Tracking Conflict in Reasoning

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

by

Revati Vijay Shivnekar
(Roll No: 19128262)

DEPARTMENT OF COGNITIVE SCIENCE
INDIAN INSTITUTE OF TECHNOLOGY KANPUR

December 2024

Certificate

It is certified that the work contained in this thesis entitled **"Tracking Conflict in Reasoning"** by **Revati Vijay Shivnekar** has been carried out under my supervision and that it has not been submitted elsewhere for a degree.

Digitally signed by Nisheeth Srivastava
DN: cn=Nisheeth Srivastava, o=IIT Kanpur, ou=CGS, email=nsrivast@iitk.ac.in, c=IN
Date: 2024.12.24 15:22:17 +05'30'

Dr. Nisheeth Srivastava
Associate Professor
CSE Department
Indian Institute of Technology Kanpur
December 2024

Declaration

This is to certify that the thesis titled **"Tracking Conflict in Reasoning"** has been authored by me. It presents the research conducted by me