. The dimensions for each image were 720x720 (WxH) pixels.

Upon completing the task, participants were asked to fill out the BDI-II to assess the presence and severity of depressive symptoms. After completing the depression questionnaire, participants were asked to identify a particular shape from an array of five shapes presented to them. Participants were then presented with a debrief sheet, outlining the objective of the experiment.

Data Analysis

Two main analyses were conducted: (1) comparison of group means to explore the trends in perceiving agency (ratings) and conducting actions, and (2) fitting an active inference model to provide a computational and cognitive explanation of the processes underlying these behaviours.

Comparing Group Means

Agency Ratings

All data pre-processing, statistical analyses, and visualisations were conducted in R (R Core Team, 2021). The *dplyr* (Wickham et al., 2019) and *ggplot2* packages (Wickham et al., 2016) were used to facilitate data wrangling and visualisations. For all analyses, the significance threshold was set at an α level of 0.05. The analyses conducted here were similar to those conducted in Chapter 4.

The analysis began with fitting a full linear mixed model (LMM) using the *lmer()* function from the *lme4* package in R (Bates et al., 2015; Bates et al., 2009). The full LMM included the nine experimental conditions, rating category (self or other), depressive symptom severity (BDI-II scores), position of participant's button (left or right) and their interactions. Depressive symptom severity was treated as a continuous variable, whereas the other predictors were treated as categorical variables with pre-defined factor levels. The Zero condition was set as the reference factor (dummy variable) for comparison in the model, allowing all other conditions to be interpreted relative to this baseline. Participant ID was included as a random intercept to account for individual variability in baseline ratings.

The full model was refined using a backward stepwise algorithm, using the *step()* function from the *lmertest* package (Kuznetsova et al., 2015), removing non-significant effects based on Akaike Information Criterion (AIC) (Venables & Ripley, 2013). Nested models were

then compared hierarchically to the resulting model to ensure the best-fitting model was selected for analysis.

Model assumptions, including normality of residuals, homoscedasticity, and linearity, were checked to ensure the validity of the final model. The linearity of the relationship between the independent and dependent variables was evaluated by plotting the observed ratings against the predicted (fitted) values. The normality of residuals was assessed using both a histogram and a Q-Q (quantile-quantile) plot. The assumption of homoscedasticity (constant variance of residuals) was checked by plotting the residuals against the fitted values to ensure that the spread of residuals remained constant across the range of predicted values. If model assumptions were met, the resulting model was analysed using a standard LMM. If model assumptions were violated, robust SEs were applied to the LMM. Specifically, a cluster-robust variance-covariance estimator was used to adjust the SEs. This adjustment ensures that the *p*-values are more reliable, making the conclusions drawn from the model robust to assumption violations.

For post-hoc analyses, estimated marginal means were computed for significant predictors using the *emmeans* package (Lenth, 2023). All predictors were treated as discrete for these analyses. Depressive symptom severity was categorised as "low depressive symptoms" (N=28, 41.8% of sample) and "high depressive symptoms" (N=39, 58.2% of sample), based on established cut-offs in non-clinical populations, where a score of 13 or below indicates minimal to no depressive symptoms, and a score above 13 suggests the presence of higher depressive symptoms (Beck et al., 1996; von Glischinski et al., 2019). Pairwise comparisons were then conducted to identify significant differences between specific factor levels. The Tukey method was applied to adjust *p*-values for multiple comparisons.

Actions

In addition to analysing the ratings, actions (i.e., button presses) were also examined to determine whether they were influenced by trial progression (trial number), experimental conditions, and depressive symptom severity. The random slope was specified as the trial number, investigating how behavioural actions evolved as the trials progressed across different experimental conditions. During the analysis of actions, the number of iterations for the LMM optimisation algorithm was increased (using the *bobyqa* optimiser with a maximum iteration limit of 100,000) due to convergence issues, likely caused by the large number of observations. This adjustment ensured that the model could converge effectively and produce stable estimates.

Finally, the participants' experienced contingency was calculated using ΔP based on the trial types (**A**, **B**, **C**, and **D**) produced by them. The difference between ratings and ΔP for the self was calculated, as well as the difference between ratings and ΔP for the other. Whether the difference scores differed across depressive symptom levels under different experimental conditions was then examined via LMM and post-hoc pairwise comparisons. This allowed us to examine whether participants with varying levels of depressive symptoms produced the four types of events related to action-outcome contingencies in different proportions and thereby altered the experienced objective contingency, and how this influenced their perceived sense of control.

Active Inference Model Fitting

This section describes two essential components of the active inference approach: the generative model and the generative process. The generative model represents the simulated agent's beliefs about the environment and its operations, while the generative process simulates behavioural task. Both components are crucial in active inference modelling: the generative process provides the simulated environment with which the agent (represented by the

generative model) interacts, mirroring the actual experimental conditions. The generative model, in turn, simulates the participant's learning and decision-making processes within this environment. This setup allows for a computational simulation of the experimental task and participant behaviour, which can then be fitted to actual behavioural data.

Model fitting is the process of adjusting the parameters of a computational model so that its outputs closely match observed data. In the context of active inference, this involves tuning the parameters of the generative model to reproduce the behaviour seen in human participants. The goal is to find the set of parameter values that best explain the observed data, allowing researchers to make inferences about the underlying cognitive processes, providing insights into how people make decisions and learn from their environment.

The active inference approach was implemented using the *JAX*-based Python package, *pymdp*, which facilitates the construction of flexible and efficient active inference models (Heins et al., 2022). To infer the model parameters via fitting the model to the human behavioural data, the *PyBefit* package was used, leveraging *NumPyro*, a wrapper for fitting Bayesian modelling (Bingham et al., 2019; Phan et al., 2019). The code for both the model specification and the fitting process can be found in Supplementary Material 5 (or see the following reference: Markovic & Pitliya, 2024). The model operated in a discrete time-space framework, where all variables and parameters were categorical.

Generative Model

In an active inference model, both hidden and controllable states are grouped into state factors, while observations are classified into observation modalities. State factors represent different dimensions of the agent’s internal states, while observation modalities correspond to what the agent perceives at each timestep. The agent’s task is to infer the hidden states and select actions based on these observations.

In this experiment, the agent (referred to as the "focal agent") received three key observations at each timestep: the action of the other agent (whether they pressed the button or not), the focal agent's own action (whether the button was pressed), and the outcome (whether the outcome was present or absent). Based on these inputs, the agent inferred the hidden states. Specifically, there were three state factors: the action of the other agent, the focal agent's own action, and agency. The first two state factors mapped directly to the observations. For example, if the focal agent observed the other agent pressing the button, it inferred that the other agent indeed pressed it.

The agency state factor was more complex, consisting of four possible states: Self Positive & Other Positive (SPOP), Self Positive & Other Negative (SPON), Self Negative & Other Positive (SNOP), and Self Negative & Other Negative (SNON). These states captured combinations of contingency (positive, negative, or zero) and control (self or other). Together, these four states encapsulated all possible agency scenarios and allowed the agent to compute beliefs about the degree of control exerted by both itself and the other agent over the outcome. These states were used to compute model-equivalent ratings for the self and other. For example, self-ratings were calculated as (SPOP + SPON) – (SNOP + SNON), while other-ratings were (SPOP + SNOP) – (SPON + SNON). If the focal agent believed the self had zero control but the other agent had positive control, the belief distribution across these four states might look like 0.4, 0.1, 0.4, and 0.1, respectively. This would result in a self-rating of zero (reflecting no control; (0.4 + 0.1) – (0.4 + 0.1) = 0) and an other-rating of 0.6 ((0.4 + 0.4) – (0.1 + 0.1) = 0.6).

The generative model's parameters included the likelihood, state transitions, preferences for observations, priors over states, and a prior over policy selection. These parameters were encoded in mathematical structures known as tensors, which are multi-dimensional arrays that efficiently capture the relationships between variables. For instance, the likelihood tensor specifies the probabilities of observing particular outcomes given

combinations of hidden states, while the state transition tensor represents how the agent expects hidden states to evolve over time. These tensors allow the model to process complex probabilistic relationships in a structured and scalable way. These tensors are commonly denoted A, B, C, D, and E, though in this thesis I avoid using these labels to prevent confusion with the **A**, **B**, **C**, and **D** types of trials introduced earlier. Figure 9 provides a visual representation of the active inference generative model structure, including the relationships between state factors, observation modalities, and the various tensors that define the model's parameters.

The likelihood tensor represented the agent's beliefs about the probability of observing certain outcomes given the hidden states. It was initialised with free parameters *p1* and *p2*, where $0 < p1 < 0.5$ and $0.5 < p2 < 1$, and these values were fitted to the data. For instance, the agent might initially believe there is a 75% chance of observing the outcome when both agents press the button in the SPOP state. As the agent accumulates data, the precision of its beliefs increases, which translates into stronger state beliefs and higher ratings. It was hypothesised that participants in the depressed group would have higher *p1* and lower *p2* values, leading to a flatter likelihood distribution and reflecting reduced certainty in the relationship between states and outcomes. This may result in reduced agency ratings across all experimental conditions, reflecting the general reduced sense of agency associated with depression.

The tensor for prior beliefs over hidden states captured the agent's initial expectations before receiving any observations. The agent had no prior biases regarding its own or the other agent's actions, meaning these tensors were uniformly distributed. However, the prior beliefs about agency states were adjusted during model fitting for both clinically and non-clinically depressed participants. It was hypothesised that both groups would show stronger prior beliefs for SPON and SNOP states, reflecting a tendency for relative validity of predictors – wherein if one agent is inferred to have control in one direction, the other agent is inferred to have

control in the opposite direction. Additionally, we predicted that depressed participants would exhibit stronger prior beliefs for SPOP and SNOP states compared to non-depressed participants, suggesting a bias towards believing that other agents have positive control across all conditions.

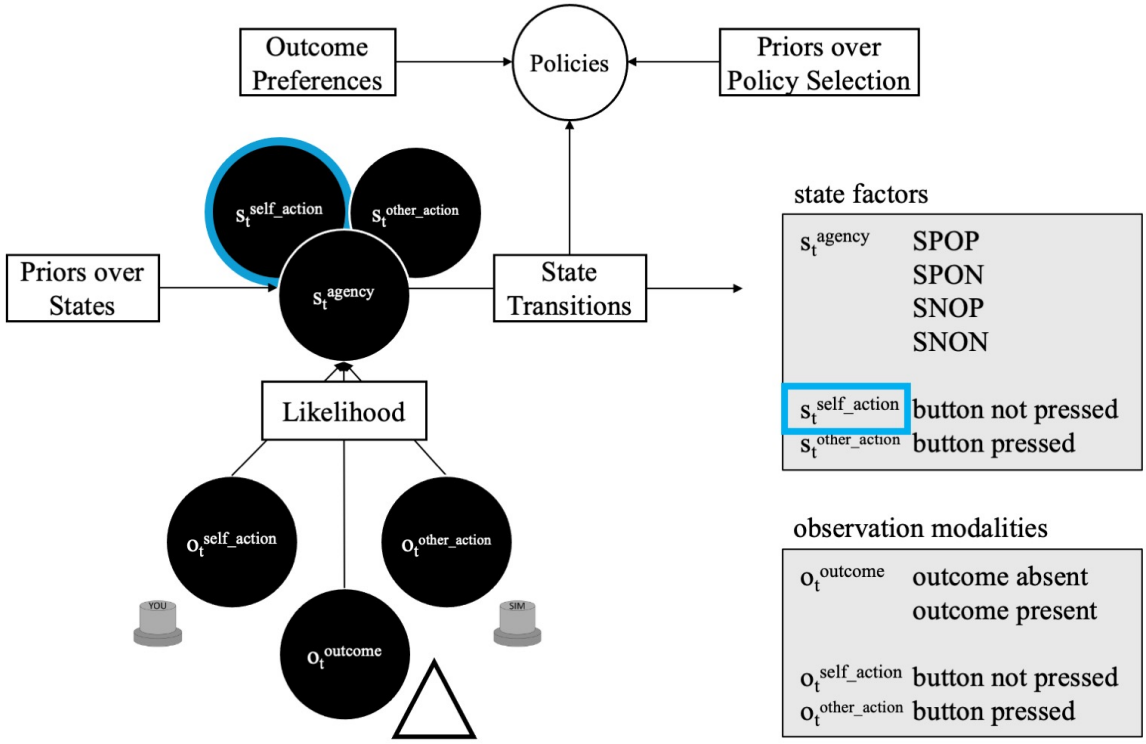
The state transition tensor encoded how the agent believed hidden states evolved over time. For agency states, the tensor remained fixed over a block of trials, and so there was no change in agency over time – it was assumed that the participants knew this. The transition tensor for the other agent's actions was set as a flat distribution, indicating the focal agent believed the other agent's actions were unpredictable. By contrast, the focal agent's own actions were fully controllable, as denoted by the blue outlined circle in Figure 9, and it could choose between pressing or not pressing the button at the next timestep. The focal agent evaluated two possible policies – pressing or not pressing the button – at each timestep, selecting the one that minimised EFE, balancing between achieving desired outcomes and reducing uncertainty about the environment.

Policy selection was influenced by a tensor representing prior beliefs over policy selection, which reflected habitual actions. A parameter called gamma modulated how much weight the agent gave to habitual actions versus decisions calculated based on EFE. Lower gamma values indicated a stronger influence of habitual behaviour. Both the value for the prior over pressing the button and the gamma value were determined through model fitting. It was hypothesised that, for the clinically depressed group, they would have weaker habits to press the button and lower gamma values, reflecting a tendency for habitual inactions to play a stronger role in decision-making.

Finally, the preference tensor encoded the agent's desire to see a particular outcome. In this task, the agent preferred outcomes where the result was present, aligning with participants' instructions to maximise the occurrence of the outcome. It was hypothesised that clinically

depressed participants would have lower preference for the outcome, consistent with diminished reward sensitivity observed in depression (Pitliya et al., 2022; Pizzagalli et al., 2009; Vrieze et al., 2013).

Figure 9.
The Active Inference Model for Experiments 4-6.



Notes. A graphical representation of the active inference generative model of the two-agent agency learning task. The states inferred in the state factor of agency (s_t^{agency}) are Self Positive Other Positive (SPOP), Self Positive Other Negative (SPON), Self Negative Other Positive (SNOP), and Self Negative Other Negative (SNON). The subscripted t refers to the trial number, which ranges from 1 to 24 here. The blue outlined state factor indicates that it is controllable by the focal agent. The model's variables are depicted as circles, while the parameters are shown as rectangles. Arrows indicate the direction of influence. For a detailed description of the variables and parameters, please refer to the main text.

Generative Process

The generative process, which represents the true environment, was defined alongside the agent's generative model. It closely mirrored the experimental task structure, incorporating three observation modalities (focal agent's action, other agent's action, and outcome) and three

state factors (focal agent's action, other agent's action, and agency states). The key difference was in the agency state factor: instead of the SPOP, SPON, SNOP, and SNON states used in the agent's generative model, the generative process specified the actual experimental conditions (Self Positive, Self Negative, Other Positive, Other Negative, each in probabilistic and deterministic versions, plus the Zero condition). Unlike the generative model, the generative process had no free parameters, as it represented the actual task dynamics. For instance, in the Probabilistic Self Positive condition, the likelihood tensor specified a 75% chance of observing the outcome when both agents pressed their buttons, and 100% for the Deterministic Self Positive condition. This setup enabled the process to accurately represent the true contingencies and control dynamics of each experimental condition, providing the environment with which the defined active inference agent interacted during simulations.

The Model Fitting Process

The model fitting process employed stochastic variational inference, which approximates the posterior distributions of model parameters. This probabilistic approach allowed capturing of parameter variations across groups and identify cognitive differences mapped onto the model parameters. The free parameters included priors over agency, the habit to press the button, gamma values, preferences for seeing the outcome, likelihood probabilities for observing the outcome given agency states, and the standard deviation for fitting ratings to beliefs.

A hierarchical setup was used for the prior distribution, employing a Regularised Horseshoe prior (Piironen & Vehtari, 2017) for group-level uncertainty and Gaussian priors for group-level means. Individual subject-specific parameters were drawn from a Gaussian distribution, conditioned on group-level means and uncertainties. This hierarchical structure balanced the need to detect group-level trends while allowing for individual differences in parameter estimates. This hierarchical prior results in estimates pulled toward group level

mean, when individual differences are non-inferable from the data, while at the same time allowing for substantial individual differences in parameter estimates when data provides sufficient evidence.

Model fitting combined this prior with a likelihood function that measured how closely model predictions aligned with observed data. The sophisticated variational inference ran for 5,000 iterations, minimising the Evidence Lower Bound (commonly referred to as ELBO), which balances model fit with complexity to find the best approximation of the true posterior distribution of model parameters. The convergence of the loss function was monitored throughout the process, and the plot can be viewed in the Supplementary Material 5. After fitting, posterior samples were analysed to compare parameters between the clinically depressed and non-depressed groups.

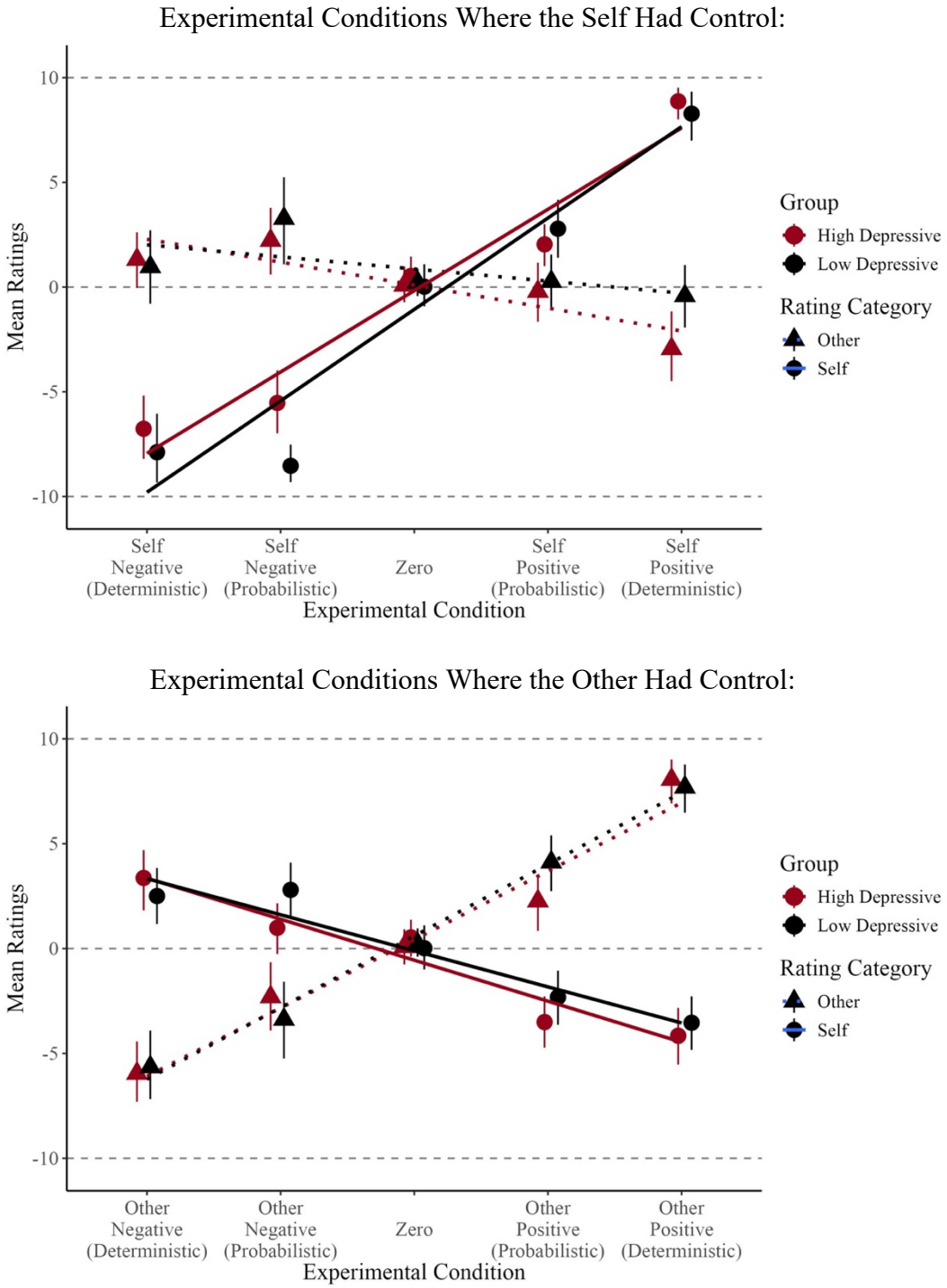
Results

Comparing Group Means

Agency Ratings

Agency ratings for both self and other participants varied significantly across experimental conditions and depressive symptom severity, suggesting that participants' perceptions of agency were sensitive to which agent had control in the experimental conditions and their individual levels of depressive symptoms. The backward stepwise algorithm identified that the interactions between the experimental conditions, rating category (self or other), and depressive symptom severity should remain in the final model ($F(8, 1105) = 2.171$, $p = 0.027$). The position of participant's button (left or right) did not influence ratings. The final model, compared against a set of nested models defined in a hierarchical manner, was confirmed as the best fitting model. Due to violations in model assumptions (see Supplementary Material 6A), robust SEs were applied, and the interactions remained significant.

Figure 10.
Results from Experiment 4: Ratings Across Experimental Conditions by Depressive Symptom Severity Group.



Notes. Mean ratings of agency attributed to the Self (circles; solid lines) and the Other (simulated participant; triangles; dotted lines) across experimental conditions. The top panel shows conditions where the Self had control, while the bottom panel shows conditions where the Other had control. Ratings are displayed for participants with low depressive symptoms (black) and high depressive symptoms (red). Each point represents the mean rating for a given condition, with error bars indicating 95% confidence intervals.

Participants generally identified which agent had control in each condition, as ratings for the self and other varied significantly across experimental conditions. As can be seen in Figure 10, participants rated the controlling agent higher than the non-controlling one and in the correct direction, with ratings generally being of a higher magnitude in the Deterministic conditions compared to the Probabilistic conditions.

In the Self Positive condition, participants rated themselves as having more positive control than the other participant under both Deterministic (β = 9.455, SE = 1.452, t(36.874) = 6.510, $p < 0.001$) and Probabilistic control (β = 3.673, SE = 1.446, t(36.874) = 2.540, p = 0.015). The larger β estimate in the Deterministic condition indicates a greater perceived difference in control between the self and other in that condition compared to the Probabilistic condition. Similarly, in the Self Negative condition, participants rated themselves as having more negative control under Deterministic (β = -9.270, SE = 1.249, t(36.874) = -7.423, $p <$ 0.001) and Probabilistic control (β = -12.774, SE = 1.606, t(36.874) = -7.953, $p < 0.001$). However, this time, the β estimates suggest a greater perceived difference in control in the Probabilistic condition compared to the Deterministic condition.

In the Other Positive condition, participants rated the other participant as having more positive control than the self under both Deterministic (β = -10.938, SE = 1.230, t(36.874) = - 8.892, $p < 0.001$) and Probabilistic control (β = -6.613, SE = 1.357, t(36.874) = -4.873, p = 0.015). Similarly, in the Other Negative condition, participants rated the other participant as having more negative control under Deterministic (β = 7.719, SE = 1.481, t(36.874) = 5.211, $p < 0.001$) and Probabilistic control (β = 6.169, SE = 1.651, t(36.874) = 3.737, p = 0.001). In both cases, the β estimates suggest a greater perceived difference in control in the Deterministic condition than the Probabilistic condition.

Depressive symptom severity had no significant effect on ratings for the self and other across all experimental conditions except the Probabilistic Self Negative condition (β = 1.784,

SE = 0.786, *t*(24.297) = 2.270, *p* = 0.032). Post-hoc analysis revealed that in the Probabilistic Self Negative condition, participants with low depressive symptom severity provided more negative ratings for the self than those in the high depressive symptom group (β = 3.002, *SE* = 1.028, *t*(1154.504) = 2.922, *p* = 0.019).

A pattern of "opposing control" emerged across experimental conditions. Participants correctly attributed agency to the controlling participant in the appropriate direction (positive or negative) but incorrectly attributed agency in the opposite direction to the non-controlling participant. For example, in the Deterministic Other Positive condition, participants accurately perceived the other participant as having positive control (*M* = 7.906, *SE* = 0.425), whilst attributing negative control to themselves (*M* = -3.900, *SE* = 0.486), despite having no actual control. The rating for the self in this condition was significantly more negative than in the Zero condition (β = 4.218, *SE* = 0.706, *t*(1105) = 5.976, *p* < 0.001), confirming that participants attributed negative control to themselves rather than perceiving themselves as having zero control.

This opposing control trend was more pronounced in conditions where the other participant had control compared to conditions where the self had control. This is reflected in Figure 10, where the "X" shape formed by the lines in the bottom panel (representing conditions where the other had control) is more distinct and symmetrical, whereas the top panel (representing conditions where the self had control) shows a less pronounced pattern, with less slanted dotted lines (representing the ratings for the other participant, who did not have control).

The trend was statistically significant for most conditions where the other participant had control, with ratings for the self differing significantly from the Zero condition in the Deterministic Other Positive ($p < 0.001$), Probabilistic Other Positive ($p < 0.001$) and Deterministic Other Negative conditions ($p < 0.001$), but not in the Probabilistic Other

Negative condition (p = 0.860). For conditions where the self had control, the trend was statistically significantly present only in the Probabilistic Self Negative condition (p = 0.040), but not the other three conditions. And since the ratings across depressive symptom severity were similar, the opposing control pattern did not differ with depressive symptom severity.

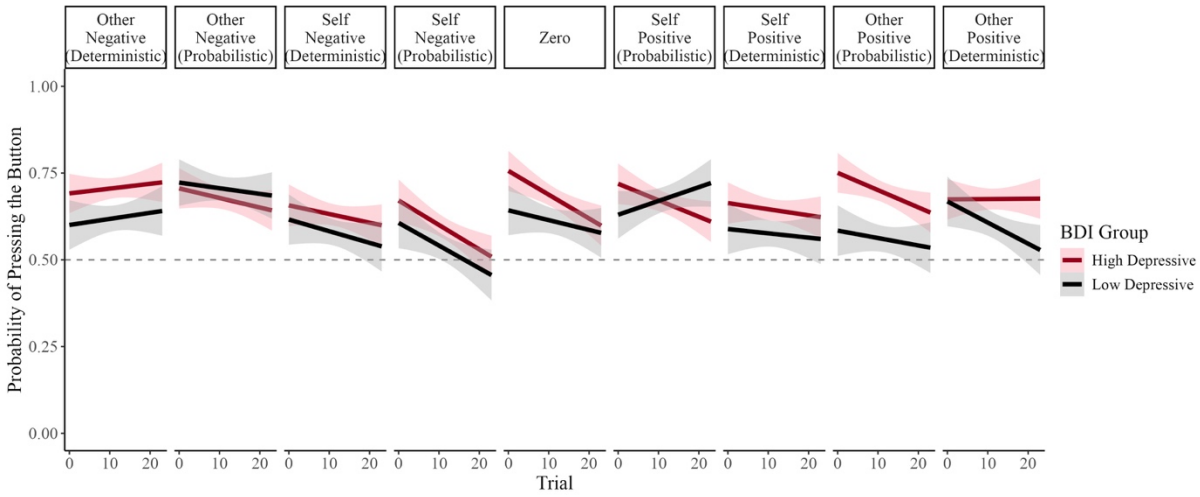
To summarise, participants were sensitive to the varying agency levels of the self and other participant across the experimental conditions. Deterministic conditions elicited stronger agency ratings than probabilistic ones. Participants consistently demonstrated an ability to correctly identify the agent with control, whilst also showing a tendency to attribute opposing control to the non-controlling agent. This opposing control effect was more pronounced when the other participant had control. Depressive symptom severity generally did not affect agency ratings, except in the Probabilistic Self Negative condition, where higher depressive symptoms were associated with less negative ratings for the self.

Actions

Pressing the Button over Time. The probability of pressing the button varied across trials, influenced by experimental conditions and depressive symptom severity. The backward stepwise algorithm revealed a significant three-way interaction between trial progression, experimental condition, and depressive symptom severity (F(8, 14306) = 2.347, p = 0.016), with comparisons to nested models confirming this model as the best fit.

As illustrated in Figure 11, button pressing generally decreased over time across all conditions. Compared to the control condition (Zero), where neither participant had control, participants further decreased their button pressing in the Probabilistic Self Negative condition (β = 0.009, SE = 0.004, t(14306) = -2.284, p = 0.022) and the Deterministic Other Positive condition (β = 0.008, SE = 0.004, t(14306) = -2.090, p = 0.037). These conditions were characterised by participants perceiving themselves as having negative control, indicating a sensitivity of their actions to their perceived level of control.

Figure 11.
Results from Experiment 4: Button-Pressing Probability Across Trials by Depressive Symptom Severity Group.



Notes. Probability of button pressing across a series of trials and experimental conditions for participants with low (black) and high (red) depressive symptom severity. The lines represent the linear trends in button-pressing probability over trials for each group, with the shaded areas indicating the standard error of the mean (SEM).

Moreover, button-pressing behaviour also varied with depressive symptom severity, with increased button pressing associated with higher severity across experimental conditions ($\beta = 0.006$, $SE = 0.002$, $t(192) = 2.192$, $p = 0.030$). This pattern is clearly visible in Figure 11, where the red lines (representing participants with high depressive symptom severity) are consistently higher and relatively flat compared to the black lines (representing those with low depressive symptom severity), which slope downwards. When comparing to the Zero condition, participants with higher depressive symptom severity pressed the button statistically significantly more than those with lower depressive symptom severity in the Deterministic Self Negative condition ($\beta = 0.004$, $SE = 0.002$, $t(14306) = 2.283$, $p = 0.022$), Probabilistic Self Negative condition ($\beta = 0.004$, $SE = 0.002$, $t(14306) = 2.266$, $p = 0.023$), and Deterministic Other Positive condition ($\beta = 0.006$, $SE = 0.002$, $t(14306) = 3.340$, $p = 0.001$). These were all conditions where the self was perceived as having negative control.

Interestingly, as shown in Figure 11, in the Probabilistic Self Positive condition, participants with low depressive symptom severity appeared to increase their button pressing

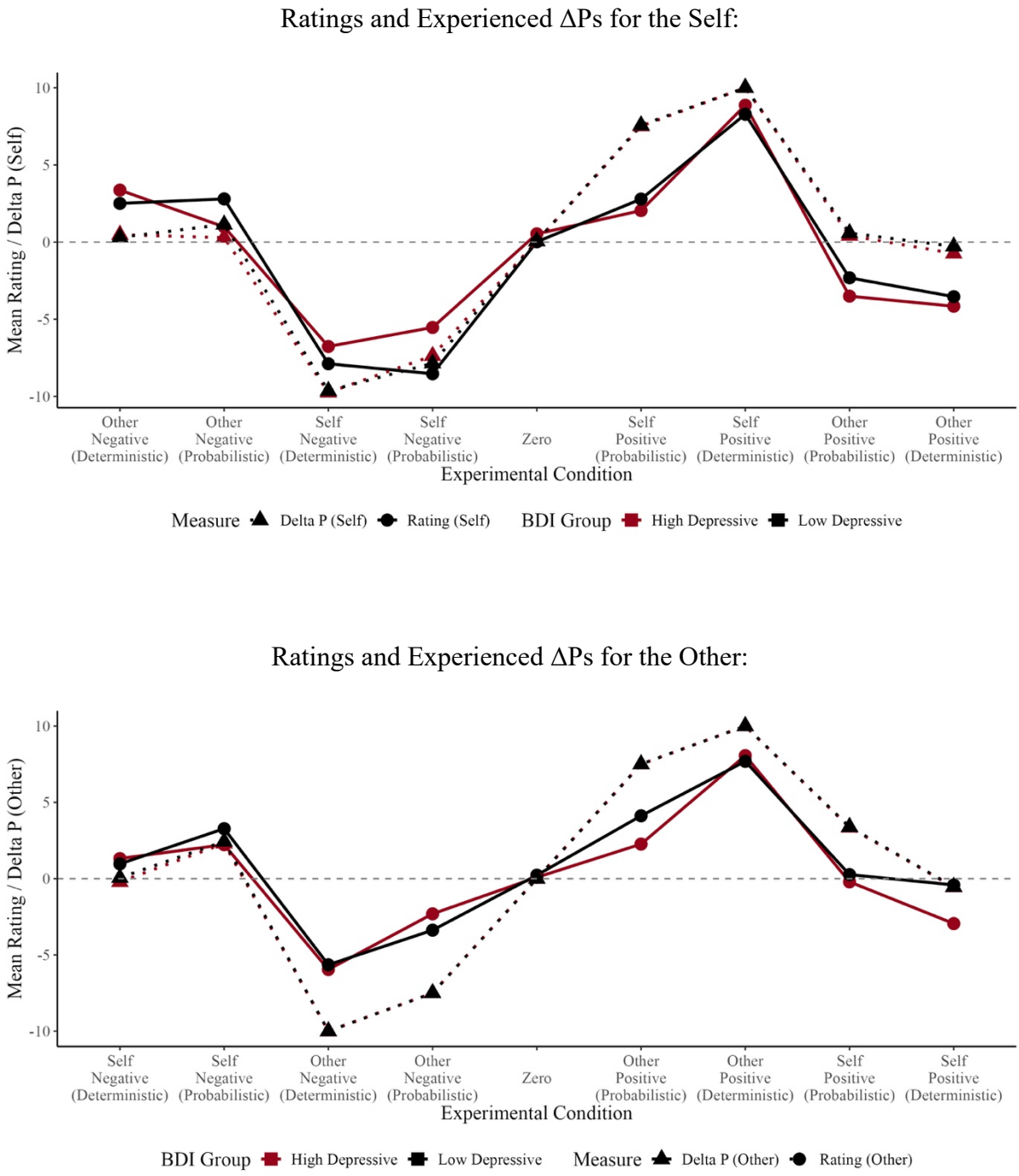
as trials progressed, while those with high depressive symptom severity continued to decrease their button pressing. However, this difference was not statistically significant ($p = 0.210$).

In summary, this study revealed interactions between experimental conditions, trial progression, and depressive symptom severity in button-pressing behaviour when learning different agents' agency levels, as visualised in Figure 11. Key findings include: (1) a general decrease in button pressing over time across all conditions, (2) further decreases in button pressing in conditions where participants perceived negative control, and (3) consistently higher button pressing among participants with higher depressive symptom severity, with this difference being most pronounced in conditions where the self was perceived as having negative control.

Experienced ΔP. Figure 12 illustrates the mean ratings and experienced ΔPs of agency attributed to the self (top graph) and the other (bottom graph) across various experimental conditions. Generally, the ratings did not consistently align with the experienced ΔPs, as indicated by the circles (ratings) and triangles (experienced ΔPs) not always overlapping.

In conditions where the participant had control, ratings tended to align more closely with the experienced ΔPs. This is particularly evident when the self was in control (Figure 12, top graph, middle five conditions). However, an exception is seen in the Probabilistic Self Positive condition, where the ratings were notably lower than the experienced ΔP, indicating a tendency for participants to perceive less control than they had there. In contrast, when considering the other agent's control (Figure 12, bottom graph, middle five conditions), the alignment between ratings and experienced ΔPs was less consistent. Here, ratings often regressed toward zero, suggesting a perception of reduced control for the other agent even when the experienced ΔPs indicated otherwise.

Figure 12.
Results from Experiment 4: Ratings and Experienced ΔPs Across Experimental Conditions, by Depressive Symptom Group.



Notes. Mean ratings (circles) and experienced ΔPs (triangles) of agency attributed to the Self (top graph) and the Other (bottom graph) across experimental conditions, for participants with low depressive symptoms (black) and high depressive symptoms (red).

For conditions where the self lacked control (two outermost conditions in the top graph), the ratings deviated from the experienced ΔPs, reflecting the "opposing control" pattern. This means that participants attributed more control to the self when it actually had none, and in the direction opposing the other agent's. In the conditions where the other agent lacked control (two outermost conditions in the bottom graph), the ratings tended to align more closely with the experienced ΔPs, reflecting a more accurate perception of the other agent's lack of control. Here, an interesting pattern emerged in the Probabilistic Self Negative and Probabilistic Self Positive conditions, where the experienced ΔP for the other agent was slightly positive instead of zero. While ratings aligned with the experienced ΔP in the Probabilistic Self Negative condition, they deviated in the Probabilistic Self Positive condition, following the opposing control pattern.

The experienced action-outcome contingencies (ΔPs) for self and other did not significantly differ with depressive symptom severity. This can be seen in Figure 12 as the black and red triangles, representing the experienced ΔPs of the low and high depressive severity groups respectively, overlap with each other. This means that when they sampled their environment, they generated similar types of information, resulting in similar experienced objective action-outcome contingencies as calculated by ΔP.

However, a notable difference emerged when comparing participants' ratings for the self and other participant to their experienced ΔPs across experimental conditions. In the Probabilistic Self Negative condition, the discrepancies between ratings and experienced ΔPs for the self varied with depressive symptom severity. Participants with high depressive symptoms demonstrated a significantly larger discrepancy between their ratings and experienced ΔPs compared to those with low depressive symptoms (β = -2.515, SE = 1.03, $t(568) = -2.446$, $p = 0.015$). Specifically, participants with high depressive symptoms perceived themselves as having less negative agency than those with low depressive symptoms, despite

both groups experiencing similar levels of negative contingencies between the actions produced by the self and the outcomes.

The difference between ratings and experienced ΔPs for the other agent also varied with depressive symptom severity. In the Deterministic Self Positive condition, participants with high depressive symptoms showed a significantly larger discrepancy between their ratings and experienced ΔPs for the other agent (β = 2.512, SE = 1.09, $t(584)$ = 2.311, p = 0.021). Participants with high depressive symptoms perceived the other agent as having more negative agency compared to those with low depressive symptoms, even though both groups experienced the contingency between the other participant's action and outcome as zero.

In summary, ratings were more aligned with experienced ΔPs when the self had control compared to when the other agent had control, and ratings deviated from experienced ΔPs when the agents did not have control in the direction opposing the agent that did have control. Whilst the experienced ΔPs did not differ significantly based on depressive symptom severity, the perception and interpretation of these contingencies varied distinctly. Participants with high depressive symptoms underestimated their own negative agency when they had negative agency and overestimated the negative agency of others when the other actually had none. These findings suggest that participants with higher depressive symptoms may have distorted perceptions of agency for the self and others, despite generating similar patterns of action-outcome data through their interactions with the task environment. This indicates that the differences in agency perception are not due to variations in the underlying data produced, but rather in how this information is processed and interpreted.

Summary

The results from this experiment revealed complex interactions in how individuals with varying depressive symptom severity perceive agency of the self and other participants, and how they interact with their environments when learning agency in multi-agent contexts.

The agency ratings demonstrated that while all participants were generally able to correctly identify who had control, they also attributed control in the opposing direction to the participant who did not have any control in that condition. This effect was particularly pronounced when the other participant had control (i.e., assigning more control to the self when the self had none in such conditions) and in deterministic conditions compared to probabilistic ones. Regarding sampling of the environment, all participants tended to decreasingly press the button as trials progressed across all conditions, but more so when they perceived to have negative control. Participants with higher depressive symptom severity pressed the button more than those with lower depressive symptom severity, i.e., less decreasingly and more uniformly.

Differences in behaviour and agency perception mainly emerged in the Probabilistic Self Negative condition. Here, participants with higher depressive symptoms provided significantly less negative ratings for themselves and consistently pressed the button more frequently, compared to those with lower depressive symptoms who rated themselves as having more negative control and who pressed the button significantly more decreasingly in this condition compared to other conditions. Notably, even though both groups experienced very similar ΔPs of negative contingencies between their actions and outcomes, participants with higher depressive symptoms demonstrated a significantly larger discrepancy between their ratings and experienced ΔPs compared to those with low depressive symptoms. The underestimation of negative agency for the self, combined with persistent environmental sampling and a larger gap between perceived and actual contingencies, suggests difficulties in processing and interpreting action-outcome relationships, particularly in scenarios involving negative self-agency.

Active Inference Model Fitting

The active inference model fitting results provided a computational account of the observed perceptual and behavioural differences between participants with low and high

depressive symptom severity. The model incorporated several free parameters, including priors over agency states, habit to press the button, gamma values (influencing the balance between habitual and goal-directed behaviour), preferences for observing outcomes, likelihood probabilities for observing outcomes given agency states (precision (confidence) of beliefs about action-outcome relationships), and the standard deviation for fitting ratings to beliefs. The analysis revealed distinct differences between the low and high depressive symptom severity groups in the parameters related to action behaviour (outcome preferences, habitual actions, and gamma values), but not in the parameters related to perceptual processes (priors over agency states and precision of beliefs about action-outcome relationships); Table 5.

Table 5.
Results from Experiment 4: Active Inference Model Fitting Results for Low and High Depressive Symptom Groups.

Parameter	Low Depressive Symptoms (N=28)	High Depressive Symptoms (N=39)
Probability of observing outcome present given agency and action states for the likelihood tensor	0.548	0.540
Preference to observe the outcome present	3.026	1.618
Habit to press the button	0.626	0.676
Gamma (weighting for habit)	3.784	2.315
Prior for SPOP state	0.132	0.111
Prior for SPON state	0.337	0.367
Prior for SNOP state	0.379	0.364
Prior for SNON state	0.151	0.158
Standard deviation for ratings regarding self	0.268	0.313
Standard deviation for ratings regarding the other	0.323	0.323

Notes. SPOP refers to Self Positive & Other Positive; SPON refers to Self Positive & Other Negative; SNOP refers to Self Negative & Other Positive; and SNON refers to Self Negative & Other Negative.

The model fitting revealed differences in outcome preferences and action habits between the two groups. Participants with low depressive symptoms showed a stronger preference to observe the outcome present (3.026) compared to participants with high depressive symptoms (1.618). Regarding button-pressing habits, participants with high depressive symptoms showed a slightly higher habit of pressing the button (0.676) compared to participants with low depressive symptoms (0.626). But this, paired with a lower gamma value for participants with high depressive symptoms (2.315, compared to 3.784 for participants with low depressive symptoms), indicated that their actions were more influenced by habits rather than goal-directed behaviour calculated based on EFE. Together, these finding align with the behavioural data, which showed that participants with high depressive symptoms maintained more and consistent button-pressing behaviour across conditions and over time, showing less behavioural flexibility across different experimental conditions.

The model fitting results for parameters pertaining to the perceptual processes revealed similarities between the parameters estimated for the low and high depressive symptoms groups. The fitted parameters for both groups showed higher values for SPON (Self Positive & Other Negative) and SNOP (Self Negative & Other Positive) states, suggesting a common bias where agency in one direction for one agent implies the opposite direction for the other agent. This finding aligns with the data regarding agency ratings, where participants showed the opposing control pattern in their ratings.

The parameter fitting results for the likelihood tensor revealed similar belief precisions between the two groups, 0.548 for the low vs 0.540 for the high depressive symptoms group. This finding aligns with our finding that agency ratings did not vary significantly according to depressive symptom severity.

Discussion

Supporting our hypotheses, the results indicate that participants were generally able to correctly identify which agent had control in each condition. They assigned higher ratings to the agent with control, corresponding to the direction of that control, and they distinguished between deterministic and probabilistic conditions. The results also demonstrated the "opposing control" pattern: participants attributed some degree of control to the non-controlling agent in the opposite direction of the controlling agent's influence. This tendency was evident across conditions, aligning with previous findings and suggesting that participants often perceive control as a relative concept between agents. The active inference model supported this behavioural pattern through the priors over agency states, which showed a shared bias where agency in one direction for one agent implies the opposite direction for the other. This alignment between the behavioural data and the model underscores a complex process of agency attribution in multi-agent contexts.

Contrary to our expectations, these patterns did not vary significantly with depressive symptom severity. Participants with higher depressive symptoms provided ratings similar to those with lower symptoms, and both demonstrated the opposing control pattern. However, one exception emerged in the Probabilistic Self Negative condition, where participants with higher depressive symptoms rated their own control as less negative compared to those with lower symptoms, supporting our hypothesis that depressive symptoms might weaken the perception of negative self-agency. This behavioural finding, however, was not reflected in the active inference model fitting as it was not clearly reflect this difference in priors; I would have expected this result to appear as a higher prior to SPON than SNOP. This suggests that the discrepancy may arise more from how participants interpret their actions and outcomes than from their underlying beliefs about agency.

Also contrary to our hypotheses, individuals with higher depressive symptoms did not engage less with the environment compared to those with lower symptoms; rather, they displayed more consistent button-pressing behaviour throughout the trials. While participants with lower depressive symptoms reduced their button presses as trials progressed and adjusted their behaviour according to the varying levels of control across conditions, those with higher depressive symptoms maintained relatively constant button-pressing behaviour.

This difference suggests that participants with lower symptoms learned over time, needing less exploration as they developed a clearer understanding of their control. And their behaviour remained somewhat goal-directed, adjusting actions to produce desired outcomes more effectively. In contrast, the more uniform behaviour of those with higher symptoms suggests less learning across trials. Despite engaging more with their environment, these actions appeared less goal-directed, potentially reflecting a broader or less focused exploration strategy and less ability to adapt behaviours. These behavioural trends were mirrored in the active inference model results, where participants with lower depressive symptoms exhibited a stronger preference for outcomes and lower reliance on habitual button-pressing, while those with higher symptoms showed the opposite: lower outcome preference and a greater emphasis on habitual actions.

Given these findings, future experiments could focus on a more streamlined design by eliminating the probabilistic conditions and retaining only the deterministic ones. This approach would allow for a clearer examination of how participants attribute agency when outcomes are more predictable, which might help further differentiate patterns of goal-directed behaviour across different groups. Additionally, extending the research to include a clinically depressed sample, which has not been explored in the experiments reported thus far in the experiments, could provide deeper insights into how severe depressive symptoms affect the perception of control and engagement with the environment to learn action-outcome

contingencies. Such modifications could clarify the mechanisms behind these behavioural differences and refine our understanding of how agency is perceived and learned in both typical and clinical populations.

Experiment 5

Participants

A total of 87 participants were recruited from three sources to take part in the experiment, and subsequently grouped into two categories for analysis: clinically depressed individuals (31% of the sample) from an in-patient clinic in Ferrara, Italy, and non-clinically depressed individuals (69% of the sample) from the general population in Ferrara and undergraduate students from the University of Oxford. The data were analysed for the three separate sample groups to verify whether the results remained consistent when merging two of these groups into a single category, and since the findings were consistent, we proceeded with the two-group categorisation for analysis.

Recruitment from the general population in Ferrara occurred via word of mouth – specifically through friends and family of the research assistants who collected data from the clinic's in-patients. Undergraduate participants were recruited via the Department of Experimental Psychology's internal recruitment platform, Sona Systems (https://www.sona-systems.com).

To ensure adequate attention to the task, participants who provided uniform responses across 90% or more of the task conditions were excluded from analysis, resulting in seven exclusions. The final analysed sample consisted of 80 participants in total, including 30 males, 47 females, 2 non-binary individuals, and 1 participant who preferred not to disclose their gender. The overall mean age of the final sample was 32.6 years (SD of 14.8).

Table 6.
Experiment 5: Demographic Information from the Three Sample Groups.

Sample Source	In-patient depressed individuals from a clinic in Ferrara, Italy	Not clinically depressed individuals in Ferrara, Italy	Not clinically depressed individuals from the University of Oxford, UK
Total number of participants	25	24	31
Number of male participants	10	9	11
Number of female participants	15	14	18
Number of participants who identify with neither of the above two genders	0	1	1
Mean age	46.0	35.5	19.5
Standard deviation of age	12.2	12.6	1.5
Mean BDI-II score	29.7	5.0	10.0
Standard deviation of BDI-II score	11.5	5.3	11.2

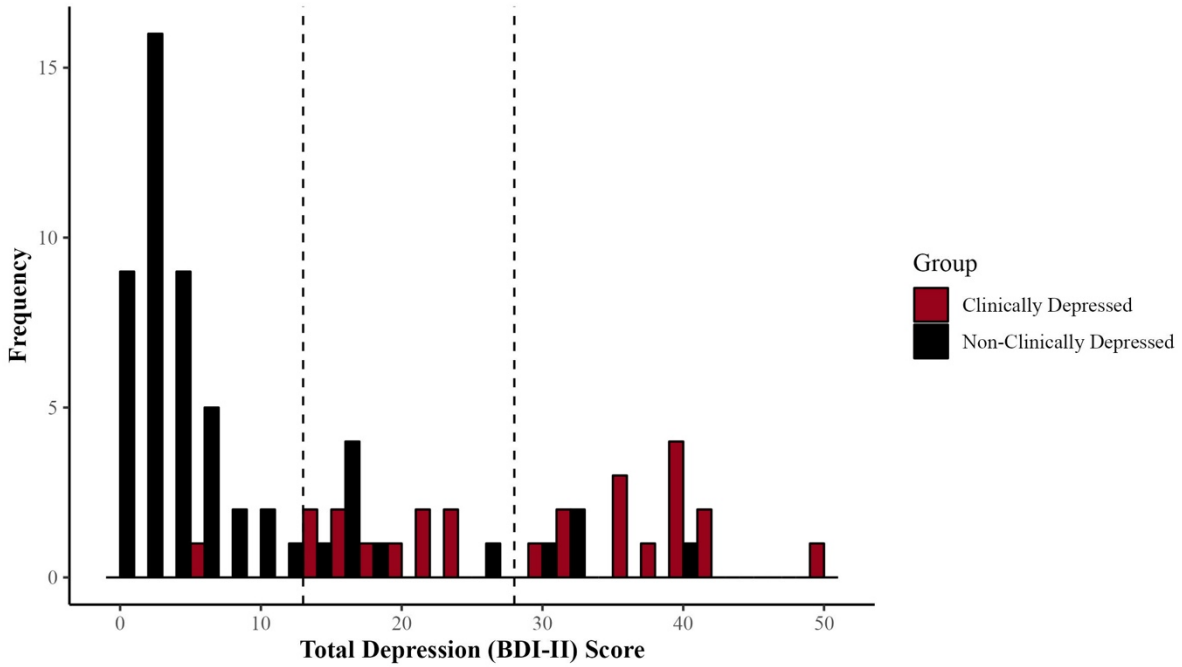
The non-clinically depressed group comprised of 55 participants, 20 of which identified as being male, 32 as being female, 2 as non-binary, and 1 participant preferred not to disclose their gender. The overall mean age of this group was 26.5 years (SD of 11.5). The clinically depressed group comprised of 25 participants, 10 of which identified as being male and 15 as being female. The overall mean age of this group was 46.0 years (SD of 12.2). See Table 6 for demographic details of the three sample groups.

Participants also completed the BDI-II. Although BDI-II scores were not used to categorise participants into groups in the present experiment, their scores are provided below for reference. The mean BDI-II score among participants was 14.663 (SD = 14.324), with scores ranging from 0 to 51. According to BDI-II categorisation (Beck et al., 1996), around 56.25% of all participants fell into the minimal depression range (scores of 0–13; N=45),

21.25% into the mild-to-moderate range (scores of 14–28; N=17), and 22.50% into the severe range (scores of 29–63; N=18). Mean BDI-II scores were 29.7 (SD = 11.5) for the clinically depressed group and 7.6 (SD = 8.7) for the non-clinically depressed group (see Figure 13 for the distribution of BDI-II scores between the two depression groups).

Figure 13.
Experiment 5: Distribution of Total Depression (BDI-II) Scores for Clinically and Non-Clinically Depressed Groups.

.



Notes. This figure displays the distribution of total depression scores as measured by the Beck Depression Inventory (BDI-II; Beck et al. (1996)) for two groups: clinically depressed (in red) and non-clinically depressed (in black). The x-axis represents the total BDI-II scores, ranging from 0 to 50, while the y-axis shows the frequency of participants within each score range. The two dashed vertical lines indicate the BDI-II categorisation thresholds: scores of 0–13 fall within the minimal depression range, scores of 14–28 correspond to the mild-to-moderate depression range, and scores of 29 and above fall into the severe depression range.

Materials, Procedure, and Task Design

At the beginning of the experiment, participants were presented with a briefing sheet detailing the task, data usage and storage, contact information for queries, and their right to withdraw. They provided informed consent and demographic information (age and gender) and confirmed they did not have a history of seizures or diagnosed motor or cognitive impairment before completing the computer task.

Participants completed the same task as the one outlined in the previous experiment. Briefly, in the task, in each block of trials, participants freely produced actions, observed another (simulated) participant's actions, integrated this information with the absence or presence of outcomes, and provided ratings about how much control both agents had, reflecting their learning of action-outcome contingencies (Figure 8).

However, in this experiment, the Probabilistic experimental conditions were removed. This resulted in five conditions: the control condition being Zero, where no one has control, and the four experimental conditions being Self Positive, Self Negative, Other Positive, and Other Negative. When one agent had control, the other had none. The contingencies were deterministic, such that when an agent had control, they had 100% control.

Participants completed a total of 15 blocks of trials (plus one practice condition), structured as three repeats of the five conditions. Participants were allowed a maximum of two five-minute breaks between blocks of trials.

Upon completing the task, participants were asked to fill out the BDI-II to assess the presence and severity of depressive symptoms. After completing the depression questionnaire, participants were asked to identify a particular shape from an array presented to them. Participants were then presented with a debrief sheet, outlining the objective of the experiment. The whole experiment took approximately 25 minutes to complete.

Data Analysis

The same analytical approach as in Experiment 4 was employed for Experiment 5, with a few key differences in the variables and model specifications.

Comparing Group Means

Agency Ratings

The analysis began with fitting a full linear mixed model (LMM), similar to Experiment 4. However, the model for Experiment 5 included depression group (clinical or non-clinical) as a categorical variable, replacing the continuous depressive symptom severity measure used in Experiment 4. The full LMM comprised the experimental condition (Zero, Self Positive, Self Negative, Other Positive, Other Negative), rating category (self or other), depression group, position of participant's button (left or right), repeat blocks of each condition, and their interactions. The repeat blocks did not have a significant effect on ratings, so the ratings were averaged over the repeated conditions for the analyses. The model refinement process (using a backward stepwise algorithm), examination of significant predictors, and post-hoc analyses (including pairwise comparisons) followed the same procedure as in Experiment 4.

Actions

The analysis of actions (button presses) mirrored that of Experiment 4, examining the influence of trial progression, experimental conditions, and depression group (rather than depressive symptom severity) on participants' behaviours. Additionally, as in Experiment 4, the participants' experienced contingency was calculated using ΔP based on the trial types they produced, and difference scores between ratings and ΔP for both self and other were computed. These difference scores were examined across depression groups under different experimental conditions using LMM and post-hoc pairwise comparisons, following the same analytical approach as in Experiment 4.

Active Inference Model Fitting

The active inference modelling approach for Experiment 5 was consistent with that of Experiment 4, maintaining the same structure for the generative model but editing the generative process to reflect the changes made in the task design in this experiment. The key difference lay in the comparison between clinically depressed and non-clinically depressed groups, rather than low and high depressive symptom severity groups.

The model fitting process also employed the same methods as in Experiment 4, with posterior samples analysed to compare parameters between the clinically depressed and non-clinically depressed groups, aiming to identify potential cognitive differences between these distinct groups.

Results

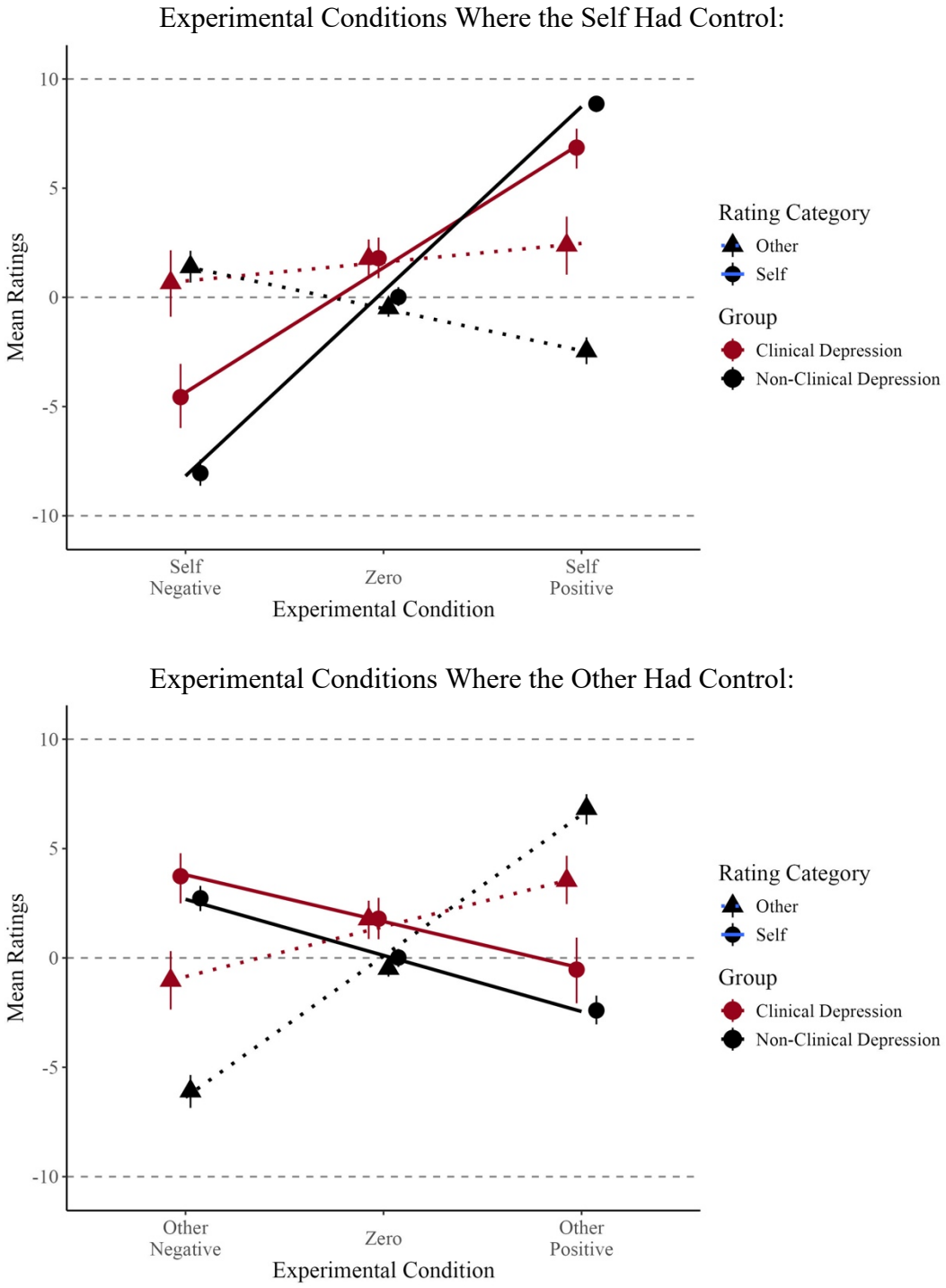
Comparing Group Means

Agency Ratings

Agency ratings for both self and other participants varied significantly across experimental conditions and depression groups, suggesting that participants' perceptions of agency were sensitive to which agent had control in the experimental conditions and whether individuals were clinically depressed or not. The backward stepwise algorithm identified that the interactions between the experimental conditions (Zero, Self Positive, Self Negative, Other Positive, Other Negative), rating category (self or other), and depression group (clinical or non-clinical) should remain in the final model ($F(4, 2302) = 36.349$, $p < 0.001$). The position of participant's button (left or right) did not influence ratings. The final model was compared against a set of nested models defined in a hierarchical manner, confirming that the final model was the best fitting model. Due to violations in model assumptions (see Supplementary Material 6B), robust SEs were applied, and the interaction effects remained significant.

Four key trends emerged from the behavioural data. First, clinically depressed participants consistently rated their own agency and the other participant's agency lower compared to non-clinically depressed participants. Second, non-clinically depressed participants attributed control in the correct direction (i.e., positive or negative) to the participant who had control but consistently attributed control in the opposite direction to the participant without control. This pattern of "opposing control" was less pronounced and less consistent among clinically depressed participants. Third, generally, clinically depressed participants rated the self and the other as having more positive control, whereas non-clinically depressed participants rated the other slightly negatively compared to their slightly positive rating for themself. Fourth, ratings for the self and other participant differed across experimental conditions, except for clinically depressed participants, whose ratings for the other did not vary significantly between conditions. These trends suggest that clinically depressed individuals exhibit a reduced but positively skewed perception of agency for both the self and the other, with an unchanging bias attributing positive control to the other. In contrast, non-clinically depressed participants exhibited more amplified perceptions of agency, adapting to the different experimental conditions, but with a slight self-serving bias, wherein the self was seen as having positive control, while the other was perceived as having negative control. The remainder of this section will explore these trends in more detail.

Figure 14.
Results from Experiment 5: Ratings Across Experimental Conditions Where the Self Had Control, by Depression Group.



Notes. Mean ratings of agency attributed to the Self (circles; solid lines) and the Other (simulated participant; triangles; dotted lines) across experimental conditions. The top panel shows conditions where the Self had control, while the bottom panel shows conditions where the Other had control. Ratings are displayed for participants that were non-clinically depressed (black) and clinically depressed (red). Each point represents the mean rating for a given condition, with error bars indicating 95% confidence intervals.

First, clinically depressed participants consistently gave lower ratings compared to the non-clinically depressed group. This pattern is evident in Figure 14, where the red points (clinically depressed participants) cluster closer to the zero line than the black points (non-clinically depressed participants), indicating a reduced overall perception of control in the clinically depressed group. This trend is supported by pairwise comparisons in the Other Positive condition, where clinically depressed participants rated the other participant positively (M = 3.541, SE = 0.572) but significantly lower than the ratings provided by non-clinically depressed participants (M = 6.830, SE = 0.352), β = -3.289, SE = 0.658, t(1040.425) = -4.997, $p < 0.001$. Meanwhile, in the same condition, the ratings given by clinically depressed participants for the self were negative (M = -0.533, SE = 0.767) but significantly higher (closer to zero) than those given by non-clinically depressed participants (M = -2.400, SE = 0.336), trending towards zero, β = 1.867, SE = 0.658, t(1040.425) = 2.836, p = 0.024. Similar trends are observed across other conditions. These findings indicate that clinically depressed participants generally perceived reduced agency for both participants compared to non-clinically depressed participants.

Second, in each experimental condition, non-clinically depressed participants attributed agency to the participant with control in the correct direction (i.e., positive or negative), but incorrectly attributed agency in the opposite direction to the participant without control. For example, in the Self Positive condition, non-clinically depressed participants correctly perceived the self as having positive control (M = 8.864, SE = 0.191), while they attributed negative control to the other participant (M = -2.455, SE = 0.320), despite the other having no control. The rating for the other in this condition was significantly more negative than in the Zero condition (β = 1.975, SE = 0.480, t(2302) = 4.115, p = 0.002), confirming that non-clinically depressed participants attributed negative control to the other, rather than perceiving

them as having zero control as they did in the Zero condition. We can see this trend in Figure 14 as the black lines across both graphs create an "X" shape.

This pattern was less pronounced among clinically depressed participants, reflected in the less clear "X" shapes formed by the red lines in Figure 14. In the Other Positive condition, the other participant was correctly rated as having positive control (M = 3.541, SE = 0.572), while the self was rated as having slightly negative control (M = -0.533, SE = 0.767). The rating for the self was significantly different from the rating for the self in the Zero condition (β = 2.328, SE = 0.712, $t(2302) = 3.271$, p = 0.036), confirming that clinically depressed participants indeed perceived the self as having negative control compared to their baseline of zero control (as in the Zero condition).

However, this trend did not hold across all experimental conditions for clinically depressed participants. For example, in the Self Positive condition, the self was correctly perceived as having positive control (M = 6.857, SE = 0.482), and they also gave the other participant a positive control rating (M = 2.388, SE = 0.666). Importantly, the latter rating did not significantly differ from the rating of the other in the Zero condition, indicating that the other was perceived as having zero control. This occurred in the Self Negative and Other Negative conditions too. These findings suggest little to no trend of opposing control demonstrated by clinically depressed participants. This is reflected in Figure 14, where the dotted red lines across both graphs slant the same way and less than the black dotted lines.

These observed deviations from the opposing control trend in the clinically depressed group leads us to the third trend observed: they tended to rate both self and other as having similar levels of positive control, and in contrast, non-clinically depressed participants viewed the other as having slightly negative control, opposing their slightly positive self-agency. This distinction was particularly evident in the Zero condition, where neither participant had control. Non-clinically depressed participants accurately perceived this lack of agency (self: M = 0.019,

SE = 0.218; other: *M* = -0.481, *SE* = 0.192), while clinically depressed participants perceived higher and similar levels of positive control for both (self: *M* = 1.794, *SE* = 0.487; other: *M* = 1.780, *SE* = 0.442. The group differences were statistically significant for both self (β = 1.776, *SE* = 0.658, *t*(1040.425) = 2.698, *p* = 0.036), and other (β = 2.261, *SE* = 0.658, *t*(1040.425) = 3.434, *p* = 0.003).

It is noteworthy that while both groups perceived the self as having more positive control compared to the other, the clinically depressed group attributed higher positive control to the self than the non-clinically depressed group did. Averaged across all conditions, clinically depressed participants rated the other participant at 1.470 (*SE* = 0.296) and themselves at 1.457 (*SE* = 0.345). In comparison, non-depressed participants gave lower, more balanced ratings: -0.159 (*SE* = 0.210) for the other participant and 0.232 (*SE* = 0.231) for the self, as supported by the significant main effect of the depression group in the LMM (β = -2.261, *SE* = 0.596, *t*(46.506) = -3.795, *p* < 0.001).

Fourth, ratings given by non-clinically depressed participants for self and other significantly differed across experimental conditions. Specifically, the ratings for self significantly differed from the Zero condition to the Self Positive (β = -8.845, *SE* = 0.480, *t*(2302) = -18.435, *p* < 0.001), Self Negative (β = 8.070, *SE* = 0.480, *t*(2302) = 16.819, *p* < 0.001), Other Positive (β = 2.419, *SE* = 0.480, *t*(2302) = 5.041, *p* < 0.001), and Other Negative (β = -2.712, *SE* = 0.480, *t*(2302) = -5.651, *p* < 0.001) conditions. And the ratings for the other agent significantly differed from the Zero condition to the Self Positive (β = 1.975, *SE* = 0.480, *t*(2302) = 4.115, *p* = 0.002), Self Negative (β = -1.879, *SE* = 0.480, *t*(2302) = -3.917, *p* = 0.004), Other Positive (β = -7.311, *SE* = 0.480, *t*(2302) = -15.237, *p* < 0.001), and Other Negative (β = 5.607, *SE* = 0.480, *t*(2302) = 11.686, *p* < 0.001) conditions.

Clinically depressed participants' ratings for the self also significantly differed across conditions, except for one. The ratings for self significantly differed from the Zero condition

to the Self Positive (β = -5.063, SE = 0.712, t(2302) = -7.113, $p < 0.001$), Self Negative (β = 6.363, SE = 0.712, t(2302) = 8.940, $p < 0.001$), Other Positive (β = 2.328, SE = 0.712, t(2302) = 3.127, p = 0.036), and Other Negative (p = 0.163) conditions. In contrast, clinically depressed participants' ratings for the other did not significantly differ between conditions, suggesting an unchanging bias that attributes positive agency to the other, regardless of condition. The ratings for the other agent significantly differed from the Zero condition only to Other Negative (β = 2.808, SE = 0.712, t(2302) = 3.945, p = 0.003) condition, but not to the Self Positive (p = 0.998), Self Negative (p = 0.866), Other Positive (p = 0.283) conditions.

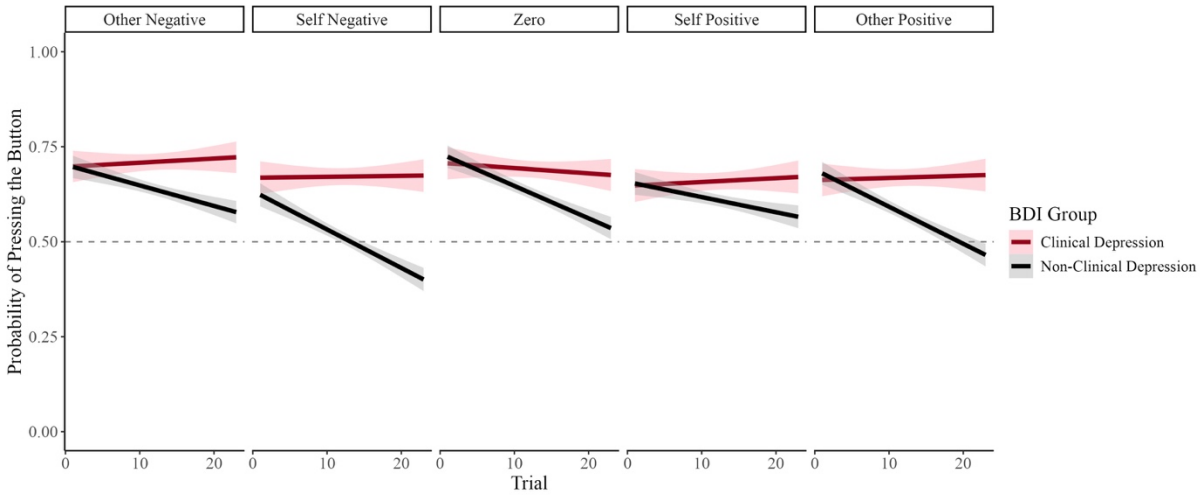
In summary, the behavioural data revealed distinct patterns of learning and perceiving agency between non-clinically depressed and clinically depressed participants in a multi-agent context. Non-clinically depressed participants showed sensitivity to the different agency levels of the self and other across experimental conditions, with their agency ratings for both self and other varying significantly across conditions. They demonstrated the opposing control pattern, where they correctly attributed agency to the participant with control, but also attributed control in the opposite direction to the participant without control. They also demonstrated a slight self-serving bias, tending to view themselves as having slightly positive control, while seeing others as having slightly negative control. In contrast, clinically depressed participants consistently gave lower ratings for agency across all conditions, both for themselves and others, suggesting a reduced overall perception of control. Unlike their non-clinical counterparts, who demonstrated a slight self-serving bias, they tended to rate both self and other as having similar levels of positive control. Importantly, clinically depressed participants' ratings for others remained relatively constant across conditions, indicating an unchanging bias that attributed positive agency to others, while the ratings for self varied across conditions albeit less sensitively than non-clinically depressed participants. The opposing control pattern was less pronounced and less consistent in this group. These findings highlight that depression is

associated with a reduced but positively skewed perception of agency, with less sensitivity to changes in others' agency, while non-depressed individuals demonstrate more varied and condition-dependent agency attributions.

Actions

Pressing the Button over Time. Overall, the probability of pressing the button varied across trials depending on both the experimental conditions and depression group, as shown in Figure 15. Across all experimental conditions, both clinically and non-clinically depressed participants pressed the button decreasingly as trials progressed. However, non-clinically depressed participants demonstrated greater responsiveness to changes in experimental conditions, adjusting their actions more as trials progressed across conditions. In contrast, clinically depressed participants displayed less sensitivity to these changes, maintaining relatively consistent actions, even when pressing the button prevented the outcome from occurring.

Figure 15.
Results from Experiment 5: Button-Pressing Probability Across Trials by Depression Group.



Notes. Probability of button pressing across a series of trials and experimental conditions for participants with non-clinical (black) and clinical (red) depression. The lines represent the linear trends in button-pressing probability over trials for each group, with the shaded areas indicating the standard error of the mean (SEM).

Statistical analysis revealed important interactions between trial progression, experimental conditions, and depression groups. Notably, the backward stepwise algorithm revealed that the three-way interaction between trial progression, experimental condition, and depression group did not need to remain in the final model ($p = 0.130$). Instead, it identified the two-way interactions between the following variables should remain in the final model: trial progression and experimental condition ($F(4, 28628) = 5.705$, $p < 0.001$), trial progression and depression group ($F(1, 78) = 17.145$, $p < 0.001$), and experimental condition and depression group ($F(4, 28628) = 11.095$, $p < 0.001$). Comparison with a set of nested models confirmed that the final model provided the best fit.

The significant interaction between trial progression and experimental conditions revealed behavioural variability when participants had different levels of agency in each experimental conditions as trials progressed across all participants. In general, button pressing decreased over time across all conditions. Compared to the control condition, Zero, where neither participant had control, participants slightly increased their button pressing in the Other Negative condition ($\beta = 0.003$, $SE = 0.001$, $t(28628) = 2.190$, $p = 0.029$) and Self Positive condition ($\beta = 0.003$, $SE = 0.001$, $t(28628) = 2.660$, $p = 0.008$) as trials progressed – both conditions where participants rated themselves as having positive control. In contrast, in the Other Positive and Self Negative condition, where participants rated themselves as having negative control, button pressing over time did not significantly differ from the Zero condition. Participants did not adjust their actions to reflect negative control (i.e., their actions were not significantly reduced), behaving instead as though they had no control at all. This suggests that participants were less able to act in line with negative control, implying a limited capacity to adjust behaviour based on a belief that their actions might actively prevent outcomes from occurring.

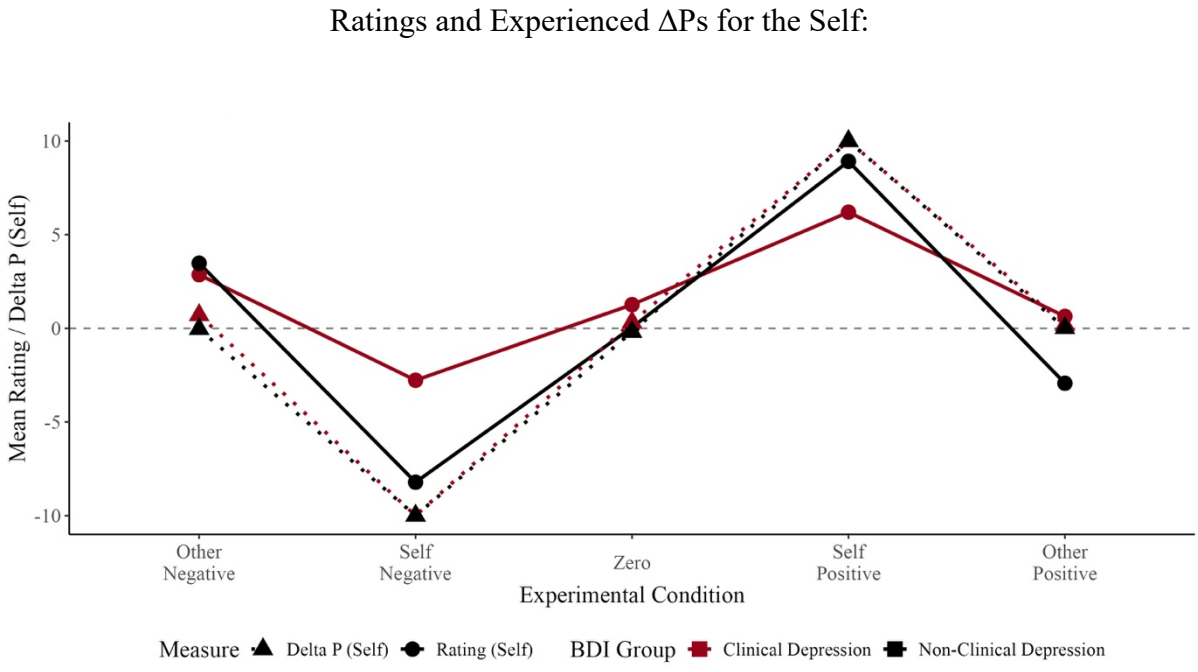
The significant interaction between trial progression and depression groups indicated that baseline probabilities of action across a block of trials differed between the clinical and non-clinical groups. As trials progressed, non-clinically depressed participants reduced their button pressing at a steeper rate compared to clinically depressed participants (β = -0.012, *SE* = 0.002, $t(78) = -4.141$, $p < 0.001$). This pattern is illustrated in Figure 10, where the red lines (representing clinically depressed participants) are consistently higher and relatively flat than the black lines (representing non-clinically depressed participants) which slope downwards across all conditions. The steeper decline in button pressing among non-clinically depressed participants may reflect faster learning of agency, reducing their need to sample the environment through button pressing. In contrast, the more consistent button-pressing behaviour of clinically depressed participants may be due to a slower learning process or a persistent uncertainty about the action-outcome contingencies.

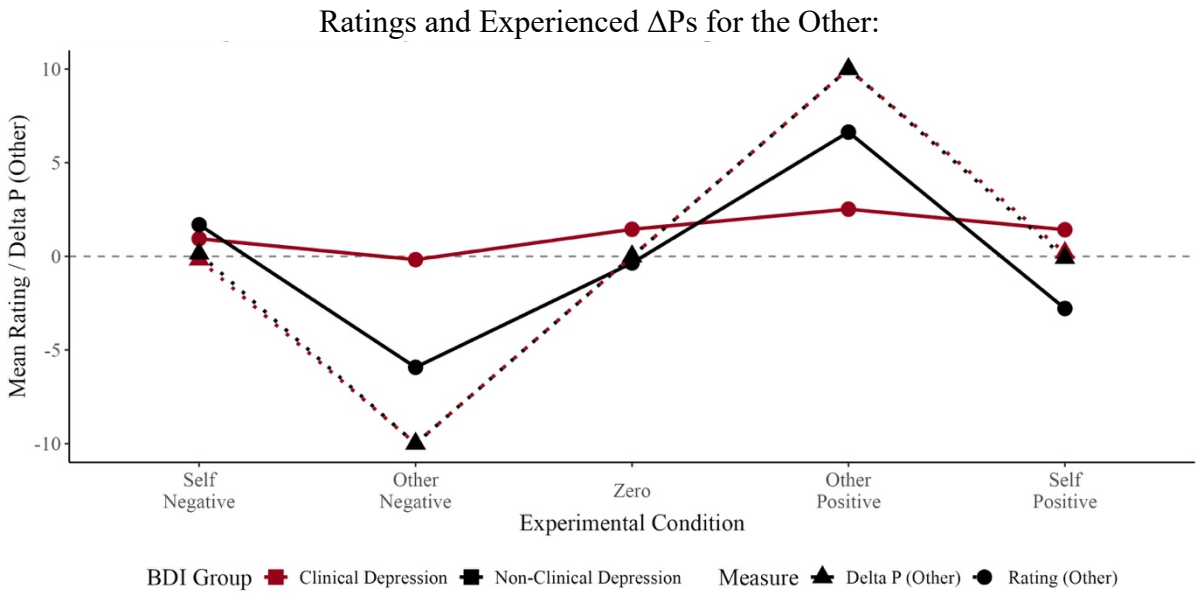
The interaction between experimental condition and depression group suggested that actions differed between the depression groups across agency levels, but not within each condition as trials progressed – this interaction instead reflects overall differences in actions averaged across each block of trials. Post hoc analysis revealed that only in the Self Negative condition did non-clinically depressed participants press the button significantly less than clinically depressed participants (β = -0.093, *SE* = 0.002, $t(28628) = -5.190$, $p < 0.001$). This finding suggests that clinically depressed participants may have difficulty recognising that their actions inhibit outcomes from occurring (negative contingency), leading to reduced goal-directed behaviour.

These findings highlight the complex relationship between depression, agency perception, and behavioural adaptation. Across all experimental conditions, participants pressed the button less as trials progressed, potentially reflecting learning of agency. Non-clinically depressed participants showed a steeper decline in button pressing, suggesting faster

learning of agency. They also demonstrated greater behavioural flexibility, adjusting their actions more readily across different experimental conditions - pressing more when they perceived positive agency and less with negative control. In contrast, clinically depressed participants displayed higher and more consistent button-pressing behaviour across conditions and over time, potentially indicating slower learning. While they also increased button pressing in perceived positive agency conditions, they did not reduce it when they had negative control. These results suggest that depression is associated with reduced behavioural flexibility and difficulty in processing changing environmental action-outcome contingencies, especially negative ones. This implies that depression may impair the ability to recognise and respond appropriately when one's actions inhibit desired outcomes, leading to less effective goal-directed behaviour.

Figure 16.
Results from Experiment 5: Mean Ratings and Experienced ΔPs Across Experimental Conditions, by Depression Group.





Notes. Mean ratings (circles) and experienced ΔPs (triangles) of agency attributed to the Self (top graph) and the Other (bottom graph) across experimental conditions, for participants with non-clinical (black) and clinical (red) depression.

Experienced ΔP. Figure 16 illustrates the mean ratings and experienced ΔPs of agency attributed to the self (top graph) and the other (bottom graph) across various experimental conditions. Generally, the ratings did not consistently align with the experienced ΔPs, as indicated by the circles (ratings) and triangles (experienced ΔPs) not always overlapping. In conditions where the participant had control, ratings tended to align more closely with the experienced ΔPs compared to conditions where the other agent had control. This is illustrated in the top graph (conditions where the self had control) of Figure 16, where the circles (ratings) are closer to the triangles (experienced ΔPs) compared to the bottom graph (conditions where the other had control) where the circles and triangles are further apart.

The experienced action-outcome contingencies (ΔPs) for self and other did not significantly differ between clinically and non-clinically depressed individuals (p = 0.153). This can be seen in Figure 16, where the black and red triangles, representing the experienced ΔPs of the non-clinically and clinically depressed groups respectively, overlap with each other. This means that when they sampled their environment, they generated similar types of

information, resulting in similar experienced objective action-outcome contingencies as calculated by ΔP.

However, significant differences emerged between the depression groups when comparing participants' ratings to their experienced ΔPs across experimental conditions. The discrepancies between ratings and experienced ΔPs for the self was larger in the clinically depressed group compared to the non-clinical depression groups. This is illustrated in Figure 16, where the red circles (ratings by clinically depressed participants) are further away from the red triangles (experienced ΔPs of clinically depressed participants), regressing towards zero, compared to the black points (non-clinically depressed participants).

For example, in the Self Negative condition, clinically depressed participants demonstrated a significantly larger discrepancy between their ratings and experienced ΔPs compared to non-clinically depressed participants (β = -5.443, SE = 1.03, $t(389)$ = -5.280, $p <$ 0.001). Specifically, the clinically depressed group perceived themselves as having less negative agency than the non-clinically depressed group, despite both groups experiencing similar levels of negative contingencies.

Similarly, in the Self Positive condition, the clinically depressed group had a significantly larger difference between ratings and experienced ΔPs (β = 2.713, SE = 1.03, $t(389)$ = 2.632, p = 0.008). Here, clinically depressed participants rated themselves as having less positive agency compared to non-clinically depressed participants, even though both groups experienced comparable positive contingencies.

Interestingly, the pattern was inversed in the Other Positive condition: non-clinically depressed participants exhibited a larger discrepancy between their self-ratings and experienced ΔPs (β = -3.403, SE = 1.03, $t(389)$ = -3.302, p = 0.001). While both groups experienced near-zero contingencies, non-clinically depressed participants rated themselves as having more negative control, whereas depressed participants rated themselves as having zero

control. There were no significant differences in ratings and experienced ΔPs for the self between the depression groups in the Zero or Other Negative conditions.

The difference between ratings and experienced ΔPs for the other participant also varied notably between clinical and non-clinical depression groups. In the Other Negative condition, clinically depressed participants showed a significantly larger discrepancy between their ratings and experienced ΔPs for the other agent (β = -5.749, SE = 1.13, $t(389)$ = -5.092, $p <$ 0.001). The clinically depressed group perceived the other agent as having less negative agency compared to the non-clinically depressed group, despite both groups experiencing similar levels of negative contingencies.

Likewise, in the Other Positive condition, the clinically depressed group demonstrated a significantly larger gap between ratings and experienced ΔPs (β = 4.115, SE = 1.13, $t(389)$ = 3.644, $p < 0.001$). Clinically depressed participants rated the other agent as having less positive agency compared to non-clinically depressed participants, even though both groups experienced comparable positive contingencies.

Conversely, in the Self Positive condition, non-clinically depressed participants showed a larger discrepancy between their ratings and experienced ΔPs for other participants (β = -3.929, SE = 1.13, $t(389)$ = -3.480, $p <$ 0.001). While both groups experienced near-zero contingencies for the other agent, non-clinically depressed participants rated the other as having more negative control. There were no significant differences in ratings and experienced ΔPs for the other between the depression groups in the Zero or Self Negative conditions.

In summary, these findings reveal distinct patterns in how clinically and non-clinically depressed participants differently perceive agency for both self and others across various experimental conditions. While experienced contingencies (ΔPs) did not differ significantly between groups, clinically depressed individuals consistently showed larger discrepancies between their ratings and experienced ΔPs when judging the participant with control,

regardless of whether it was the self or other. This trend was evident in both positive and negative contingency scenarios.

Interestingly, when rating the participant without control in conditions where the participant in control had positive agency, non-clinically depressed participants exhibited an opposing control trend, perceiving the participant without control as having negative control. In contrast, clinically depressed participants did not demonstrate this trend to the same extent, instead showing a bias towards perceiving more positive control for all participants. Consequently, clinically depressed participants had smaller discrepancies between ratings and experienced ΔPs in these conditions, whereas non-clinically depressed participants showed larger discrepancies.

Summary

The experiment explored agency perception in multi-agent contexts, comparing clinically depressed and non-clinically depressed participants. Results revealed differences in how these groups perceived agency and interacted with their environment to learn about the agency of the self and other participant across various conditions.

Non-clinically depressed participants demonstrated greater sensitivity to changing agency levels. Their ratings showed an opposing control pattern, where they correctly attributed agency to the participant with control whilst attributing opposite control to the participant without control. They also exhibited a slight self-serving bias, viewing themselves as having slightly positive control and others as having slightly negative control. This sensitivity was reflected in their behaviour, as they pressed buttons less frequently as trials progressed and adapted their actions based on perceived agency. Their experienced contingencies (ΔPs), which were calculated from the combinations of presence and absence of actions and outcomes (trial types) generated from their actions, aligned more closely with their ratings.

In contrast, clinically depressed participants showed reduced sensitivity to changing agency conditions. They consistently gave lower agency ratings for both self and others, but with a bias towards perceiving more positive control overall for both agents. This positive bias was reflected in their behaviour and difficulty processing negative contingencies. While they increased button pressing in perceived positive agency conditions, they did not reduce it when their actions were preventing outcomes (negative contingencies), demonstrating an impaired ability to adjust their behaviour when their actions inhibited desired results. Interestingly, while their experienced ΔPs did not differ significantly from non-depressed participants, they showed larger discrepancies between their ratings and experienced ΔPs when judging the agent with control, regardless of whether it was self or other. This trend persisted across both positive and negative contingency scenarios, highlighting a consistent misalignment between their positively biased perceptions and the actual contingencies they experienced.

These findings underscore that depression is associated with a reduced but positively skewed perception of agency, coupled with less sensitivity to changes in others' agency and diminished behavioural flexibility. The results suggest that depression may impair the ability to recognise and respond appropriately to changing environmental contingencies, particularly negative ones, potentially leading to less effective goal-directed behaviour.

Active Inference Model Fitting

The active inference model fitting results provided a computational account of the observed behavioural differences between clinically depressed and non-clinically depressed individuals. The model incorporated several free parameters, including priors over agency states, habit to press the button, gamma values (influencing the balance between habitual and goal-directed behaviour), preferences for observing outcomes, likelihood probabilities for observing outcomes given agency states, and the standard deviation for fitting ratings to beliefs. The analysis revealed distinct differences between the clinically and non-clinically depressed

groups in prior beliefs of agency, precision (confidence) of beliefs about action-outcome relationships, outcome preferences, habitual actions, and balance between habitual and goal-directed behaviour; see Table 7.

The model fitting results for prior beliefs over agency states revealed both similarities and differences between the parameters estimated for the depressed and non-depressed groups. The fitted parameters for both groups showed higher values for SPON (Self Positive & Other Negative) and SNOP (Self Negative & Other Positive) states, suggesting a common bias where agency in one direction for one agent implies the opposite direction for the other agent. This finding aligns with the behavioural data, where participants showed the opposing control pattern in their ratings. However, these values were notably higher in the non-clinically depressed group (SPON: 0.368, SNOP: 0.303) compared to the clinically depressed group (SPON: 0.288, SNOP: 0.290), reflecting the amplified opposing control pattern observed in the non-clinical group's behavioural data. This finding aligns with the behavioural results, where non-clinically depressed participants showed a more pronounced opposing control pattern in their ratings.

Other group-specific biases were also observed. The depressed group showed stronger prior beliefs for the SPOP (Self Positive & Other Positive) state, with values of 0.280 for SPOP compared to 0.159 in the non-depressed group. This aligns with the behavioural finding that clinically depressed participants tended to rate both self and other as having similar levels of positive control. In contrast, the non-depressed group exhibited stronger prior beliefs for the SPON state (0.368 compared to 0.288 in the depressed group), consistent with their slight self-serving bias observed in the behavioural data.

Table 7.
Results from Experiment 5: Active Inference Model Fitting Results for Non-Clinically and Clinically Depressed Groups.

Parameter	Non-Clinically Depressed (N=55)	Clinically Depressed (N=25)
Probability of observing outcome present given agency and action states for the likelihood tensor	0.565	0.547
Preference to observe the outcome present	1.689	0.551
Habit to press the button	0.594	0.687
Gamma (weighting for habit)	4.509	3.544
Prior for SPOP state	0.159	0.280
Prior for SPON state	0.368	0.288
Prior for SNOP state	0.303	0.290
Prior for SNON state	0.171	0.142
Standard deviation for ratings regarding self	0.246	0.288
Standard deviation for ratings regarding the other	0.286	0.305

Notes. SPOP refers to Self Positive & Other Positive; SPON refers to Self Positive & Other Negative; SNOP refers to Self Negative & Other Positive; and SNON refers to Self Negative & Other Negative

These differences in prior beliefs led to distinct patterns of agency attribution. The depressed group had a higher prior to attribute more positive control to both self and others. Their attribution of agency to the self, calculated as (SPOP + SPON) – (SNOP + SNON), was 0.136, while their attribution of agency to the other, (SPOP + SNOP) – (SPON + SNON), was 0.140. This closely mirrors the behavioural finding that clinically depressed participants gave similar positive ratings for both self and other. In contrast, the non-clinically depressed group displayed less biased attribution, with a slight tendency to attribute more positive agency to self (0.053) and more negative agency to others (-0.077), consistent with the self-serving bias observed in their behavioural ratings.

The parameter fitting results for the likelihood tensor and rating fit revealed differences in belief precision between the two groups. The parameters fitted to the clinically depressed group showed slightly lower values for the probability of the outcome being present in the

likelihood tensor (0.547 compared to 0.565 for the non-clinically depressed group). This flatter distribution suggests less precise beliefs about the relationship between actions and outcomes in the clinically depressed group, which aligns with the behavioural finding that clinically depressed participants consistently gave lower agency ratings across all conditions.

The model fitting also revealed significant differences in outcome preferences and action habits between the two groups. Non-depressed participants showed a stronger preference to observe the outcome present (1.689) compared to depressed participants (0.551). Regarding button-pressing habits, the depressed group showed a higher habit of pressing the button (0.687) compared to the non-depressed group (0.594). Moreover, the clinically depressed group had a lower gamma value (3.544, compared to 4.509 for the non-clinically depressed group), indicating that their actions were more influenced by habits rather than goal-directed behaviour calculated based on EFE. Together, these finding corresponds with the behavioural data, which showed that clinically depressed participants maintained more and consistent button-pressing behaviour across conditions and over time, showing less behavioural flexibility across different experimental conditions.

To summarise, the active inference model fitting results revealed distinct differences between clinically depressed and non-clinically depressed individuals. Both groups showed priors for opposing control between agents, with a bias for positive control for the self, but differed in their perception of the other agent. The non-clinically depressed group had a prior of the other having slightly negative control, while the clinically depressed group saw the other agent as also having positive control. The positive bias for both agents in the clinically depressed group was notably higher than the self-positive bias in the non-clinically depressed group. Additionally, the clinically depressed group showed less precise beliefs about action-outcome relationships, as indicated by the flatter likelihood distribution in the fitted parameters. Regarding action selection, the parameters fitted to the clinically depressed group indicated

more persistent button presses but less goal-oriented behaviour, suggesting a more aimless pattern of actions. This was evidenced by the higher habit parameter, lower gamma value, and lower outcome preference, indicating less reliance on goal-directed decision-making. In contrast, the non-clinically depressed group's fitted parameters reflected fewer actions over time but more goal-oriented behaviour. These computational findings provide potential explanations for the observed behavioural patterns, including differences in agency ratings, control attributions, and action selection strategies between the two groups.

Discussion

The results of this experiment supported our hypotheses, revealing that participants could accurately discern which agent had control in each condition, assigning higher ratings to the controlling agent in line with the direction of their control. This pattern, consistent across conditions, reaffirmed the presence of the opposing control effect observed previously: participants tended to attribute some degree of control to the non-controlling agent, but in a direction opposite to that of the agent who had actual control. The active inference model corroborated this pattern, highlighting a shared bias in the priors where perceived agency for one agent implied the inverse for the other. These findings suggest that, in multi-agent contexts, individuals perceive each agent's control relative to one another.

Importantly, supporting our hypotheses but contrasting the results from Experiment 4, these patterns varied significantly between the non-clinically and clinically depressed groups. Clinically depressed participants gave lower ratings overall, tending towards zero, which suggests a general reduction in their confidence about the agents' influence. This is reflected in the less precise action-outcome likelihood mappings in the active inference model. This also resulted in a less distinct opposing control pattern.

Additionally, the clinically depressed group demonstrated weakened perceptions of negative contingencies, with ratings skewed towards more positive values even when negative

control was expected. This was consistent with the active inference model's priors, which indicated a bias towards positive agency states for this group, whereas the non-clinically depressed participants showed a slight self-serving bias, perceiving themselves as having greater positive influence while others as having a negative influence.

Contrary to our initial predictions, and consistent with the previous experiment, individuals with clinical depression did not decrease their engagement over time; their button-pressing behaviour remained steady across trials. This suggests that, while they engaged more with their environment, these actions were less purposeful or adaptive, reflecting a more aimless approach to interacting with the environment. In contrast, non-clinically depressed participants reduced their button presses as they learned, reflecting fewer actions that were more targeted and goal-directed, perhaps more intentional. This difference was mirrored in the active inference model: clinically depressed participants showed a stronger habit to press the button and a lower preference for observing outcomes, whereas non-clinically depressed participants demonstrated a higher outcome preference with reduced reliance on habits. This pattern indicates that those with clinical depression may maintain consistent engagement but in a less strategic way, while non-clinically depressed participants adjust their behaviour more deliberately to achieve desired outcomes.

These findings offer important insights into how agency perception may operate in multi-agent contexts, especially among individuals with differing levels of depression. The variability in how control is perceived and acted upon suggests that agency attribution is not only influenced by the presence of other agents but also by the underlying cognitive biases and decision-making tendencies of the learner. This highlights the potential for tailored interventions that could improve agency perception and contingency learning in clinical populations, especially by helping individuals shift from habitual to more goal-directed engagement with their environment.

To further our understanding of how participants learn about their own and others' levels of agency, the next experiment will introduce a running score of outcomes during the trials. This adjustment aims to provide participants with immediate feedback on the consequences of their actions, potentially making the connection between actions and outcomes more explicit. Such real-time scoring could help clarify how participants integrate this feedback into their perception of agency, particularly for those with clinical depression, who may benefit from more tangible indicators of their influence over outcomes. By making the effects of each action more apparent, we can better assess whether individuals with higher depressive symptoms continue to exhibit less adaptive, habitual behaviour or if this direct feedback might encourage a shift towards more goal-directed actions.

Additionally, by gamifying the task through a scoring system, we aim to elicit more goal-directed action patterns from non-clinically depressed participants. In previous experiments, participants' actions tended to decrease over trials, even when they had positive control over outcomes, albeit to a lesser degree. The gamified element is designed to motivate participants to maintain or even increase their actions to achieve higher scores, fostering more strategic behaviour and exploitation of their control.

We also extended the number of trials per condition, allowing participants more time to demonstrate exploitative behaviour. This increase in number of trials in each block also aimed to explore whether the differing behavioural patterns between depression groups across a longer period might lead to variations in experienced ΔPs, as the behavioural trends observed in earlier experiments suggested potential differences that may have been obscured by shorter trial blocks. Moreover, to simulate a real-life learning situation, two experimental conditions were added where the other participant's actions were no longer random (50%) but adjusted based on their control type: either increasing or decreasing to produce the outcome as much as possible. This addition aims to mimic the other participant learning and adapting their

behaviour to maximise their perceived control, providing a richer context for understanding how participants assess their own agency relative to others.

The next section will present the results from the full experiment conducted with non-clinically depressed participants and the pilot study with clinically depressed participants, offering a comparative analysis of how varying levels of feedback and simulated learning conditions influence behaviour across these groups.

Experiment 6

Participants

A total of 90 participants were recruited from three sources to take part in the experiment and were subsequently grouped into two categories for analysis: clinically depressed individuals (11% of the sample) from a clinic in Parma, Italy, and non-clinically depressed individuals (89% of the sample) from the general population in the UK, USA, and Canada, as well as undergraduate students from the University of Oxford. Recruitment for the general population in the UK, USA, and Canada was conducted via Prolific (https://www.prolific.com; Palan and Schitter (2018)), while undergraduate participants were recruited through the Department of Experimental Psychology's internal recruitment platform, Sona Systems.

To ensure adequate attention to the task, participants who provided uniform responses across 90% or more of the task conditions were excluded from analysis; one participant was excluded based on this criterion. The final analysed sample consisted of 89 participants, including 36 males and 53 females. The overall mean age of the final sample was 32.2 years (SD of 12.4). The non-clinically depressed group comprised of 79 participants, 32 of which identified as being male and 47 as being female. The overall mean age of this group was 30.1 years (SD of 10.7). The clinically depressed group comprised of 10 participants, 4 of which

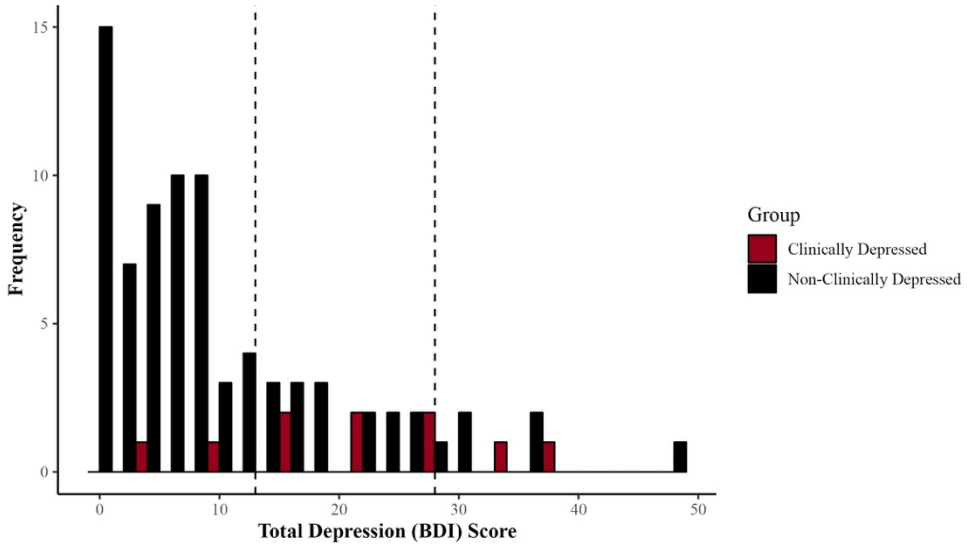
identified as being male and 6 as being female. The overall mean age of this group was 48.8 years (SD of 12.6). See Table 8 for demographic details of the three sample groups.

Participants also completed the BDI-II. Although BDI-II scores were not used to categorise participants into groups in the present experiment, their scores are provided below for reference. The mean BDI-II score among participants was 11.494 (SD = 10.738), with scores ranging from 0 to 48. According to BDI-II categorisation (Beck et al., 1996), around 67.41% of all participants fell into the minimal depression range (scores of 0–13; N=60), 23.60% into the mild-to-moderate range (scores of 14–28; N=21), and 8.99% into the severe range (scores of 29–63; N=8). Mean BDI-II scores were 22.300 (SD = 10.478) for the clinically depressed group and 10.127 (SD = 10.027) for the non-clinically depressed group (see Figure 17 for the distribution of BDI-II scores between the two depression groups).

Table 8.
Experiment 6: Demographic Information from the Three Sample Groups.

Sample Source	In-patient depressed individuals from a clinic in Parma, Italy	Not clinically depressed individuals from Prolific	Not clinically depressed individuals from the University of Oxford, UK
Total number of participants	10	61	18
Number of male participants	4	29	3
Number of female participants	6	32	15
Mean age	48.8	33.4	19.0
Standard deviation of age	12.6	10.0	0.7
Mean BDI-II score	22.3	8.1	16.9
Standard deviation of BDI-II score	10.5	8.3	12.4

Figure 17.
Experiment 6: Distribution of Total Depression (BDI-II) Scores for Clinically and Non-Clinically Depressed Groups.



Notes. This figure displays the distribution of total depression scores as measured by the Beck Depression Inventory (BDI-II; Beck et al. (1996)) for two groups: clinically depressed (in red) and non-clinically depressed (in black). The x-axis represents the total BDI-II scores, ranging from 0 to 50, while the y-axis shows the frequency of participants within each score range. The two dashed vertical lines indicate the BDI-II categorisation thresholds: scores of 0–13 fall within the minimal depression range, scores of 14–28 correspond to the mild-to-moderate depression range, and scores of 29 and above fall into the severe depression range.

Materials, Procedure, and Task Design

At the beginning of the experiment, participants were presented with a briefing sheet detailing the task, data usage and storage, contact information for queries, and their right to withdraw. They provided informed consent and demographic information (age and gender) and confirmed they did not have a history of seizures or diagnosed motor or cognitive impairment before completing the computer task.

Participants completed the same task as the one outlined in the previous experiment. Briefly, in the task, in each block of trials, participants freely produced actions, observed another (simulated) participant's actions, integrated this information with the absence or presence of outcomes, and provided ratings about how much control both agents had, reflecting their learning of action-outcome contingencies (Figure 8).

However, in this experiment, there were three key differences: (i) there was a count displayed on the screen for the number of outcomes produced across trials in a block of trials to gamify the task so that participants had to increase their score, (ii) the number of trials increased from 24 to 40, and (iii) two experimental conditions were added where the other participant's actions were no longer random (50%) but rather increased or decreased according to the type of control they had to simulate the other participant learning how much control they had and adjusting their behaviour to produce the outcome as much as possible. This resulted in seven conditions: the control condition being Zero, where no one has control, and the six experimental conditions being Self Positive, Self Negative, Other Positive, Other Positive Learning, Other Negative, Other Negative Learning. When one agent had control, the other had none. The contingencies were deterministic, such that when an agent had control, they had 100% control. Participants completed a total of 7 blocks of trials (plus one practice condition). Participants were allowed a maximum of two five-minute breaks between blocks of trials. The whole experiment took approximately 15 minutes to complete.

Upon completing the task, participants were asked to fill out the BDI-II to assess the presence and severity of depressive symptoms. After completing the depression questionnaire, participants were asked to identify a particular shape from an array presented to them. Participants were then presented with a debrief sheet, outlining the objective of the experiment.

Data Analysis

The same analytical approach as in Experiment 5 was employed for Experiment 6, with a few key differences in the variables and model specifications.

Comparing Group Means

Agency Ratings

The analysis began with fitting a full linear mixed model (LMM), similar to Experiment 4 and 5. The full LMM comprised the experimental condition, rating category (self or other),

depression group, position of participant's button (left or right), and their interactions. The model refinement process (using a backward stepwise algorithm), examination of significant predictors, and post-hoc analyses (including pairwise comparisons) followed the same procedure as in Experiment 4 and 5.

Actions

The analysis of actions (button presses) mirrored that of Experiment 4 and 5, examining the influence of trial progression, experimental conditions, and depression group on participants' behaviours. Additionally, as in Experiment 4 and 5, the participants' experienced contingency was calculated using ΔP based on the trial types they produced, and difference scores between ratings and ΔP for both self and other were computed. These difference scores were examined across depression groups under different experimental conditions using LMM and post-hoc pairwise comparisons, following the same analytical approach as in Experiment 4 and 5.

Active Inference Model Fitting

The active inference modelling approach for this experiment was consistent with that of Experiment 4 and 5, maintaining the same structure for the generative model but editing the generative process to reflect the changes made in the task design in this experiment. The comparison was made between clinically depressed and non-clinically depressed groups.

The model fitting process employed the same methods as in Experiment 4 and 5, with posterior samples analysed to compare parameters between the clinically depressed and non-clinically depressed groups, aiming to identify potential cognitive differences between these distinct groups.

Results

Comparing Group Means

Agency Ratings

Agency ratings for both self and other varied significantly across experimental conditions and depression groups, indicating that participants' perceptions of agency were sensitive to which agent had control. This sensitivity differed between clinically depressed and non-clinically depressed participants. The backward stepwise algorithm supported these observed behavioural trends, identifying that interactions between experimental conditions, rating category (self or other), and depression group (clinical or non-clinical) should remain in the final mode ($F(6, 1131) = 2.774$, $p = 0.011$). The position of the participant's button (left or right) did not influence ratings. The final model was validated against a set of nested models defined hierarchically, confirming it as the best-fitting model. Due to violations in model assumptions (see Supplementary Material 6C), robust standard errors were applied, and the interaction effects remained significant. It is important to note that the clinically depressed group had a small sample size, which may affect the clarity of their behavioural trends during statistical analyses, though these trends are depicted in the graphs.

Participants generally identified which agent had control in each condition, as ratings for the self and other varied significantly across experimental conditions. As illustrated in Figure 18, participants rated the controlling agent higher than the non-controlling one, aligning with the correct direction. Although the clinically depressed group did not provide statistically significantly lower ratings than the non-clinically depressed group, the small sample size should be considered. The visualisation in Figure 18 suggests that ratings from clinically depressed participants are closer to zero, indicating a trend towards lower ratings, even though the difference is not yet statistically significant.

Ratings by non-clinically depressed participants for both self and other varied significantly across experimental conditions. For self-ratings, significant differences were observed between the Zero condition and Self Positive (β = -7.848, SE = 0.748, t(1131) = -10.490, $p < 0.001$), Self Negative (β = 7.709, SE = 0.748, t(1131) = 10.304, $p < 0.001$), Other Positive (β = 3.646, SE = 0.748, t(1131) = 4.873, $p < 0.001$), Other Positive Learning (β = 3.747, SE = 0.748, t(1131) = 5.008, $p < 0.001$), Other Negative (β = -3.734, SE = 0.748, t(1131) = -4.991, $p < 0.001$), and Other Negative Learning (β = -4.835, SE = 0.748, t(1131) = -6.463, $p < 0.001$) conditions. Ratings for the other also differed significantly between the Zero condition and Self Positive (β = 2.797, SE = 0.748, t(1131) = 3.739, p = 0.014), Other Positive (β = -6.228, SE = 0.748, t(1131) = -8.325, $p < 0.001$), Other Positive Learning (β = -6.987, SE = 0.748, t(1131) = -9.340, $p < 0.001$), Other Negative (β = 5.848, SE = 0.748, t(1131) = 7.817, $p < 0.001$), and Other Negative Learning (β = 5.848, SE = 0.748, t(1131) = 7.817, $p < 0.001$) conditions, but not Self Negative (p = 0.996). Clinically depressed participants' ratings for the self and other did not yet show significant differences across conditions.

Interestingly, the non-clinically depressed group gave similar ratings in conditions where the other agent behaved randomly versus when simulating learning behaviour by acting more optimally over time. However, clinically depressed participants tended to give lower ratings to the random behaviour condition compared to the learning conditions. For instance, in the Other Negative Learning condition versus the Other Negative condition, non-clinically depressed participants rated the other's agency as -5.620 (0.582) and -5.620 (0.557), respectively, whereas clinically depressed participants rated the other's agency as -5.100 (1.320) and -2.000 (1.220).

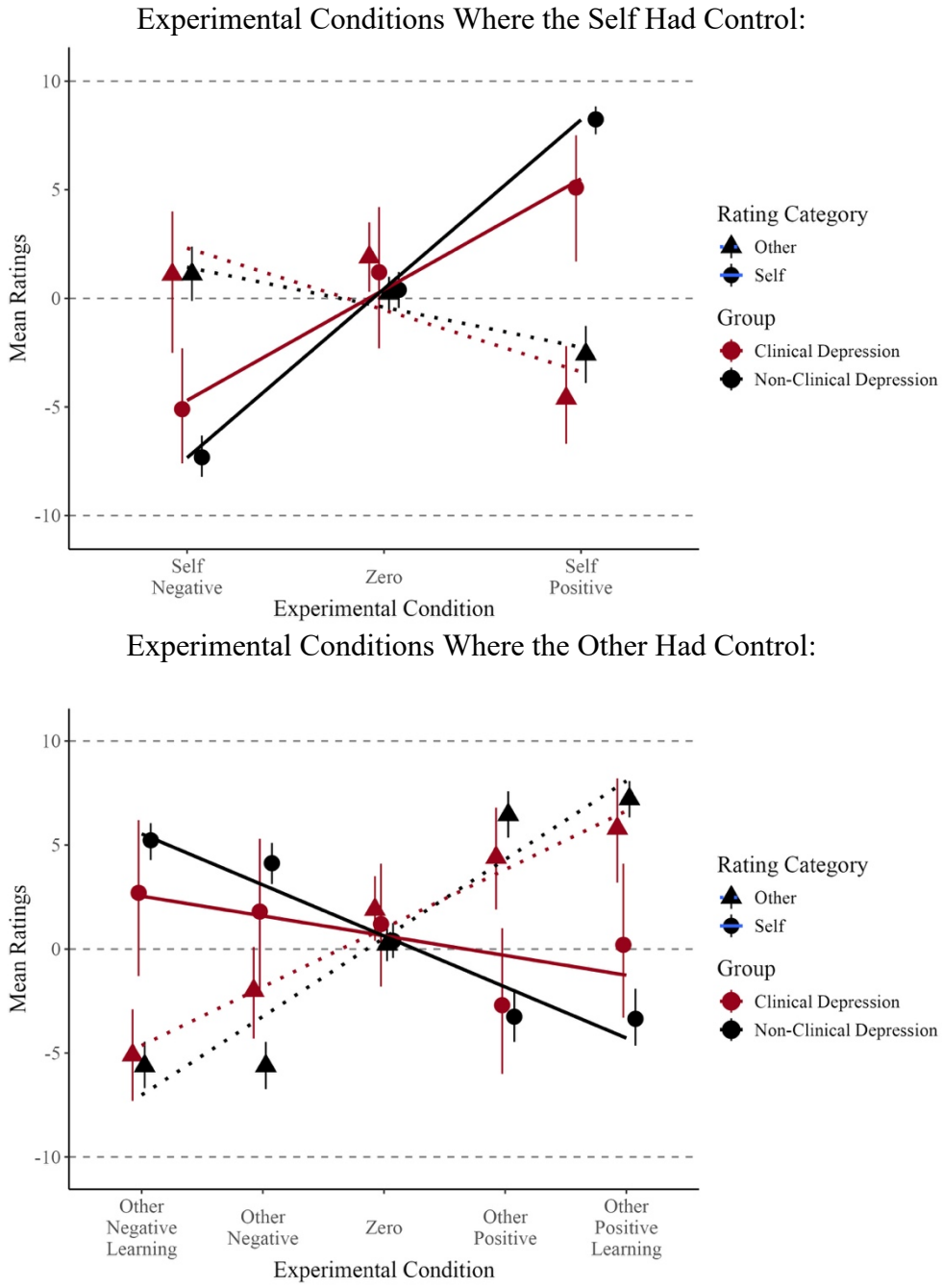
Additionally, non-clinically depressed participants attributed agency to the controlling participant accurately (i.e., in the correct positive or negative direction) but incorrectly attributed agency in the opposite direction to the non-controlling participant. For example, in

the Self Positive condition, non-clinically depressed participants correctly perceived the self as having positive control (M = 8.241, SE = 0.336), but attributed negative control to the other participant (M = -2.570, SE = 0.656), despite the other having no control. This negative attribution was significantly stronger than in the Zero condition (β = 2.797, SE = 0.748, $t(1131)$ = 3.739, p = 0.014), confirming that non-clinically depressed participants attributed negative control to the other, rather than perceiving them as having zero control as they did in the Zero condition.

For clinically depressed participants, this difference did not reach significance. This pattern was less pronounced among clinically depressed participants. In the Self Positive condition, the self was correctly rated as having positive control (M = 5.100, SE = 1.676), while the other participant was rated as having slightly negative control (M = -4.600, SE = 1.301). However, this rating was not significantly different from the rating for the self in the Zero condition (p = 0.110). The visualisation in Figure 18 suggests that the trend for clinically depressed participants' ratings of the non-controlling participant moving away from zero.

Overall, both groups perceived the self as having more positive control compared to the other. Averaging across all conditions, clinically depressed participants rated the other participant at 0.214 (SE = 0.675) and themselves at 0.457 (SE = 0.768). In comparison, non-depressed participants rated the other participant at 0.172 (SE = 0.294) and themselves at 0.580 (SE = 0.294) for the for the self. These differences were not statistically significant (p = 0.699).

Figure 18.
Results from Experiment 6: Ratings Across Experimental Conditions Where the Self Had Control, by Depression Group.



Notes. Mean ratings of agency attributed to the Self (circles; solid lines) and the Other (simulated participant; triangles; dotted lines) across experimental conditions. The top panel shows conditions where the Self had control, while the bottom panel shows conditions where the Other had control. Ratings are displayed for participants that were non-clinically depressed (black) and clinically depressed (red). Each point represents the mean rating for a given condition, with error bars indicating 95% confidence intervals.

Actions

Pressing the Button over Time. Overall, the probability of pressing the button varied across trials, influenced by both the experimental conditions and the depression group, as shown in Figure 19. Across conditions, participants adjusted their button-pressing behaviour over time, sometimes increasing and sometimes decreasing their actions as trials progressed. Consistent with the previous two experiments, non-clinically depressed participants showed greater responsiveness to changes in experimental conditions, adjusting their actions more dynamically across trials. Clinically depressed participants, however, exhibited less sensitivity to these changes, maintaining relatively stable behaviour. Notably, unlike the previous experiments, clinically depressed participants pressed the button less frequently than non-clinically depressed participants across all conditions.

Statistical analysis identified key interactions between trial progression, experimental conditions, and depression groups. The backward stepwise algorithm determined that the three-way interaction between these variables was not necessary for the final model ($p = 0.651$). Instead, the analysis highlighted the importance of two-way interactions between trial progression and experimental condition ($F(6, 24733) = 7.905$, $p < 0.001$), experimental condition and depression group ($F(6, 24733) = 11.750$, $p < 0.001$), and trial progression and depression group ($F(1, 87) = 4.056$, $p = 0.047$). A comparison with nested models confirmed that this final model provided the best fit.

The significant interaction between trial progression and experimental conditions revealed variability in participants' behaviour as they adjusted to differing levels of agency over time. Across all conditions, participants did not consistently reduce their button-pressing as trials progressed; instead, they sometimes increased their actions. This pattern differed from the previous experiments, where participants more consistently reduced their button presses over time. When considering both non-clinically and clinically depressed participants together,

the differences in their button-pressing behaviour were not statistically significant. The small sample size of the clinically depressed group, due to piloting this experiment on that sample, may have introduced variability, making it difficult to detect statistical differences despite clear visual trends. To address this, we analysed the effect separately by group.

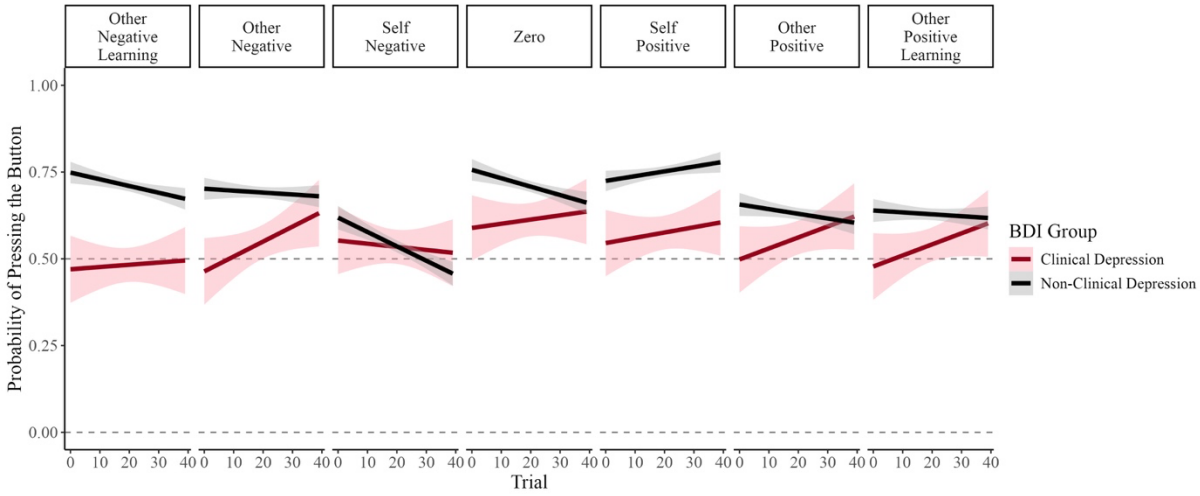
Non-clinically depressed participants showed a significant increase in button-pressing in the Self Positive condition compared to the Zero condition ($\beta = 0.003$, $SE = 0.001$, $t(21960) = 3.650$, $p < 0.001$), with the upward trend evident in Figure 19. Unlike the previous two experiments, this trend was distinctly positive rather than simply showing a less pronounced decrease. The trend was cubic ($\beta = 0.001$, $SE = 0.001$, $t(21960) = 3.457$, $p < 0.001$), characterised by an initial decrease in actions, followed by a reversal to an increase, before reaching a plateau towards the end of the trials. In contrast, non-clinically depressed participants pressed the button significantly less in the Self Negative condition compared to the Zero condition ($\beta = -0.002$, $SE = 0.001$, $t(21960) = -2.426$, $p = 0.015$). This pattern also showed a cubic trend ($\beta = -0.005$, $SE = 0.001$, $t(21960) = -2.510$, $p = 0.012$), characterised by an initial decline in button-pressing, followed by a slight increase, and then a levelling off, before trending downwards again, suggesting that participants adjusted their responses as they adapted to the Self Negative condition. In conditions where participants had no actual control, their behaviour did not differ significantly from the Zero condition, despite their ratings suggesting a perception of slight control. This suggests that their actions aligned more with a perception of having no control.

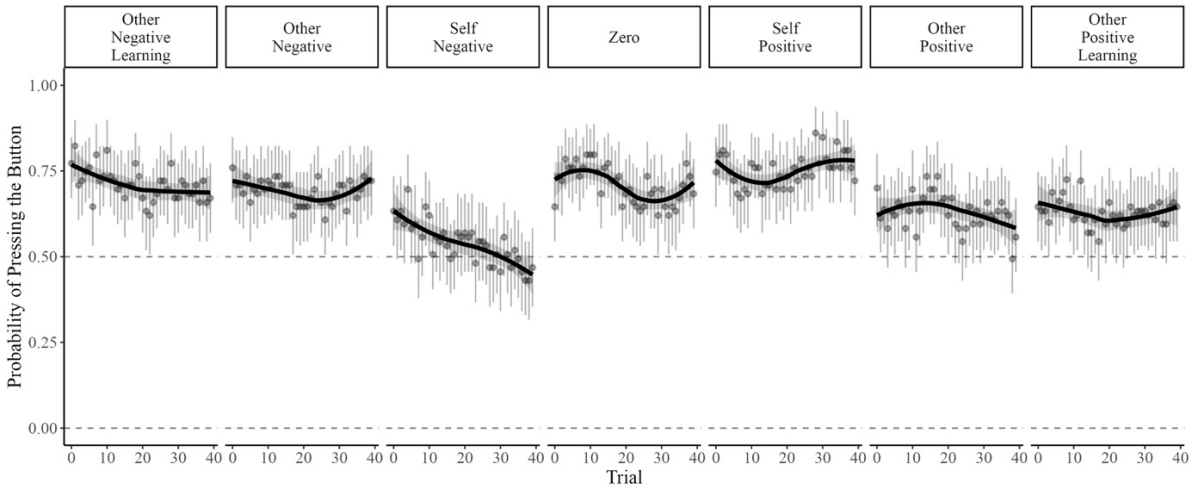
While the clinically depressed participants did not show statistically significant differences in behaviour between conditions (likely due to the small sample size), visual inspection indicated varying trends across conditions. For instance, there were increases in button-pressing in conditions such as Self Positive, Other Negative, and Other Positive

Learning, where participants seemed to perceive some positive agency as indicated by their ratings.

The significant two-way interactions between experimental condition and depression group ($F(6, 24733) = 11.750$, $p < 0.001$) and trial progression and depression group ($F(1, 87) = 4.056$, $p = 0.047$) highlighted differences in action patterns across trials between clinical and non-clinical groups. Importantly, and contrary to previous experiments, non-clinically depressed participants consistently pressed the button more than clinically depressed participants ($\beta = 0.02$, $SE = 0.008$, $t(14720) = 3.409$, $p < 0.001$). This can be seen in Figure 19, where the black lines representing the non-clinically depressed group are generally above the red lines representing the clinically depressed group across conditions. This effect does not occur in the Self Negative condition, where the lines intersect, suggesting that, as in previous experiments, clinically depressed participants may have difficulty recognising when their actions inhibit outcomes (negative contingency), which could contribute to reduced goal-directed behaviour.

Figure 19.
Results from Experiment 6: Button-Pressing Probability Across Trials by Depression Group.



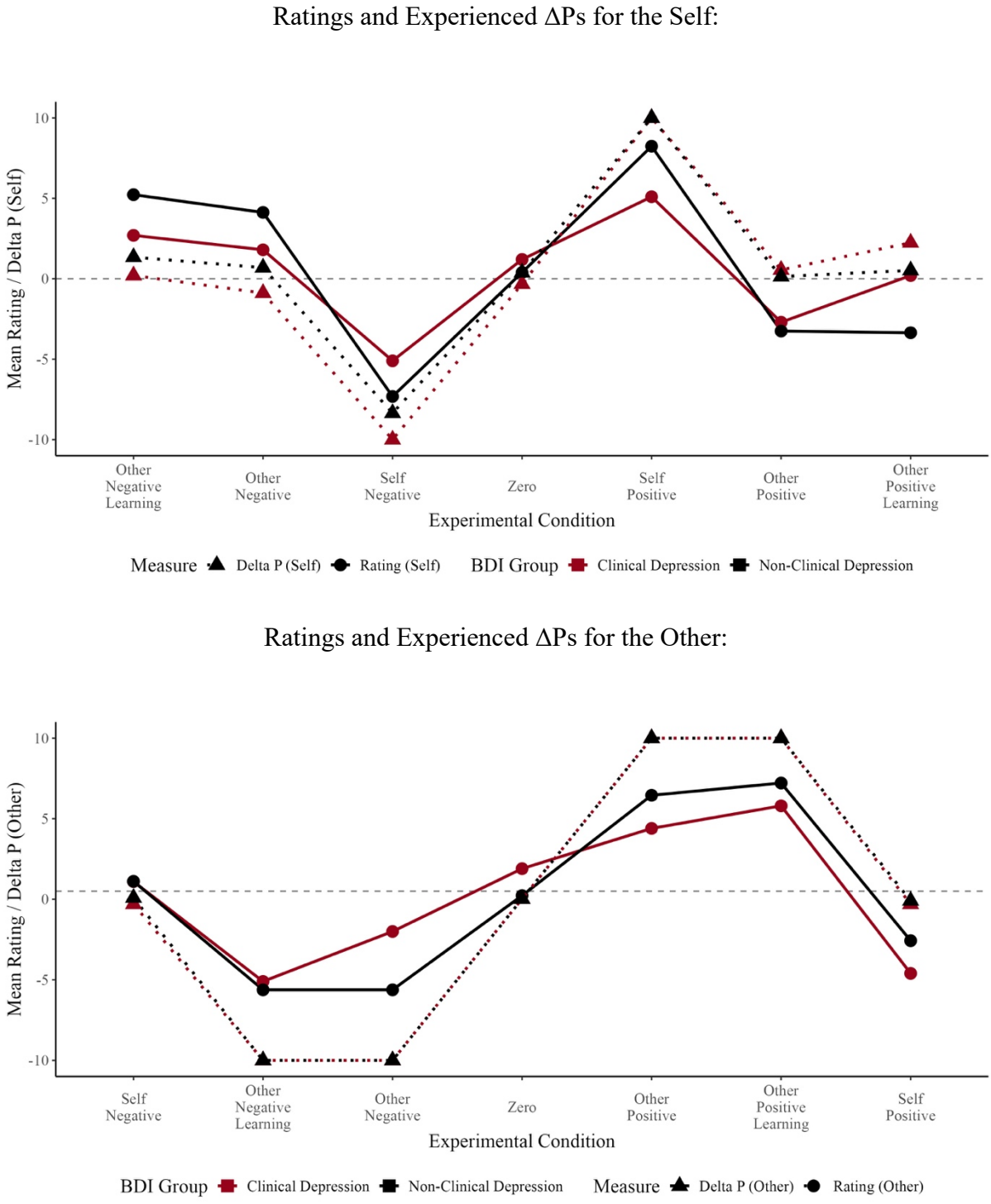


Notes. Probability of button pressing across a series of trials and experimental conditions for participants with non-clinical (black) and clinical (red) depression. The lines represent the linear trends in button-pressing probability over trials for each group, with the shaded areas indicating the standard error of the mean (SEM).

Experienced ΔP.

Figure 20 illustrates the mean ratings and experienced ΔPs of agency attributed to the self (top graph) and the other (bottom graph) across various experimental conditions. Generally, the ratings did not align with the experienced ΔPs, as indicated by the circles (ratings) and triangles (experienced ΔPs) not always overlapping. In conditions where the participant had control, ratings tended to align more closely with the experienced ΔPs compared to conditions where the other agent had control. This is illustrated in the top graph (conditions where the self had control) of Figure 20, where the circles (ratings) are closer to the triangles (experienced ΔPs) compared to the bottom graph (conditions where the other had control) where the circles and triangles are further apart.

Figure 20.
Results from Experiment 6: Ratings and Experienced ΔPs Across Experimental Conditions, by Depression Group.



Notes. Mean ratings (circles) and experienced ΔPs (triangles) of agency attributed to the Self (top graph) and the Other (bottom graph) across experimental conditions, for participants with non-clinical (black) and clinical (red) depression.

Contrary to the previous two experiment, the experienced action-outcome contingencies (ΔPs) for self significantly differed between clinically and non-clinically depressed individuals, while the experienced ΔP for the other agent remained similar across all conditions. The difference was significant in the Other Positive Learning condition (β = -3.364, SE = 1.292, t(522) = -2.604, p = 0.009), with the experienced ΔP being more positive for the clinically depressed group than the non-clinically depressed group. The difference in experienced ΔPs was not statistically significant in the Other Negative Learning, Other Negative, and Self Negative conditions, however, the difference was visibly clear in Figure 20 (top graph), where the experienced ΔPs were more negative for the clinically depressed group than the non-clinically depressed group.

Moreover, significant differences emerged when comparing participants' ratings to their experienced ΔPs across experimental conditions. The discrepancies between ratings and experienced ΔPs for the self was larger in the clinically depressed group compared to the non-clinical depression groups. This is illustrated in Figure 20, where the red circles (ratings by clinically depressed participants) are further away from the red triangles (experienced ΔPs of clinically depressed participants), regressing towards zero, compared to the black points (non-clinically depressed participants).

For example, clinically depressed participants demonstrated a significantly larger discrepancy between their ratings and experienced ΔPs compared to non-clinically depressed participants in the Self Positive condition (β = -4.660, SE = 2.306, t(522) = -2.021, p = 0.043) and in the Self Negative condition (β = -3.862, SE = 2.306, t(522) = -2.246, p = 0.025). Specifically, the clinically depressed group perceived themselves as having less positive and negative agency, respectively, than the non-clinically depressed group, despite both groups experiencing similar levels of contingencies as calculated by ΔP. In other words, ratings were generally lower (closer to zero) compared to the experienced ΔPs.

The difference between ratings and experienced ΔPs for the other participant also varied between clinical and non-clinical depression groups, but only in the Other Negative condition. Generally, across all conditions, the difference between ratings and experienced ΔPs for the other agent remained relatively similar between non-clinically and clinically depressed participants. The clinically depressed consistently perceived lower agency for the other agent, despite both groups experiencing similar levels of contingencies as calculated by ΔP. In other words, ratings were generally lower (closer to zero) compared to the experienced ΔPs, as clearly seen in Figure 20 (bottom graph) where the red line is closer to zero than the black line. However, in the Other Negative condition, clinically depressed participants showed a significantly larger discrepancy between their ratings and experienced ΔPs for the other agent (β = -3.620, SE = 1.64, t(609) = -2.207, p = 0.028). The clinically depressed group perceived the other agent as having less negative agency compared to the non-clinically depressed group, despite both groups experiencing similar levels of negative contingencies.

In summary, these findings reveal distinct patterns in how clinically and non-clinically depressed participants differently perceive agency for both self and others across various experimental conditions. While experienced contingencies (ΔPs) generally did not differ significantly between groups, clinically depressed individuals consistently showed larger discrepancies between their ratings and experienced ΔPs when judging the participant with control, regardless of whether it was the self or other. This trend was evident in the negative contingency scenarios.

Active Inference Model Fitting

The active inference model fitting results provided a computational account of the observed behavioural differences between clinically depressed and non-clinically depressed individuals. The analysis revealed distinct differences between the clinically and non-clinically

depressed groups in prior beliefs of agency, outcome preferences, habitual actions, and balance between habitual and goal-directed behaviour; see Table 9.

Table 9.
Results from Experiment 6: Active Inference Model Fitting Results for Non-Clinically and Clinically Depressed Groups.

Parameter	Non-Clinically Depressed (N=79)	Clinically Depressed (N=10)
Probability of observing outcome present given agency and action states for the likelihood tensor	0.514	0.509
Preference to observe the outcome present	5.364	0.100
Habit to press the button	0.732	0.629
Gamma (weighting for habit)	7.966	1.346
Prior for SPOP state	0.207	0.258
Prior for SPON state	0.297	0.242
Prior for SNOP state	0.295	0.243
Prior for SNON state	0.201	0.257
Standard deviation for ratings regarding self	0.096	0.059
Standard deviation for ratings regarding the other	0.094	0.060

Notes. SPOP refers to Self Positive & Other Positive; SPON refers to Self Positive & Other Negative; SNOP refers to Self Negative & Other Positive; and SNON refers to Self Negative & Other Negative

The model fitting results for prior beliefs over agency states revealed that the non-clinically depressed group showed higher values for SPON (0.297) and SNOP (0.295) states, suggesting a common bias where agency in one direction for one agent implies the opposite direction for the other agent. This finding aligns with the behavioural data, where participants showed the opposing control pattern in their ratings. However, the clinically depressed group had relatively flat priors, with a slightly inverse distribution compared to the non-clinically depressed group as the prior values for SPON and SNOP were slightly lower than the SPOP and SNON prior values. This suggests that the opposing control bias is not as strong in the clinically depressed group and, corroborating with the previous experiment's findings, that they tend to view both agents as having control in the same direction.

However, the attribution of agency to the self, calculated as (SPOP + SPON) – (SNOP + SNON), and to the other, (SPOP + SNOP) – (SPON + SNON), were zero for both groups, suggesting no prior bias towards the self or other.

The parameter fitting results for the likelihood tensor and rating fit revealed slight differences in belief precision between the two groups. The parameters fitted to the clinically depressed group showed slightly lower values for the probability of the outcome being present in the likelihood tensor (0.509 compared to 0.514 for the non-clinically depressed group). This flatter distribution suggests less precise beliefs about the relationship between actions and outcomes in the clinically depressed group, which aligns with the behavioural finding that clinically depressed participants consistently gave lower agency ratings across all conditions. The large variance caused by the small sample size in the clinically depressed group may have resulted in the less distinct precision values between clinically and non-clinically depressed groups.

The model fitting also revealed significant differences in outcome preferences and action habits between the two groups, similar to the findings from the previous two experiments. Non-depressed participants showed a stronger preference to observe the outcome present (5.364) compared to depressed participants (0.100). Regarding button-pressing habits, the depressed group showed a lower habit of pressing the button (0.629) compared to the non-depressed group (0.732). However, the clinically depressed group had a lower gamma value (1.346), compared to 7.699 for the non-clinically depressed group, indicating that their actions were more influenced by habits rather than goal-directed behaviour calculated based on EFE. Together, these finding corresponds with the behavioural data, which showed that clinically depressed participants maintained more and consistent button-pressing behaviour across conditions and over time, showing less behavioural flexibility across different experimental conditions.

Discussion

The results of this experiment supported our hypotheses and aligned with findings from previous research, showing that participants accurately identified which agent had control in each condition. Participants consistently assigned higher ratings to the controlling agent according to the direction of their control. This pattern, observed across conditions, reaffirmed the presence of the previously identified opposing control effect: participants attributed some degree of control to the non-controlling agent, but in a direction opposite to that of the agent with actual control. These rating patterns were more pronounced among the non-clinically depressed participants, while clinically depressed participants exhibited a weaker sensitivity to variations in agency and a less distinct tendency to perceive opposing control between agents.

The active inference model further corroborated these findings, illustrating a difference in prior beliefs (i.e., biases) between groups. For the non-clinically depressed group, perceived agency for one agent implied a corresponding inverse perception for the other, reflecting a clear contrast between agents. In contrast, the clinically depressed group exhibited a more uniform prior distribution, with a slightly higher tendency to attribute agency to both agents in the same direction. This suggests that in multi-agent contexts, the perception of each agent's control is influenced by depressive symptoms, affecting how control is attributed relative to other agents. Additionally, the action-outcome likelihood mappings in the active inference model were less precise for the clinically depressed group, suggesting a broader uncertainty in their learning and perception of control.

Notably, in the experimental conditions where we simulated adaptive behaviour for the other participant (e.g., increased button-pressing in the Other Positive Learning condition to maximise outcomes), we observed a significant difference in ratings between the clinically depressed and non-clinically depressed groups. While non-clinically depressed participants' ratings remained stable regardless of whether the other agent's behaviour appeared adaptive,

those with clinical depression showed sensitivity to these changes, providing higher ratings in the conditions where adaptive behaviour for the other agent as simulated. This implies that non-clinically depressed participants may rely less on the behaviour of others to determine their agency levels, whereas clinically depressed individuals appear more influenced by how others interact with the environment.

Contrary to earlier findings but consistent with our initial hypotheses, individuals with clinical depression engaged less with the environment over time, while non-clinically depressed participants showed increased engagement. Both groups, however, demonstrated more goal-directed behaviour compared to the previous experiment, likely influenced by the gamified task element that presented a running score of outcomes. This suggests that providing real-time feedback may be a promising intervention to enhance goal-directed actions, although it is unclear whether these would improve perceptions.

Nevertheless, the clinically depressed group's engagement remained steadier and less adaptive than that of the non-clinically depressed group. To elaborate, the non-depressed participants initially reduced their actions, then increased their button-pressing as they explored and learned the relationship between actions and outcomes (i.e., exploring consequences of action and inaction), eventually shifting towards a more exploitative approach to maximise outcomes. This contrasts with the clinically depressed group's more consistent yet less strategic behaviour throughout the trials.

These behavioural trends were mirrored in the active inference model: non-clinically depressed participants demonstrated a stronger habit tendency for button-pressing, alongside a higher preference for observing outcomes, but they showed reduced reliance on habitual responses. Conversely, clinically depressed participants displayed a lower preference for outcomes and a greater reliance on habitual actions, though the magnitude of the habit was slightly lower than that of their non-clinically depressed counterparts. This suggests that

individuals with clinical depression may struggle to shift from habitual to more goal-oriented strategies, while those without clinical depression adjust their actions more deliberately to achieve desired outcomes.

These findings provide valuable insights into how agency perception functions in multi-agent scenarios, particularly among individuals with varying levels of depression. The differences in how control is perceived and learned via actions suggest that agency attribution is shaped not only by the actions of others but also by individual cognitive biases and decision-making styles. The gamification element and adaptive behaviour simulations revealed important differences between the two groups in learning agency, suggesting that targeted feedback and an understanding of others' actions may be key to improving agency perception in those with higher depressive symptoms. The findings from this experiment highlight the potential for tailored interventions that could enhance agency perception and contingency learning in clinical populations by helping individuals transition from habitual to more goal-directed engagement with their surroundings.

The next chapter will provide a more comprehensive discussion of the findings from these three experiments (Experiments 4-6) in the context of the overall research aims of this thesis. It will explore how the results from this experiment align with those from previous studies, and the other experiments reported in this thesis (Experiments 1-3), integrating them into a broader understanding of how agency perception and action-outcome contingencies are processed across varying levels of depressive symptoms. This general discussion will also address the theoretical implications of these findings, the limitations of the current research, and potential directions for future studies to advance our understanding of agency learning in multi-agent contexts.

Chapter 7:

General Discussion

This thesis aimed to elucidate the intricate processes underlying agency perception, particularly within multi-agent contexts, and to explore how these processes might be altered in individuals experiencing depressive symptoms. It particularly focuses on how individuals learn the extent of their control over outcomes in their environment via action-outcome contingency learning. The importance of agency in human behaviour and cognition cannot be overstated, as it plays a crucial role in exploratory behaviour, flexibility, goal-directed actions, and psychological well-being (DeNeve & Cooper, 1998; Penton et al., 2018; Verschure et al., 2014). Our research has provided valuable insights into how individuals learn about their agency, particularly in multi-agent contexts where the interplay between self and other becomes crucial.

Our initial experiments (Experiments 1-3) focused on the perceptual aspects of contingency learning, examining how the frequency and content of **D** events, the co-absence of target predictor and target outcome) influence perceived contingencies. Contrary to predictions derived from statistical and associative learning models, we found that increasing **D** trials consistently reduced perceived contingencies across all baseline conditions (negative, zero, and positive). This effect was particularly pronounced when a separate, non-target cue and outcome relation were presented during the **D** trials. These findings challenge existing statistical and associative learning models, which struggle to account for the observed reduction in perceived contingencies due to **D** events. Instead, our results lend support to inferential reasoning models, wherein how the causal structure of the environment is interpreted determines how **D** events impact the target predictor-outcome contingency.

A consistent finding across these experiments was that individuals with higher levels of depressive symptoms showed diminished processing of **D** events, particularly in trials without non-target stimuli. This builds upon previous research indicating reduced processing of inter-trial intervals in depression, which are analogous to **D** events. However, inter-trial intervals and **D** events have distinct properties pertaining to the regularity of presentation and

durations of presentation compared to the other trials in the block. Our study is thereby the first to show this reduced processing in a paradigm that directly manipulates **D** events, suggesting a more generalised difficulty in processing absent cue-outcome relations among those with higher depressive symptoms.

Furthermore, we observed that individuals with higher depressive symptoms demonstrated a specific difficulty in processing negative contingencies. This finding extends previous research on action-outcome contingencies to more general cue-outcome relationships, indicating a broader deficit in processing inhibitory relationships in depression. This means that they struggle to process when their actions are preventing the outcome from occurring and when their lack of action is what actually produces the outcome. This difficulty in perceiving negative contingencies could have significant implications for how individuals with depression navigate their social and physical environments, potentially contributing to maladaptive behaviours or cognitive patterns. For instance, a university student with depression might continue to isolate themselves in their room, failing to recognise that this does not improve their mood and academic performance, leading to vicious cycles of counterproductive behaviour.

Experiments 4-6 expanded our investigation to explore agency perception in multi-agent contexts, where participants interacted with a simulated agent. These experiments revealed several key insights into how individuals perceive and attribute control in social scenarios, and how this process might differ in depression. Our findings from these experiments extend our understanding by suggesting that reduced contingency learning in depression applies broadly to both cue-outcome and action-outcome relations contingency learning.

The well-established link between depression and reduced agency perception was reflected in Experiments 4-6, with depressed participants reporting reduced agency for both self and other agents across all conditions. The active inference model captured this through

lower probability values of outcome presence in the likelihood tensor, resulting in a flatter distribution. This implies that depressed participants may be less certain about their agency given the same observations, aligning with findings of reduced certainty or confidence in other cognitive behavioural tasks (Fu et al., 2005; Herskovic et al., 1986).

A key finding is the differential attribution of control to self and other agents between depressed and non-depressed individuals. Non-depressed participants perceived their control as independent of the external agent, and in more opposing directions of control. This self-serving bias, which also manifests as an "other-serving" bias, may help individuals feel a balance of control and perhaps a healthy sense of competition. It may serve to preserve the sense of independence and agency for all agents in the environment, potentially contributing to motivation and adaptive behaviour in various life situations.

Depressed participants, on the other hand, tended to show this opposing control bias less. While they often perceived others as having positive control, aligning with research linking depressive symptoms to an external locus of control (Aiken & Baucom, 1982; Benassi et al., 1988), they also perceived themselves as having positive control, matching that of external agents. This pattern may be attributed to several interrelated factors in depressed individuals: a blurred self-other distinction, which has been associated with increased mirroring behaviour; an increased sense of shared agency, perhaps reducing the sense of independence resulting in more social loafing behaviour; and potentially blunted cognitive processing of complex social scenarios involving theory of mind (Eddy, 2022; Karau & Williams, 1993). These factors could collectively contribute to more homogeneous attributions of control.

Additionally, we observed interesting and differing patterns of environmental engagement between depressed and non-depressed participants. Non-clinically depressed participants showed reduced pressing of the button while exhibiting more goal-directed behaviour. On the other hand, clinically depressed participants showed increased button-

pressing but constant button-pressing, suggesting heightened but potentially aimless environmental sampling. This could indicate non-adaptive environmental responses or a compensatory mechanism driven by heightened uncertainty. Such behaviour might reflect a desperate attempt to exert control over the environment, in the attenuated processing of goals or expectations about outcomes. In contrast, non-depressed participants tended to press the button less frequently, possibly due to quicker, more confident learning about agency.

The active inference modelling results revealed potential differences in observation preferences and environmental engagement. Non-clinically depressed participants showed a higher preference for observing the outcome, whereas clinically depressed participants showed a lower preference, aligning with the link between depression and blunted sensitivity to goals and rewards (Pitliya et al., 2022; Pizzagalli et al., 2009; Vrieze et al., 2013). The increased button-pressing in depressed participants was reflected in higher priors to select the action to press the button, and a lower gamma value, portraying higher reliance on this prior and therefore less adaptability.

An intriguing finding emerged when we introduced a gamified element to the task by displaying a running score of the number of outcomes produced. This modification led to a reversal in the action patterns of clinically and non-clinically depressed participants. The non-clinical group began operating at the level previously observed in the clinical group (i.e., higher frequency of button pressing), while the clinical group's behaviour shifted to resemble that of the non-clinical group in the original task (i.e., lower frequency of button pressing). However, both groups presented more goal-directed behaviour, adapting their actions to produce the outcome more compared to the original task. This suggests that outsourcing the tracking of outcomes or making it more explicit may be more beneficial for depressed individuals, as their actions became less monotonous and slightly more adaptive in accordance with the goal. Furthermore, we saw the non-clinically depressed group initially pressing the button less and

then increasing it and then either continuing to press the button or not depending on whether the participant had positive or negative control, respectively. This suggests more explorative behaviour first, followed by exploitative.

While Experiments 4 and 5 indicated that differing actions between non-clinically and clinically depressed participants did not result in different experienced objective contingencies as calculated by ΔP, Experiment 6 suggested otherwise. This discrepancy may be attributed to the longer duration of trials in Experiment 6, which allowed for the emergence of differences in experienced ΔPs despite the differing action patterns. Clinically depressed participants tended to produce lower action-outcome contingencies (ΔPs), which may reflect a reduced ability to create consistent patterns of behaviour that lead to predictable outcomes.

These findings underscore the importance of considering agency perception within a social context, revealing complexities in agency attribution and behavioural strategies not apparent in single-agent paradigms. This provides a more ecologically valid model of how individuals, particularly those with depression, navigate agency in real-world social environments.

Nonetheless, there are limitations of our study to consider. The simplified agency representation, with only one causal structure (two potential causes and one outcome), limits generalisability. Although we studied what happens in the presence of a separate cue-outcome relation in Experiments 1-3, we did not examine this in an agency context. Real-world environments often involve multiple causes and multiple outcomes, which were not explored in this study. Future research could explore more complex causal structures to better reflect the intricacies of real-world agency scenarios.

The use of a simulated agent as the “other” in our multi-agent context, while allowing for controlled experimental conditions, may limit the ecological validity of our findings. Real-world social interactions involve dynamic, reciprocal influences between agents, which our

paradigm did not capture. Future studies could involve two real participants performing the task with real buttons to increase ecological validity and potentially yield more nuanced data on social agency perception.

The absence of an emotional or semantic context further limits ecological validity. We used neutral, abstract stimuli that did not have an inherent connection with each other. While this approach is valuable for exploring the foundations of contingency learning without the influence of prior beliefs, future research should incorporate more meaningful stimuli to better reflect real-world agency scenarios where actions and outcomes often carry emotional or social significance.

Additionally, while we identified group-level trends, our study does not account for the heterogeneity within depressed and non-depressed groups. Depression is a complex and multifaceted condition, with various subtypes that may influence agency perception and learning differently. Future studies should explore individual differences within groups, acknowledging the diversity of experiences within depression. This could lead to a more nuanced understanding of how different manifestations of depression affect agency perception and potentially inform more targeted interventions.

However, with that said, our research approach can be extended to assess agential learning in multi-agent contexts across various conditions associated with altered senses of agency, potentially leading to more tailored interventions across a range of mental health conditions. For example, individuals with obsessive-compulsive disorder (OCD) demonstrate loss of context-dependent goal-directed behaviour, resulting in compulsively repeated actions to produce outcomes (Szalai, 2019). Those with depersonalisation disorder experience a diminished sense of agency due to a lack of sense of self, and while this concerns the implicit sense of agency more (Ciaunica et al., 2022; Simeon & Abugel, 2006), it would be interesting to see whether the development of an explicit sense of agency occurs differently to the general

population. Individuals experiencing schizophrenia and delusions may have a reduced sense of agency of the self and an exaggerated perception of agency of another agent (Daprati et al., 1997; Griffiths et al., 2019). And those with gambling disorders may have an enhanced perception of agency as they continue to produce actions even when it has no bearing on the outcome (Henslin, 1967). Exploring agency perception across these diverse conditions could provide valuable insights into the underlying mechanisms of agency and inform targeted therapeutic approaches.

Furthermore, sampling could also be improved in future studies. When we examined non-clinically depressed individuals in Experiments 1-4, we could only investigate low vs high depressive symptoms categorised via the cut-off score of 13 in the BDI-II. Future research could collect more data to ensure a more distributed sample across the 3-4 BDI-II categories (Beck et al., 1996): minimal depression range (scores of 0–13), mild-to-moderate range (scores of 14–28), and severe range (scores of 29–63). When we examined clinical vs non-clinical depression, we had relatively small and unequal sample sizes, which likely contributed to the lack of definitive statistical evidence for group differences. Larger and more balanced samples would increase the statistical power and reliability of the findings.

Longitudinal designs could help elucidate the causal relationships between depression and changes in agency perception over time. Such studies could track individuals as they move through different stages of depression, providing insights into how agency perception changes with the course of the condition and potentially identifying key periods for intervention.

In conclusion, our novel behavioural task and active inference modelling highlight the intricate relationship between perceptual and action processes that point to the interplay between agency inference and depression in multi-agent contexts. By elucidating the mechanisms underlying agency inference in depression within social contexts, we can better inform therapeutic strategies, such as focusing on enhancing identification of negative relations

in the environment. Cognitive-behavioural therapies could also focus on helping depressive individuals recognise and challenge their beliefs about agency in social contexts, focusing on differentiating and highlighting the independence of the self and other. Mindfulness-based interventions might increase awareness of one's actions and their consequences, potentially addressing the tendency towards aimless environmental sampling by explicit methods of tracking the outcome and fostering a stronger sense of independent agency.

This research contributes to our understanding of how individuals, particularly those with depression, perceive and learn about their agency in complex social environments. It underscores the importance of considering agency not just as an individual phenomenon but as a dynamic process shaped by social contexts. As we continue to unravel the intricacies of agency perception, we can move closer to developing more effective interventions that can help individuals navigate their world with a stronger, more adaptive sense of agency.

References

Abdolmanafi, A., Besharat, M. A., Farahani, H., & Khodaii, M. R. (2011). The moderating role of locus of control on the relationship between anger and depression in patients with major depression disorder. *Procedia-Social and Behavioral Sciences*, *30*, 297-301.

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, *87*(1), 49-74. https://doi.org/10.1037/0021-843X.87.1.49

Ackermann, R., & DeRubeis, R. J. (1991). Is depressive realism real? *Clinical Psychology Review*, *11*(5), 565-584. https://doi.org/https://doi.org/10.1016/0272-7358(91)90004-E

Aiken, P. A., & Baucom, D. H. (1982). Locus of control and depression: That confounded relationship. *Journal of Personality Assessment*, *46*(4), 391-395.

Albarracin, M., & Pitliya, R. J. (2022). The nature of beliefs and believing [Opinion]. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.981925

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147-149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*(3), 435.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *34*(1), 1.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*(4), 381-405.

Allan, L. G., Siegel, S., & Hannah, S. (2007). The sad truth about depressive realism. *Quarterly Journal of Experimental Psychology*, *60*(3), 482-495.

Alloy, L. B., Abramson, L., & Kossman, D. (1985). The judgment of predictability in depressed and nondepressed college students. *Affect, conditioning, and cognition: Essays on the determinants of behavior*, 229-246.

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of experimental psychology: General*, *108*(4), 441-485. https://doi.org/10.1037/0096-3445.108.4.441

Alloy, L. B., & Abramson, L. Y. (1982). Learned helplessness, depression, and the illusion of control. *Journal of personality and social psychology*, *42*(6), 1114.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological review*, *91*(1), 112.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing. https://doi.org/https://doi.org/10.1176/appi.books.9780890425596

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, *52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of experimental psychology: General*, *112*(1), 117.

Ath, G. D., Everson, R. M., Rahat, A. A. M., & Fieldsend, J. E. (2021). Greed Is Good: Exploration and Exploitation Trade-offs in Bayesian Optimisation. *ACM Trans. Evol. Learn. Optim.*, *1*(1), Article 1. https://doi.org/10.1145/3425501

Aust, F., Haaf, J. M., & Stahl, C. (2019). A memory-based judgment account of expectancy-liking dissociations in evaluative conditioning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(3), 417.

Baetu, I., & Baker, A. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 153.

Baetu, I., Baker, A. G., Darredeau, C., & Murphy, R. A. (2005). A comparative approach to cue competition with one and two strong predictors. *Learning & behavior*, *33*(2), 160-171. https://doi.org/10.3758/BF03196060

Baker, A. G., Berbrier, M. W., & Vallee-Tourangeau, F. (1989). Judgements of a 2 × 2 Contingency Table: Sequential processing and the Learning Curve. *The Quarterly Journal of Experimental Psychology Section B*, *41*(1b), 65-97. https://doi.org/10.1080/14640748908401184

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 414-432. https://doi.org/10.1037/0278-7393.19.2.414

Baker, A. G., Murphy, R., Mehta, R., & Baetu, I. (2005). Mental models of causation: A comparative view. *New directions in human associative learning*, 11-40.

Baker, A. G., Murphy, R., & Vallée-Tourangeau, F. (1996). Associative and normative models of causal induction: Reacting to versus understanding cause.

Baker, A. G., Murphy, R. A., Vallée-Tourangeau, F., & Mehta, R. (2000). Contingency learning and causal reasoning. In *Handbook of contemporary learning theories* (pp. 255-306). Psychology Press.

Baker, A. G., Vallée-Tourangeau, F., & Murphy, R. A. (2000). Asymptotic judgment of cause in a relative validity paradigm. *Memory & Cognition*, *28*(3), 466-479. https://doi.org/10.3758/BF03198561

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American psychologist*, *37*(2), 122-147. https://doi.org/10.1037/0003-066X.37.2.122

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., & Green, P. (2009). Package 'lme4'. *URL http://lme4. r-forge. r-project. org*.

Beck, A. T. (1967). *Depression: clinical, experimental, and theoretical aspects*. Staples Press.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Beck depression inventory.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 238.

Béghin, G., & Markovits, H. (2022). Reasoning strategies and prior knowledge effects in contingency learning. *Memory & Cognition*, *50*(6), 1269-1283. https://doi.org/10.3758/s13421-022-01319-w

Benassi, V. A., & Mahler, H. I. (1985). Contingency judgments by depressed college students: Sadder but not always wiser. *Journal of personality and social psychology*, *49*(5), 1323.

Benassi, V. A., Sweeney, P. D., & Dufour, C. L. (1988). Is there a relation between locus of control orientation and depression? *Journal of Abnormal Psychology*, *97*(3), 357.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., & Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, *20*(28), 1-6.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature neuroscience*, *1*(7), 635-640.

Blanco, F., Matute, H., & Vadillo, M. A. (2009). Depressive realism: Wiser or quieter? *The Psychological Record*, *59*(4), 551-562.
Blanco, F., Matute, H., & Vadillo, M. A. (2011). Making the uncontrollable seem controllable: The role of action in the illusion of control. *Quarterly Journal of Experimental Psychology*, *64*(7), 1290-1304.
Blanco, F., Matute, H., & Vadillo, M. A. (2012). Mediating role of activity level in the depressive realism effect.
Blanco, F., Matute, H., & Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & behavior*, *41*, 333-340.
Bollini, A. M., Walker, E. F., Hamann, S., & Kestler, L. (2004). The influence of perceived control and locus of control on the cortisol and subjective responses to stress. *Biological psychology*, *67*(3), 245-260.
Bowie, C. R., Milanovic, M., Tran, T., & Cassidy, S. (2017). Disengagement from tasks as a function of cognitive load and depressive symptom severity. *Cogn Neuropsychiatry*, *22*(1), 83-94. https://doi.org/10.1080/13546805.2016.1267617
Buffett, W. (1997). *Berkshire Hathaway 1996 Shareholder Letter.* https://www.berkshirehathaway.com/letters/1996.html
Bugelski, B. R. (1962). Presentation time, total time, and mediation in paired associate learning. *Journal of Experimental Psychology*, *63*(4), 409.
Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, *6*(1), 3-5.
Caputo, G., & Guerra, S. (1998). Attentional selection by distractor suppression. *Vision Research*, *38*(5), 669-689. https://doi.org/https://doi.org/10.1016/S0042-6989(97)00189-2
Carruthers, G. (2012). The case for the comparator model as an explanation of the sense of agency and its breakdowns. *Consciousness and Cognition*, *21*(1), 30-45.
Carter, T. (2007). An introduction to information theory and entropy. *Complex systems summer school, Santa Fe*.
Castiello, S., Miller, R. R., Witnauer, J. E., Alcaide, D. M., Fung, E., Pitliya, R. J., Morrissey, D. K. C., & Murphy, R. A. (2022). Benefiting from trial spacing without the cost of prolonged training: Frequency, not duration, of trials with absent stimuli enhances perceived contingency. *J Exp Psychol Gen*, *151*(8), 1772-1792. https://doi.org/10.1037/xge0001166
Castiello, S., Senan, S., Msetfi, R. M., & Murphy, R. A. (2020). Traits for depression related to agentic and external control. *Learning and Motivation*, *72*, 101684. https://doi.org/https://doi.org/10.1016/j.lmot.2020.101684
Chambon, V., Moore, J. W., & Haggard, P. (2015). TMS stimulation over the inferior parietal cortex disrupts prospective sense of agency. *Brain Structure and Function*, *220*, 3627-3639.
Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 837.
Chatlosh, D. L., Neunaber, D. J., & Wasserman, E. A. (1985). Response-outcome contingency: Behavioral and judgmental effects of appetitive and aversive outcomes with college students. *Learning and Motivation*, *16*(1), 1-34. https://doi.org/https://doi.org/10.1016/0023-9690(85)90002-5
Chekroun, P., & Brauer, M. (2002). The bystander effect and social control behavior: The effect of the presence of others on people's reactions to norm violations. *European Journal of Social Psychology*, *32*(6), 853-867.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review*, *104*(2), 367.

Ciaunica, A., Seth, A., Limanowski, J., Hesp, C., & Friston, K. J. (2022). I overthink—Therefore I am not: An active inference account of altered sense of self and agency in depersonalisation disorder. *Consciousness and Cognition*, *101*, 103320.

Crump, M. J. C., Hannah, S. D., Allan, L. G., & Hord, L. K. (2007). Contingency judgements on the fly. *Quarterly Journal of Experimental Psychology*, *60*(6), 753-761. https://doi.org/10.1080/17470210701257685

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition*, *65*(1), 71-86.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology*, *8*(4p1), 377.

Darredeau, C., Baetu, I., Baker, A. G., & Murphy, R. A. (2009). Competition between multiple causes of a single outcome in causal reasoning. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(1), 1-14. https://doi.org/10.1037/a0012699

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & behavior*, *37*(1), 1-20.

De Houwer, J., & Beckers, T. (2002a). Higher-order retrospective revaluation in human causal learning. *The Quarterly Journal of Experimental Psychology Section B*, *55*(2b), 137-151.

De Houwer, J., & Beckers, T. (2002b). A review of recent developments in research and theories on human contingency learning. *The Quarterly Journal of Experimental Psychology Section B*, *55*(4), 289-310. https://doi.org/10.1080/02724990244000034

De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(3), 965-985.

Declerck, C. H., Boone, C., & De Brabander, B. (2006). On feeling in control: A biological theory for individual differences in control perception. *Brain and Cognition*, *62*(2), 143-176. https://doi.org/https://doi.org/10.1016/j.bandc.2006.04.004

Delorme, A., Poncet, M., & Fabre-Thorpe, M. (2018). Briefly flashed scenes can be stored in long-term memory. *Frontiers in neuroscience*, *12*, 688.

DeNeve, K. M., & Cooper, H. (1998). The happy personality: a meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin*, *124*(2), 197.

Denniston, J. C., Savastano, H. I., Blaisdell, A. P., & Miller, R. R. (2003). Cue competition as a retrieval deficit. *Learning and Motivation*, *34*(1), 1-31. https://doi.org/https://doi.org/10.1016/S0023-9690(02)00505-2

Denniston, J. C., Savastano, H. I., & Miller, R. R. (2000). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. *Handbook of contemporary learning theories*, 65-117.

Di Lollo, V., Kawahara, J.-i., Shahab Ghorashi, S. M., & Enns, J. T. (2005). The attentional blink: Resource depletion or temporary loss of control? *Psychological research*, *69*(3), 191-200. https://doi.org/10.1007/s00426-004-0173-x

Dickinson, A. (1996). Within compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology: Section B*, *49*(1), 60-80.

Dickinson, A. (2001). The 28th Bartlett memorial lecture causal learning: An associative analysis. *The Quarterly Journal of Experimental Psychology Section B*, *54*(1b), 3-25.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *The Quarterly Journal of Experimental Psychology*, *36*(1), 29-50.

Dobson, K. S., & Pusch, D. (1995). A test of the depressive realism hypothesis in clinically depressed subjects. *Cognitive therapy and research*, *19*(2), 179-194. https://doi.org/10.1007/BF02229693

Eckstein, M. K., Dahl, R., Wilbrecht, L., & Collins, A. G. (2019). Modeling the development of learning strategies in a volatile environment.

Eddy, C. M. (2022). The Transdiagnostic Relevance of Self-Other Distinction to Psychiatry Spans Emotional, Cognitive and Motor Domains. *Frontiers in psychiatry*, *13*, 797952. https://doi.org/10.3389/fpsyt.2022.797952

Engel, J. F. (1963). Are Automobile Purchasers Dissonant Consumers? *Journal of Marketing*, *27*(2), 55-58. https://doi.org/10.2307/1248373

Ferster, C. B. (1973). A functional analysis of depression. *American psychologist*, *28*(10), 857.

Fresco, D. M., Alloy, L. B., & Reilly–Harrington, N. (2006). Association of attributional style for negative and positive events and the occurrence of life events with depression and anxiety. *Journal of social and clinical psychology*, *25*(10), 1140-1160.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological cybernetics*, *104*(1), 137-160.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, *7*, 598.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127-138.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural computation*, *29*(1), 1-49. https://doi.org/10.1162/NECO_a_00912

Fu, T., Koutstaal, W., Fu, C. H., Poon, L., & Cleare, A. J. (2005). Depression, confidence, and decision: Evidence against depressive realism. *Journal of Psychopathology and Behavioral Assessment*, *27*, 243-252.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences*, *4*(1), 14-21.

Gallistel, C. R., & Balsam, P. D. (2014). Time to rethink the neural mechanisms of learning and memory. *Neurobiology of learning and memory*, *108*, 136-144.

Gallistel, C. R., Craig, A. R., & Shahan, T. A. (2014). Temporal contingency. *Behavioural Processes*, *101*, 89-96.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Geng, J. J., & DiQuattro, N. E. (2010). Attentional capture by a perceptually salient non-target facilitates target processing through inhibition and rapid rejection. *Journal of Vision*, *10*(6), 5-5. https://doi.org/10.1167/10.6.5

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34-42. https://doi.org/https://doi.org/10.1016/j.cognition.2017.12.014

Gibson, E. J. (1988). Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual Review of Psychology*, *39*(1), 1-42.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in cognitive sciences*, *7*(1), 43-48.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085.

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: current status and future directions. *Annu Rev Clin Psychol*, *6*, 285-312. https://doi.org/10.1146/annurev.clinpsy.121208.131305

Griffiths, O., Shehabi, N., Murphy, R. A., & Le Pelley, M. E. (2019). Superstition predicts perception of illusory control. *British Journal of Psychology*, *110*(3), 499-518. https://doi.org/https://doi.org/10.1111/bjop.12344

Gureckis, T. M., & Markant, D. B. (2012). Self-Directed Learning:A Cognitive and Computational Perspective. *Perspectives on Psychological Science*, *7*(5), 464-481. https://doi.org/10.1177/1745691612454304

Haggard, P. (2005). Conscious intention and motor cognition. *Trends in cognitive sciences*, *9*(6), 290-295.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature neuroscience*, *5*(4), 382-385.

Hagmayer, Y., & Waldmann, M. R. (2007). Inferences about unobserved causes in human contingency learning. *Quarterly Journal of Experimental Psychology*, *60*(3), 330-355. https://doi.org/10.1080/17470210601002470

Hannah, S. D., & Beneteau, J. L. (2009). Just tell me what to do: bringing back experimenter control in active contingency tasks with the command-performance procedure and finding cue density effects along the way. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *63*(1), 59.

Hannah, S. D., Crump, M. J., Allan, L. G., & Siegel, S. (2009). Cue-interaction effects in contingency judgments using the streamed-trial procedure. *Can J Exp Psychol*, *63*(2), 103-112. https://doi.org/10.1037/a0013521

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, e4794.

Heins, C., Millidge, B., Demekas, D., Klein, B., Friston, K., Couzin, I., & Tschantz, A. (2022). pymdp: A Python library for active inference in discrete state spaces. *arXiv preprint arXiv:2201.03904*.

Henslin, J. M. (1967). Craps and magic. *American journal of Sociology*, *73*(3), 316-330.

Herskovic, J. E., Kietzman, M. L., & Sutton, S. (1986). Visual flicker in depression: response criteria, confidence ratings and response times. *Psychological medicine*, *16*(1), 187-197. https://doi.org/10.1017/S0033291700002622

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259-285.

Hovenkamp-Hermelink, J. H., Jeronimus, B. F., Spinhoven, P., Penninx, B. W., Schoevers, R. A., & Riese, H. (2019). Differential associations of locus of control with anxiety, depression and life-events: A five-wave, nine-year study to test stability and change. *Journal of Affective Disorders*, *253*, 26-34.

Hull, C. L. (1943). Principles of behavior: an introduction to behavior theory.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (Vol. 84). Routledge.

Jenkins, H., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological monographs: General and applied*, *79*(1), 1.

Johnson-Laird, P. N., & Khemlani, S. (2017). Mental models and causation. *Oxford handbook of causal reasoning*, 1-42.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*(4-6), 549-559.
Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.
Kamin, L. J. (1968). Attention-like processes in classical conditioning. Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation, Coral Gables, Florida.
Kang, P., Tobler, P. N., & Dayan, P. (2024). Bayesian reinforcement learning: A basic overview. *Neurobiology of Learning and Memory*, *211*, 107924. https://doi.org/https://doi.org/10.1016/j.nlm.2024.107924
Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1363.
Kapçi, E. G., & Cramer, D. (1999). Judgement of control revisited: Are the depressed realistic or pessimistic? *Counselling Psychology Quarterly*, *12*(1), 95-105.
Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology*, *65*(4), 681.
Karsh, N., & Eitam, B. (2015). I control therefore I do: Judgments of agency influence action selection. *Cognition*, *138*, 122-131. https://doi.org/https://doi.org/10.1016/j.cognition.2015.02.002
Kayser, A. S., Mitchell, J. M., Weinstein, D., & Frank, M. J. (2015). Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology*, *40*(2), 454-462.
Keller, A. S., Leikauf, J. E., Holt-Gosselin, B., Staveland, B. R., & Williams, L. M. (2019). Paying attention to attention in depression. *Translational psychiatry*, *9*(1), 279. https://doi.org/10.1038/s41398-019-0616-1
Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449-460.
Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *2*, 398.
Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & behavior*, *36*(3), 210-226.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest'. *R package version*, *2*(0), 734.
Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164-174. https://doi.org/https://doi.org/10.1016/j.cogsys.2010.07.007
Lennox, S. S., Bedell, J. R., Abramson, L. Y., Raps, C., & Foley, F. W. (1990). Judgement of Contingency: A Replication withe Hospitalized Depressed, Schizophrenic and Normal Samples. *Journal of Social Behavior and Personality*, *5*(4), 189.
Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. In https://CRAN.R-project.org/package=emmeans
Lewinsohn, P. M., Sullivan, J. M., & Grosscup, S. J. (1980). Changing reinforcing events: An approach to the treatment of depression. *Psychotherapy: Theory, Research & Practice*, *17*(3), 322-334. https://doi.org/10.1037/h0085929
MacDonald, K., & Frank, M. C. (2016). When does passive learning improve the effectiveness of active learning? Proceedings of the Annual Meeting of the Cognitive Science Society,
Macho, S., & Burkart, J. (2002). Recursive retrospective revaluation of causal judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1171.

Maier, S. F., & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of experimental psychology: General*, *105*(1), 3-46. https://doi.org/10.1037/0096-3445.105.1.3

Maier, S. F., & Seligman, M. E. P. (2016). *Learned helplessness at fifty: Insights from neuroscience* [doi:10.1037/rev0000033].

Marien, H., Aarts, H., & Custers, R. (2015). The interactive role of action-outcome learning and positive affective information in motivating human goal-directed behavior. *Motivation Science*, *1*(3), 165.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of experimental psychology: General*, *143*(1), 94-122. https://doi.org/10.1037/a0032108

Markovic, D., & Pitliya, R. J. (2024). *Agency Task Model Fitting with PyBefit*. Retrieved 2024 from https://github.com/riddhipits/pybefit/blob/main/examples/pymdp/simulating_and_fitting_agency_task_IWAI.ipynb

Martin, D. J., Abramson, L. Y., & Alloy, L. B. (1984). Illusion of control for self and others in depressed and nondepressed college students. *J Pers Soc Psychol*, *46*(1), 125-136. https://doi.org/10.1037//0022-3514.46.1.125

Mata, A., Garcia-Marques, L., Ferreira, M. B., & Mendonça, C. (2015). Goal-driven reasoning overcomes cell D neglect in contingency judgements. *Journal of Cognitive Psychology*, *27*(2), 238-249.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39.

Matute, H. (1996). Illusion of Control: Detecting Response-Outcome Independence in Analytic but Not in Naturalistic Conditions. *Psychological science*, *7*(5), 289-293. https://doi.org/10.1111/j.1467-9280.1996.tb00376.x

McGaugh, J. L. (1966). Time-dependent processes in memory storage. *Science (New York, N.Y.)*, *153*(3742), 1351-1358.

McLaren, I. P. L., McAndrew, A., Angerer, K., McLaren, R., Forrest, C., Bowditch, W., Monsell, S., & Verbruggen, F. (2018). Mackintosh lecture—: Association and cognition: Two processes, one system. *Quarterly Journal of Experimental Psychology*, *72*(2), 98-117. https://doi.org/10.1177/1747021818766287

Mehta, M. M., Na, S., Gu, X., Murrough, J. W., & Morris, L. S. (2023). Reward-related self-agency is disturbed in depression and anxiety. *PloS one*, *18*(3), e0282727.

Melcher, D. (2001). Persistence of visual memory for scenes. *Nature*, *412*(6845), 401-401. https://doi.org/10.1038/35086646

Mill, J. S., & Nagel, E. (2008). *Philosophy of Scientific Methods*. Genesis Publishing Pvt Ltd.

Miller, R. R., & Escobar, M. (2002). Associative interference between cues and between outcomes presented together and presented apart: an integration. *Behavioural Processes*, *57*(2), 163-185. https://doi.org/https://doi.org/10.1016/S0376-6357(02)00012-8

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of experimental psychology: General*, *125*(4), 370.

Moore, J. W., Middleton, D., Haggard, P., & Fletcher, P. C. (2012). Exploring implicit and explicit aspects of sense of agency. *Consciousness and Cognition*, *21*(4), 1748-1753.

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: A review. *Consciousness and Cognition*, *21*(1), 546-561. https://doi.org/https://doi.org/10.1016/j.concog.2011.12.002

Msetfi, R. M., Murphy, R. A., & Simpson, J. (2007). Depressive realism and the effect of intertrial interval on judgements of zero, positive, and negative contingencies. *Quarterly Journal of Experimental Psychology*, *60*(3), 461-481. https://doi.org/10.1080/17470210601002595

Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive Realism and Outcome Density Bias in Contingency Judgments: The Effect of the Context and Intertrial Interval. *Journal of experimental psychology: General*, *134*(1), 10-22. https://doi.org/10.1037/0096-3445.134.1.10

Murphy, R. A., Baker, A., & Fouquet, N. (2001). Relative validity effects with either one or two more valid cues in Pavlovian and instrumental conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*(1), 59.

Murphy, R. A., Witnauer, J. E., Castiello, S., Tsvetkov, A., Li, A., Alcaide, D. M., & Miller, R. R. (2021). More frequent, shorter trials enhance acquisition in a training session: There is a free lunch! *Journal of experimental psychology: General*.

Musca, S. C., Vadillo, M. A., Blanco, F., & Matute, H. (2010). The role of cue information in the outcome-density effect: Evidence from neural network simulations and a causal learning experiment. *Connection Science*, *22*(2), 177-192.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4), 979.

Nieuwenstein, M. R., Chun, M. M., van der Lubbe, R. H., & Hooge, I. T. (2005). Delayed attentional engagement in the attentional blink. *Journal of Experimental Psychology: Human perception and performance*, *31*(6), 1463.

Nieuwenstein, M. R., Potter, M. C., & Theeuwes, J. (2009). Unmasking the attentional blink. *Journal of Experimental Psychology: Human perception and performance*, *35*(1), 159.

Ophir, E. A., Sherman, E., & Lamy, D. (2020). An attentional blink in the absence of spatial attention: a cost of awareness? *Psychological research*, *84*, 1039-1055.

Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche*, *13*(1), 1-30.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of behavioral and experimental finance*, *17*, 22-27.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

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.

Pearl, J. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, *19*(2), 3.

Penn, D. C., & Povinelli, D. J. (2007). Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.*, *58*(1), 97-118.

Penton, T., Wang, X., Coll, M.-P., Catmur, C., & Bird, G. (2018). The influence of action–outcome contingency on motivation from control. *Experimental Brain Research*, *236*, 3239-3249.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic bulletin & review*, *14*, 577-596.

Peter, F., Been, S.-L., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition*, *31*(1), 133-142. https://doi.org/10.3758/BF03196088

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors.

Pineño, O., & Miller, R. R. (2007). Comparing associative, statistical, and inferential reasoning accounts of human contingency learning. *Quarterly Journal of Experimental Psychology*, *60*(3), 310-329. https://doi.org/10.1080/17470210601000680

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, *12*(1), 6587.

Pitliya, R. J., & Murphy, R. A. (2023). A Model of Agential Learning Using Active Inference. International Workshop on Active Inference,

Pitliya, R. J., Nelson, B. D., Hajcak, G., & Jin, J. (2022). Drift-Diffusion Model Reveals Impaired Reward-Based Perceptual Decision-Making Processes Associated with Depression in Late Childhood and Early Adolescent Girls. *Research on Child and Adolescent Psychopathology*, 1-14.

Pizzagalli, D. A., Holmes, A. J., Dillon, D. G., Goetz, E. L., Birk, J. L., Bogdan, R., Dougherty, D. D., Iosifescu, D. V., Rauch, S. L., & Fava, M. (2009). Reduced caudate and nucleus accumbens response to rewards in unmedicated individuals with major depressive disorder. *American Journal of Psychiatry*, *166*(6), 702-710.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of cs in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1-5. https://doi.org/10.1037/h0025984

Rescorla, R. A. (1969). Conditioned inhibition of fear resulting from negative CS-US contingencies. *Journal of comparative and physiological psychology*, *67*(4), 504.

Rescorla, R. A. (1972). Informational variables in Pavlovian conditioning. In *Psychology of learning and motivation* (Vol. 6, pp. 1-46). Elsevier.

Rescorla, R. A., & Wagner, A. H. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*, *2*, 64-69.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103-113. https://doi.org/https://doi.org/10.1111/j.2041-210X.2010.00012.x

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, *116*(28), 13903-13908.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of experimental psychology: General*, *110*(1), 101.

Seligman, M. E. (1974). *Depression and learned helplessness*. John Wiley & Sons.

Shaklee, H. (1983). Human covariation judgment: Accuracy and strategy. *Learning and Motivation*, *14*(4), 433-448.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory & Cognition*, *8*, 459-467.

Shanks, D. (1985a). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition*, *13*(2), 158-167. https://doi.org/10.3758/bf03197008

Shanks, D. (1985b). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1-21.

Shanks, D. (1986). Selective attribution and the judgment of causality. *Learning and Motivation*, *17*(4), 311-334.

Shanks, D. (1993). Human instrumental learning: A critical review of data and theory. *British Journal of Psychology*, *84*(3), 319-354.

Shanks, D. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology*, *60*(3), 291-309.

Shanks, D., & Dickinson, A. (1988). Associative accounts of causality judgment. In *Psychology of learning and motivation* (Vol. 21, pp. 229-261). Elsevier.

Shanks, D., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition*, *19*(4), 353-360. https://doi.org/10.3758/BF03197139

Shanks, D., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139-159.

Simeon, D., & Abugel, J. (2006). *Feeling unreal: Depersonalization disorder and the loss of the self*. Oxford University Press, USA.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of personality and social psychology*, *36*(11), 1202.

Sobel, D. M., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, *34*, 411-419.

Spellman, B. A. (1996a). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological science*, *7*(6), 337-342.

Spellman, B. A. (1996b). Conditionalizing causality. In *Psychology of learning and motivation* (Vol. 34, pp. 167-206). Elsevier.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219-239.

Szalai, J. (2019). The sense of agency in OCD. *Review of Philosophy and Psychology*, *10*(2), 363-380.

Tangen, J. M., & Allan, L. G. (2004). Cue interaction and judgments of causality: Contributions of causal and associative processes. *Memory & Cognition*, *32*(1), 107-124.

Team, R. C. (2021). *R: A language and environment for statistical computing.* In R Foundation for Statistical Computing. https://www.R-project.org/

Tenenbaum, J., & Griffiths, T. (2002). Theory-based causal inference. *Advances in neural information processing systems*, *15*.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309-318.

Thompson, A., Papas, A., Bartholomeusz, C., Nelson, B., & Yung, A. (2013). Externalized attributional bias in the Ultra High Risk (UHR) for psychosis population. *Psychiatry research*, *206*(2-3), 200-205.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science (New York, N.Y.)*, *211*(4481), 453-458. http://www.jstor.org/stable/1685855

Urcelay, G. P., & Miller, R. R. (2014). The functions of contexts in associative learning. *Behavioural Processes*, *104*, 2-12. https://doi.org/10.1016/j.beproc.2014.02.008

Urcelay, G. P., Wheeler, D. S., & Miller, R. R. (2009). Spacing extinction trials alleviates renewal and spontaneous recovery. *Learning & behavior*, *37*(1), 60-73.

Vallée-Tourangeau, F., Murphy, R. A., & Baker, A. G. (1998). Causal Induction in the Presence of a Perfect Negative Cue: Contrasting Predictions from Associative and Statistical Models. *The Quarterly Journal of Experimental Psychology Section B*, *51*(2b), 173-192. https://doi.org/10.1080/713932673

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*(2), 127-151.

Vazquez, C. (1987). Judgment of contingency: Cognitive biases in depressed and nondepressed subjects. *Journal of personality and social psychology*, *52*(2), 419.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Verschure, P. F., Pennartz, C. M., & Pezzulo, G. (2014). The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Philos Trans R Soc Lond B Biol Sci*, *369*(1655). https://doi.org/10.1098/rstb.2013.0483

von Glischinski, M., von Brachel, R., & Hirschfeld, G. (2019). How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). *Quality of Life Research*, *28*, 1111-1118.

Vrieze, E., Pizzagalli, D. A., Demyttenaere, K., Hompes, T., Sienaert, P., de Boer, P., Schmidt, M., & Claes, S. (2013). Reduced reward learning predicts outcome in major depressive disorder. *Biological psychiatry*, *73*(7), 639-645. https://doi.org/10.1016/j.biopsych.2012.10.014

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *J Exp Psychol*, *76*(2), 171-180. https://doi.org/10.1037/h0025414

Waldmann, M. R. (1996). Knowledge-based causal induction. In *Psychology of Learning and Motivation* (Vol. 34, pp. 47-88). Elsevier.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of experimental psychology: General*, *121*(2), 222-236. https://doi.org/10.1037/0096-3445.121.2.222

Waldmann, M. R., & Martignon, L. (1998). A Bayesian Network Model of Causal Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *20*.

Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. *Brazilian Journal of Psychiatry*, *35*, 416-431.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content* (Vol. 86). Harvard University Press.

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. In *Psychology of learning and motivation* (Vol. 26, pp. 27-82). Elsevier.

Wasserman, E. A., & Castro, L. (2005). Surprise and change: Variations in the strength of present and absent cues in causal learning. *Learning & behavior*, *33*(2), 131-146.

Wasserman, E. A., Chatlosh, D., & Neunaber, D. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation*, *14*(4), 406-432.

Wasserman, E. A., Dorner, W., & Kao, S. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 509.

Wasserman, E. A., Elek, S., Chatlosh, D., & Baker, A. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 174.

Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association.

Wasserman, E. A., Neunaber, D., Chatlosh, D., & OHara, M. (1983). Judgement of cause-effect relations by college students. Psychonomic Society Meeting, San Diego.

Wen, W. (2019). Does delay in feedback diminish sense of agency? A review. *Consciousness and Cognition*, *73*, 102759. https://doi.org/https://doi.org/10.1016/j.concog.2019.05.007

West, B. T., Welch, K. B., & Galecki, A. T. (2022). *Linear Mixed Models: A Practical Guide Using Statistical Software*. CRC Press. https://books.google.co.uk/books?id=k3GwEAAAQBAJ

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. *Create elegant data visualisations using the grammar of graphics. Version*, *2*(1), 1-189.

Wickham, H., François, R., Henry, L., Müller, K., & Wickham, M. H. (2019). Package 'dplyr'. *A Grammar of Data Manipulation. R package version*, *8*.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *J Exp Psychol Gen*, *143*(6), 2074-2081. https://doi.org/10.1037/a0038199

Witcher, E. S., & Ayres, J. J. (1980). Systematic manipulation of CS-US pairings in negative CS-US correlation procedures in rats. *Animal Learning & Behavior*, *8*, 67-74.

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, *58*(6), 967-973.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 593-602.

Yoshie, M., & Haggard, P. (2013). Negative emotional outcomes attenuate sense of agency over voluntary actions. *Current Biology*, *23*(20), 2028-2032.

Supplementary Materials

Supplementary Material 1.

Instructions on the computer task for Experiments 1-3.

Participants were first presented with the consent form and a participant information sheet, detailing the task, data usage and storage, contact information for queries, and their right to withdraw. Then, participants were asked to provide demographic information, including their age and gender, through a drop-down menu. The age options ranged from "younger than 18," individual numbers from 18 to 50, and "older than 50." The gender options included "male," "female," "prefer not to say," and an option to specify gender if it was not listed. Participants then provided consent, confirmed that they did not have a history of seizures and were completing the study on a laptop or desktop computer by selecting the appropriate checkboxes. After submitting this demographic information, participants were presented with the following text, divided across multiple slides:

If you have a history of seizures, are not between the ages of 18-50 years old, or do not speak fluent English, you are not permitted to participate in this study. *The following task includes rapidly moving stimuli that create some risk of inducing a seizure in people who are prone to have seizures.*

The following task is not mobile phone compatible. You may not participate using anything other than a desktop or laptop computer. Unfortunately, this task cannot run on Internet Explorer or Microsoft Edge. ***Please withdraw from the study if you do not have a suitable device.***

The task will take ***around 20 minutes (excluding any breaks)*** *and requires your absolute full attention. There will be a message on the screen telling you when you can take a break. You may not take more than two breaks and no break can be more than 10 minutes.*

We ask that you please ***maximize your browser window and minimize distractions*** *(e.g., put your phone on silent mode and put your computer on 'Do Not Disturb' mode so no notifications appear, such as email notifications). Do not write anything during the experiment or use your keyboard for anything other than responding to questions.*

In this experiment, you will be presented with a rapid stream of shapes, drawings and symbols. At the end of each stream, you will be asked some questions pertaining to the images you saw in the most recent stream.

You will first be asked about the relationship between the images of shapes and drawings. You will indicate the relatedness on a scale of -10 to 10. *Specifically, -10 indicates a strong negative relationship where the presence of one image predicts the absence of the other image, 0 indicates no relationship between the two images, and 10 indicates a strong positive relationship where the presence of one image predicts the presence of the other image.*

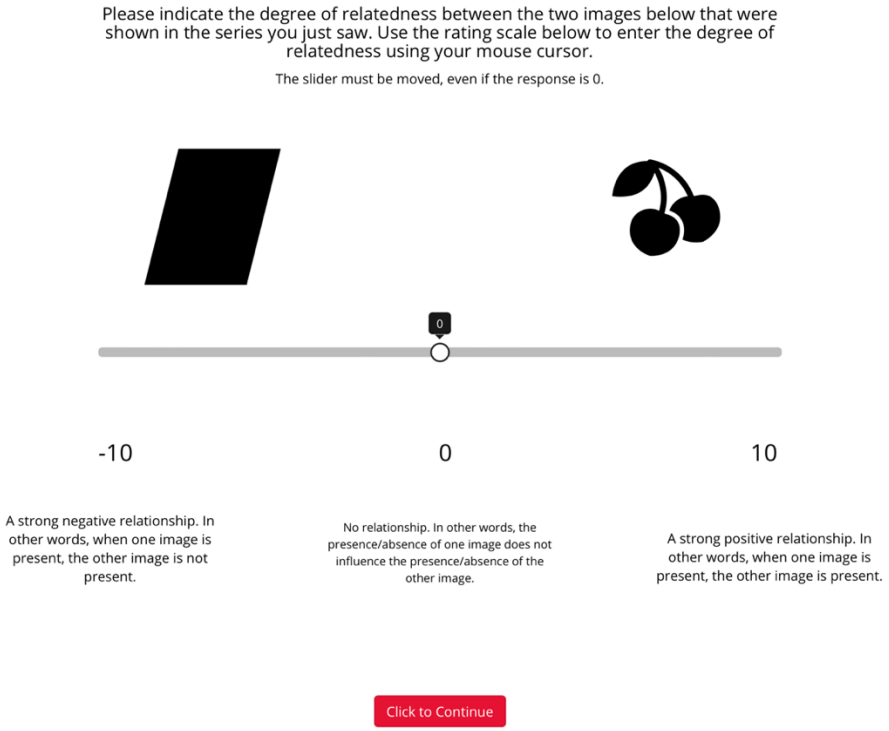
You will then be asked if you remember seeing particular symbols and you will respond by clicking on the "Yes" or "No" buttons. Therefore, you must pay attention to all images presented to you in this experiment.

Please keep your eyes focused on the fixation cross in the center of the screen.

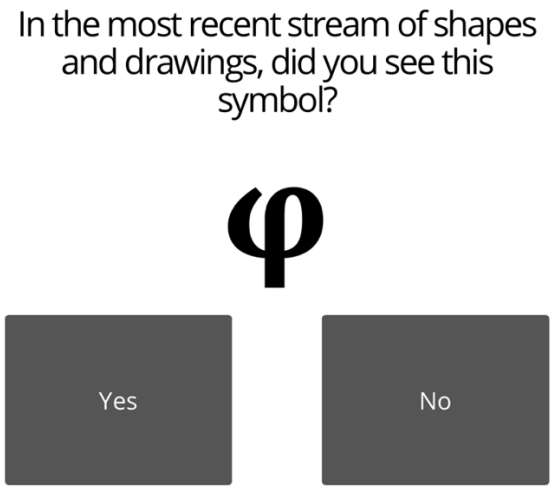
Take a break if need be when the screen indicates you can take a break. Ensure your break is no longer than 10 minutes.

There will be a 2-minute questionnaire after the computer task and then the study will end.

The participant clicked on a "Continue" button to begin the experiment. After each stream, participants were asked to "*Press the Continue button to proceed to the rating screen".* Participants responded on a slider; see the screenshot below:



The next two questions were whether the participant saw a non-target stimulus in the most recent stream of images:



In between each condition, the following text would appear:

You may take a break now, for 10 minutes maximum. Please remember that you may not take more than two breaks in total through the whole experiment.

Please click on Start Stream button to continue to the next series of shapes and objects.

Your mouse cursor will disappear during the next series of shapes and objects after pressing the Start Stream button. This is to remove any visual interference from your cursor. Your mouse cursor will reappear when questions pertaining to the series of shapes and objects appear.

After completing all conditions, participants were required to click on a square amongst other shapes in order to ensure that the participant was not a bot. Then, participants were asked "*Did you give your full attention to this experimental task (avoiding checking your mobile phone or other distracting activities)? Answer 'No' if you were distracted. Please answer honestly: you will receive credit regardless of your answer.".* Participants responded to the 'distraction question' with 'yes' or 'no' buttons presented on the screen. Then, the participants were thanked for their participation and progressed to receiving compensation for their participation.

Supplementary Material 2.

For each experiment, the number of **A**, **B**, **C**, and **D** trials in each frequency by contingency condition (stream of trials) and their respective ΔP value, which equals to (**A** / (**A**+**B**)) – (**C** / (**C**+**D**)). Content was not included in these tables as it does not impact the ΔP value.

A. Experiment 1

		D trial Frequency		
		0	**16 (baseline)**	**72**
Baseline Contingency	**Negative**	4 **A**, 44 **B**, 16 **C**, 0 **D** (ΔP = -0.92)	4 **A**, 44 **B**, 16 **C**, 16 **D** (ΔP = -0.42)	4 **A**, 44 **B**, 16 **C**, 72 **D** (ΔP = -0.10)
	Positive	44 **A**, 4 **B**, 16 **C**, 0 **D** (ΔP = -0.08)	44 **A**, 4 **B**, 16 **C**, 16 **D** (ΔP = 0.42)	44 **A**, 4 **B**, 16 **C**, 72 **D** (ΔP = 0.73)

B. Experiment 2

		D trial Frequency		
		0	**16 (baseline)**	**216**
Baseline Contingency	**Zero**	16 **A**, 16 **B**, 16 **C**, 0 **D** (ΔP = -0.50)	16 **A**, 16 **B**, 16 **C**, 16 **D** (ΔP = 0.00)	16 **A**, 16 **B**, 16 **C**, 216 **D** (ΔP = 0.32)

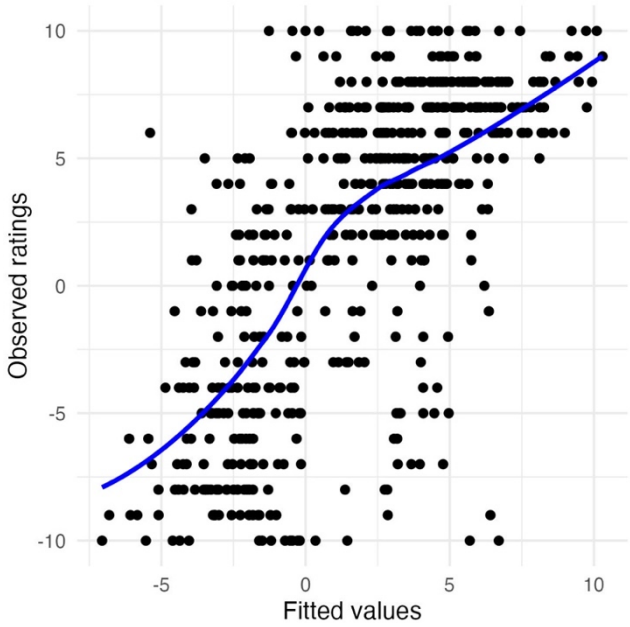
C.

		D trial Frequency		
		0	**16 (baseline)**	**216**
Baseline Contingency	**Negative**	4 **A**, 44 **B**, 16 **C**, 0 **D** ($\Delta P = -0.92$)	4 **A**, 44 **B**, 16 **C**, 16 **D** ($\Delta P = -0.42$)	4 **A**, 44 **B**, 16 **C**, 216 **D** ($\Delta P = 0.01$)
	Zero	24 **A**, 24 **B**, 16 **C**, 0 **D** ($\Delta P = -0.50$)	24 **A**, 24 **B**, 16 **C**, 16 **D** ($\Delta P = 0.00$)	24 **A**, 24 **B**, 16 **C**, 216 **D** ($\Delta P = 0.43$)
	Positive	44 **A**, 4 **B**, 16 **C**, 0 **D** ($\Delta P = -0.08$)	44 **A**, 4 **B**, 16 **C**, 16 **D** ($\Delta P = 0.42$)	44 **A**, 4 **B**, 16 **C**, 216 **D** ($\Delta P = 0.85$)

Supplementary Material 3.

A. Visualisations and results of analyses conducted to check model assumptions for Experiment 1.

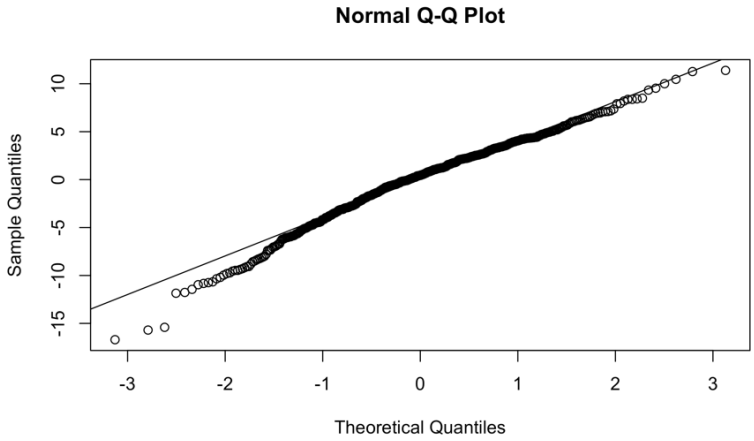
Assumption of Linearity Check



	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
study1_lme_resulting	22	3447.9	3543.5	-1701.9	3403.9			
study1_lme_resulting_freqpoly2	23	3449.0	3549.0	-1701.5	3403.0	0.8609	1	0.3535
study1_lme_resulting_bdipoly2	29	3455.4	3581.4	-1698.7	3397.4	5.6246	6	0.4665
study1_lme_resulting_bdipoly3	35	3461.5	3613.6	-1695.8	3391.5	5.8655	6	0.4384

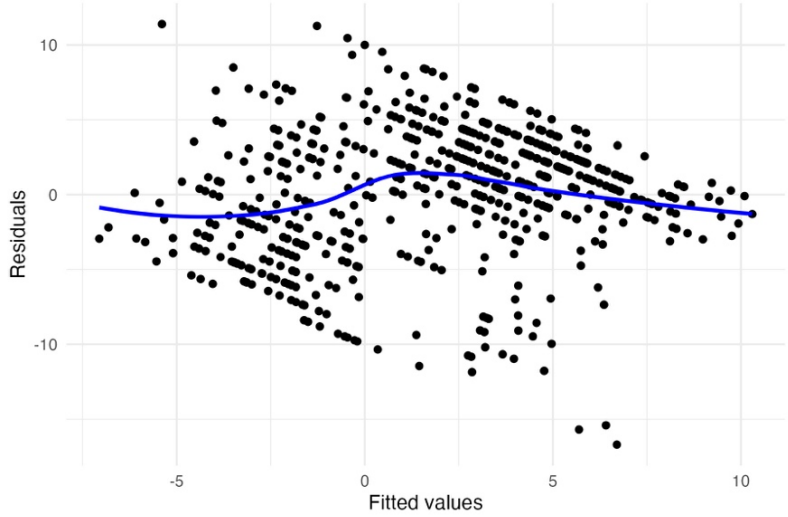
Notes. The linearity of the relationship between the independent and dependent variables was evaluated through a plot comparing the fitted values against observed ratings. Although the plot showed a very slight curvature, comparisons of the original model with polynomial models were not significant, supporting the linearity assumption.

Assumption of Normality of Residuals Check



Notes. The Normal Q-Q plot displays the distribution of residuals from the model, comparing the sample quantiles to the theoretical quantiles of a normal distribution. Most points align closely with the 45-degree reference line, suggesting that the residuals follow an approximately normal distribution in the centre. However, there are noticeable deviations at the tails, particularly in the lower quantiles where the points diverge below the line, indicating some skewness. These deviations suggest that while the residuals largely meet the assumption of normality, there may be minor departures at the extremes.

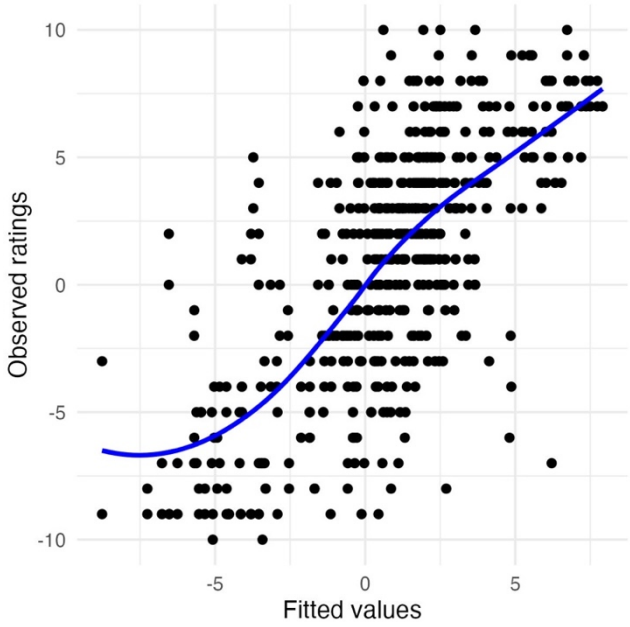
Assumption of Homoscedasticity Check



Notes. The residuals versus fitted values plot displays the residuals from the model against the predicted (fitted) values. Ideally, residuals should be randomly scattered around zero without any discernible pattern, indicating that the model fits the data well and the assumptions of linearity and homoscedasticity are met. In this plot, however, the residuals show a downward trend as the fitted values increase, suggesting heteroscedasticity – non-constant variance of the residuals across levels of the fitted values. This violation of the homoscedasticity assumption warrants using robust standard errors to obtain more accurate statistical inferences.

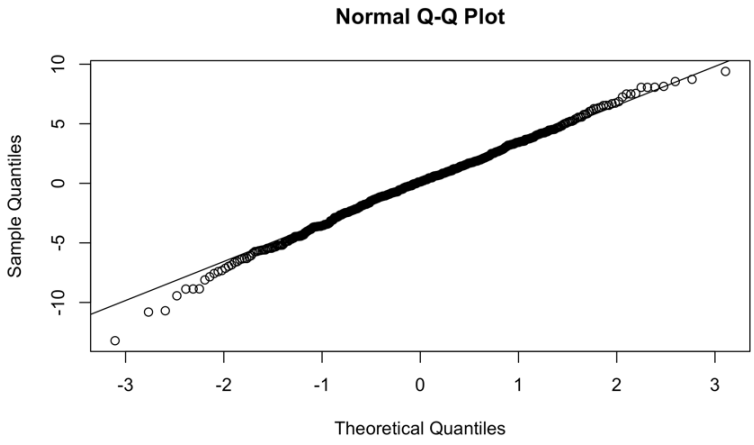
B. Visualisations and results of analyses conducted to check model assumptions for Experiment 2.

Assumption of Linearity Check

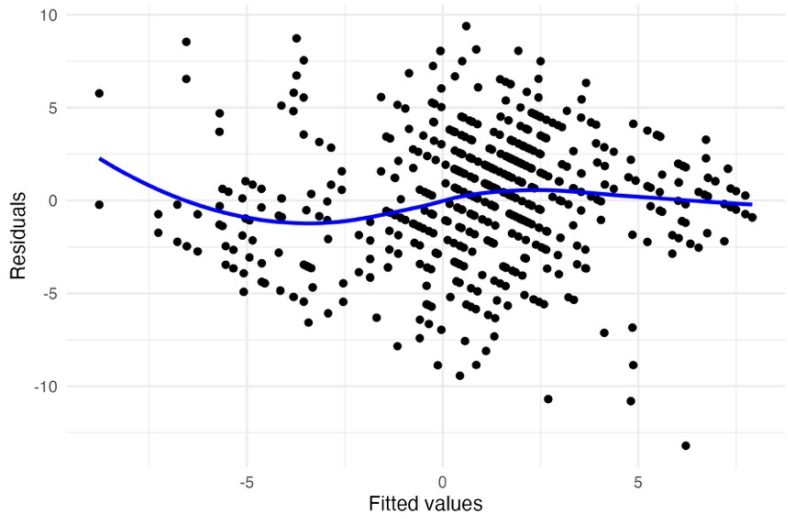


	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)	
study2_lme_resulting	14	2974.9	3034.7	-1473.5	2946.9				
study2_lme_resulting_freqpoly2	14	2974.9	3034.7	-1473.5	2946.9	0.0000	0		
study2_lme_resulting_bdipoly2	19	2977.0	3058.2	-1469.5	2939.0	7.9277	5	0.1603	
study2_lme_resulting_bdipoly3	24	2974.9	3077.5	-1463.5	2926.9	12.0807	5	0.0337	*

Assumption of Normality of Residuals Check

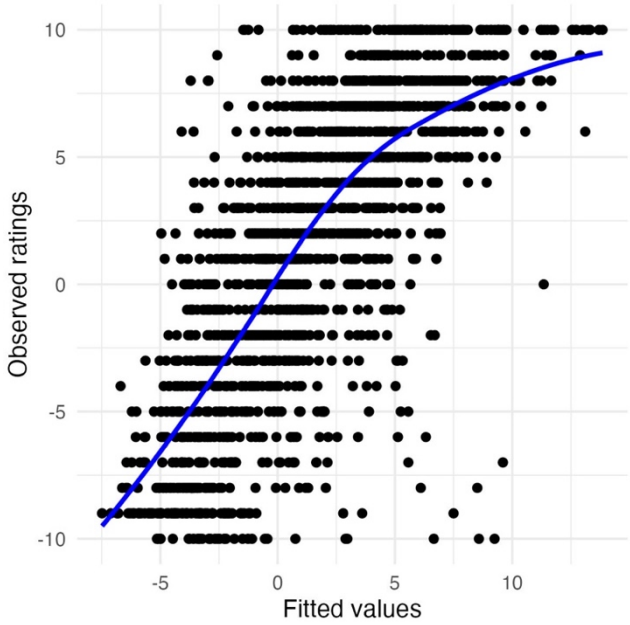


Assumption of Homoscedasticity Check



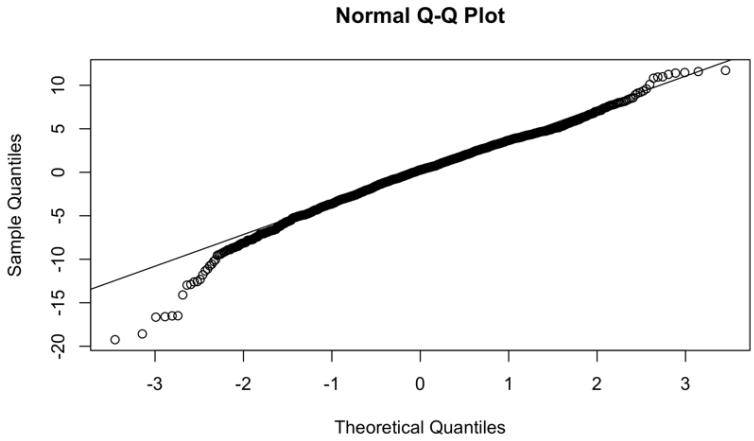
C. Visualisations and results of analyses conducted to check model assumptions for Experiment 3.

Assumption of Linearity Check

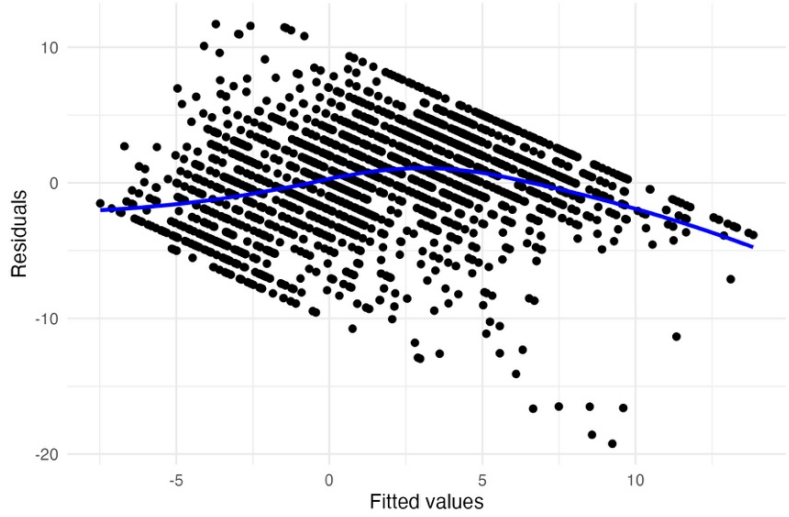


	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)	
study3_lme_resulting	24	10285	10417	-5118.5	10237				
study3_lme_resulting_freqpoly2	30	10294	10458	-5116.8	10234	3.5658	6	0.7352	
study3_lme_resulting_contpoly2	32	10276	10452	-5105.8	10212	21.9144	2	1.743e-05	***
study3_lme_resulting_bdipoly2	35	10278	10470	-5103.9	10208	3.8181	3	0.2818	
study3_lme_resulting_bdipoly3	46	10216	10469	-5062.2	10124	83.4736	11	3.131e-13	***

Assumption of Normality of Residuals Check



Assumption of Homoscedasticity Check

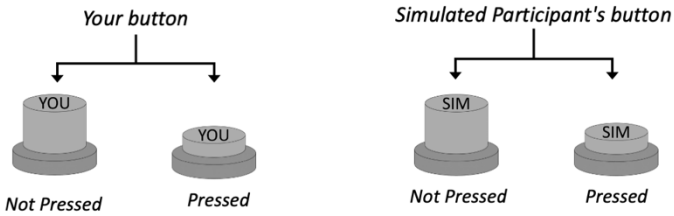


Supplementary Material 4.

Instructions for the computer task used in Experiments 4-6.

There will be two buttons presented on your screen. You have control over one button and a simulated participant has control over the other button.

This is what the buttons look like when pressed and not pressed:



You can control your button by pressing the letter **'B'** on your keyboard.

The buttons may or may not result in the filling of a shape. For example:



Unfilled shape *Filled in shape*

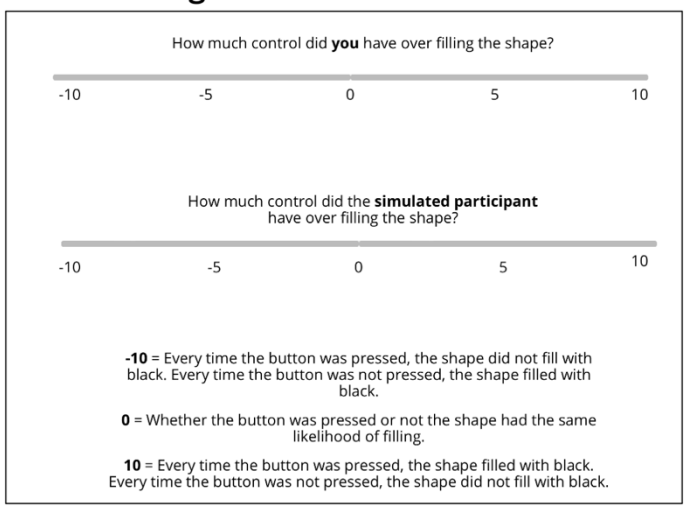
We would like you to learn how much control you and the simulated participant have over filling the shape.

In each session, you will have 1.5 minutes to learn how much control you and the simulated participant have. You can try pressing the button and not pressing the button, and see what happens.

There will be 10 sessions in total (15 minutes).

Each session will have a different shape for you to learn to fill. The amount of control you and the simulated participant have over filling in the shape will differ for each shape/session.

After each session, you will be asked to rate the amount of control you and the simulated participant have over filling the shape on your screen. The rating scale will look like this:



You have two main goals:

(1) Accurately learn the amount of control you and a simulated participant have over filling the shape on your screen. Note that you are not competing against the simulated participant.

(2) Please fill the shape in as much as possible.

We will begin with a practice round.

Supplementary Material 5.

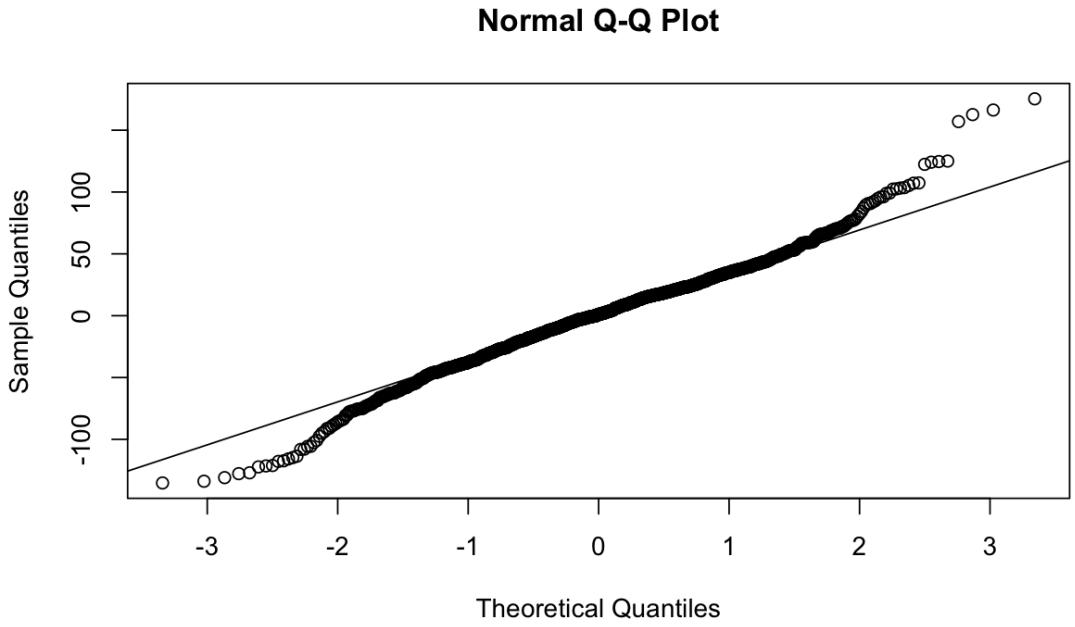
Code for the active inference model fitting process for Experiments 4-6.

The script for Experiment 5 can be viewed and downloaded at:
https://github.com/riddhipits/pybefit/blob/main/examples/pymdp/simulating_and_fitting_agency_task.ipynb

Supplementary Material 6.

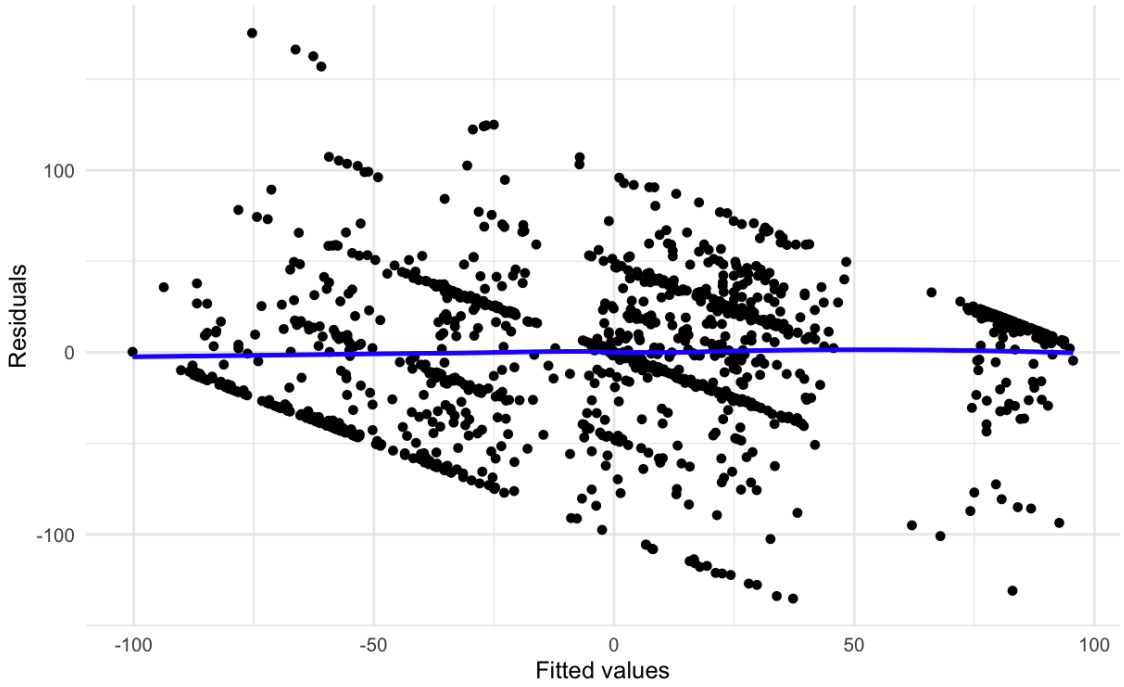
A. Visualisations and results of analyses conducted to check model assumptions for Experiment 1.

Assumption of Normality of Residuals Check



Notes. The Normal Q-Q plot displays the distribution of residuals from the model, comparing the sample quantiles to the theoretical quantiles of a normal distribution. Most points align closely with the 45-degree reference line, suggesting that the residuals follow an approximately normal distribution in the centre. However, there are noticeable deviations at the tails, particularly in the upper quantiles where the points diverge above the line, indicating some skewness. These deviations suggest that while the residuals largely meet the assumption of normality, there may be minor departures at the extremes.

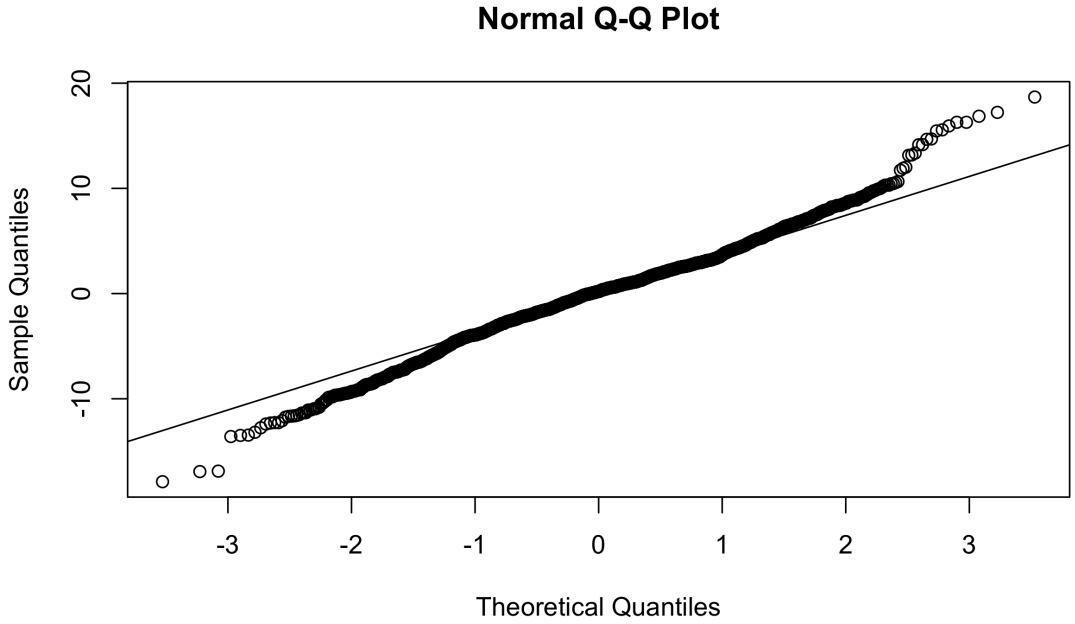
Assumption of Homoscedasticity Check



Notes. The residuals versus fitted values plot displays the residuals from the model against the predicted (fitted) values. Ideally, residuals should be randomly scattered around zero without any discernible pattern, indicating that the model fits the data well and the assumptions of linearity and homoscedasticity are met. In this plot, however, the residuals show a downward trend as the fitted values increase, suggesting heteroscedasticity – non-constant variance of the residuals across levels of the fitted values. This violation of the homoscedasticity assumption warrants using robust standard errors to obtain more accurate statistical inferences.

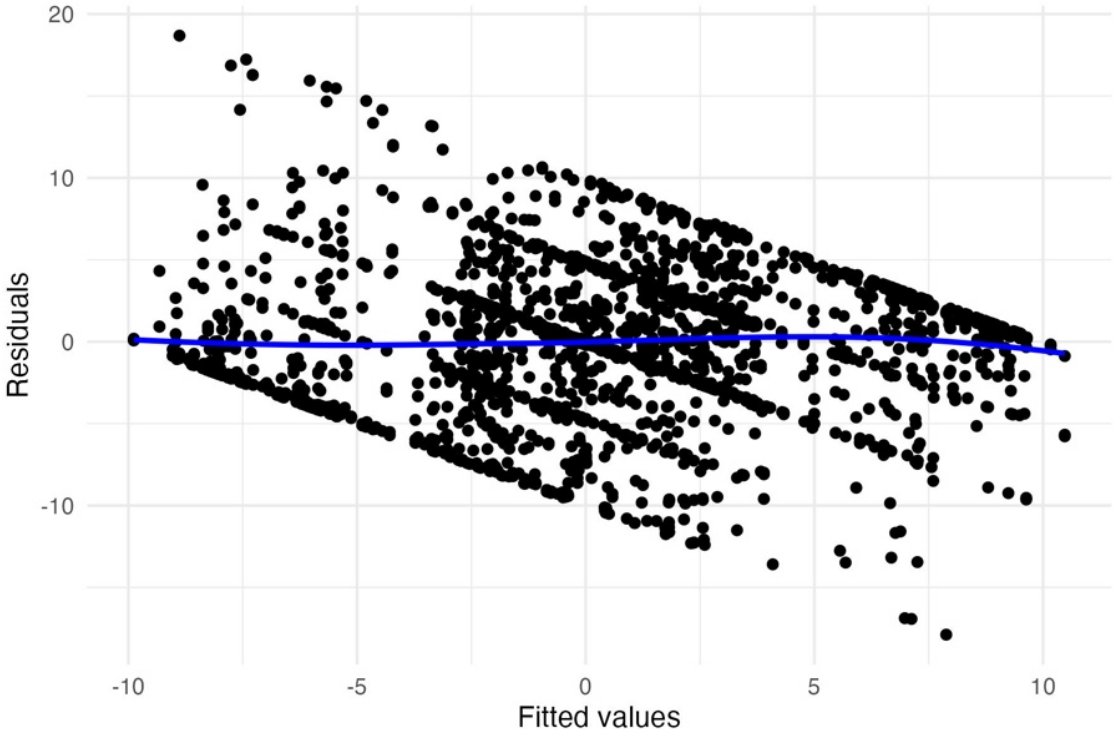
B. Visualisations and results of analyses conducted to check model assumptions for Experiment 2.

Assumption of Normality of Residuals Check



Notes. The Normal Q-Q plot displays the distribution of residuals from the model, comparing the sample quantiles to the theoretical quantiles of a normal distribution. Most points align closely with the 45-degree reference line, suggesting that the residuals follow an approximately normal distribution in the centre. However, there are noticeable deviations at the tails, particularly in the upper quantiles where the points diverge above the line, indicating some skewness. These deviations suggest that while the residuals largely meet the assumption of normality, there may be minor departures at the extremes.

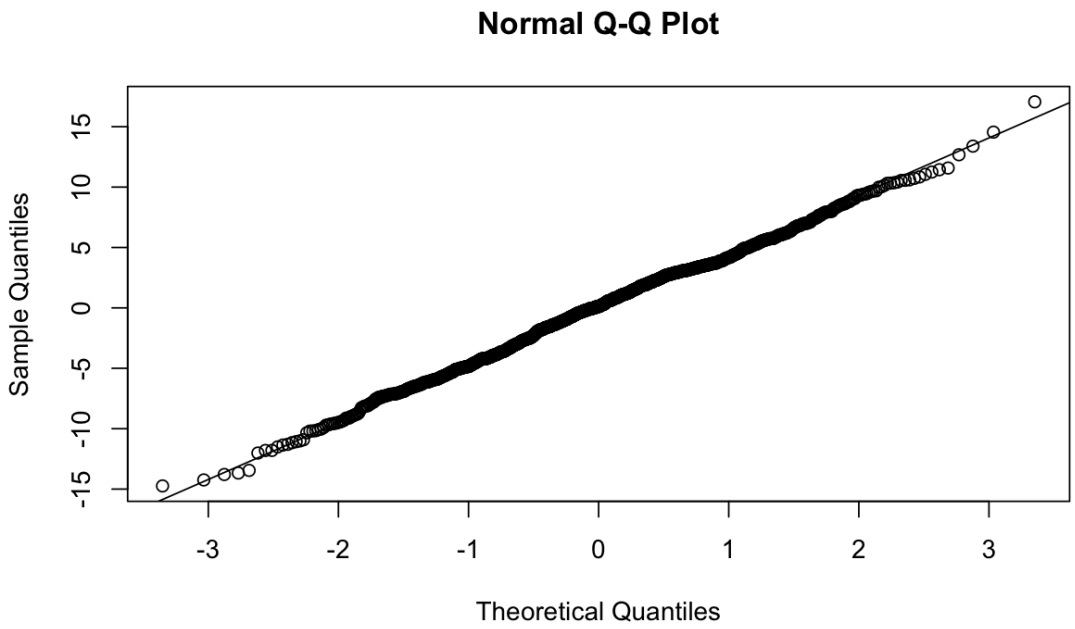
Assumption of Homoscedasticity Check



Notes. The residuals versus fitted values plot displays the residuals from the model against the predicted (fitted) values. Ideally, residuals should be randomly scattered around zero without any discernible pattern, indicating that the model fits the data well and the assumptions of linearity and homoscedasticity are met. In this plot, however, the residuals show a downward trend as the fitted values increase, suggesting heteroscedasticity – non-constant variance of the residuals across levels of the fitted values. This violation of the homoscedasticity assumption warrants using robust standard errors to obtain more accurate statistical inferences.

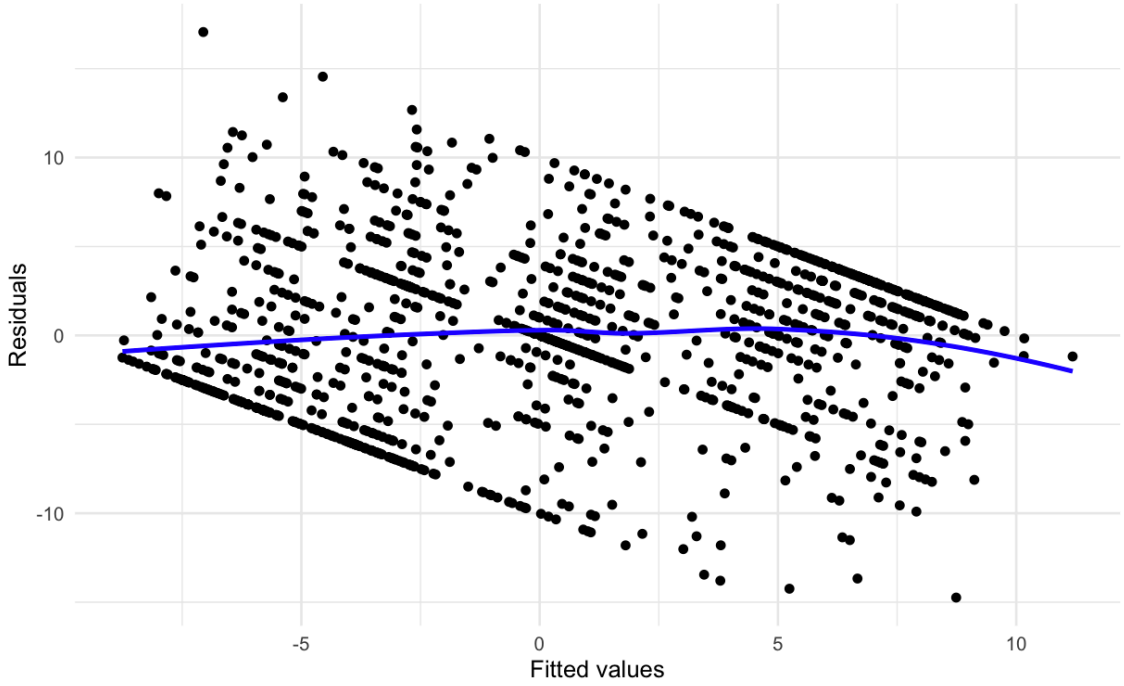
C. Visualisations and results of analyses conducted to check model assumptions for Experiment 3.

Assumption of Normality of Residuals Check



Notes. The Normal Q-Q plot displays the distribution of residuals from the model, comparing the sample quantiles to the theoretical quantiles of a normal distribution. All points align closely with the 45-degree reference line, suggesting that the residuals follow a normal distribution, meeting the assumption of normality.

Assumption of Homoscedasticity Check



Notes. The residuals versus fitted values plot displays the residuals from the model against the predicted (fitted) values. Ideally, residuals should be randomly scattered around zero without any discernible pattern, indicating that the model fits the data well and the assumptions of linearity and homoscedasticity are met. In this plot, however, the residuals show a downward trend as the fitted values increase, suggesting heteroscedasticity – non-constant variance of the residuals across levels of the fitted values. This violation of the homoscedasticity assumption warrants using robust standard errors to obtain more accurate statistical inferences.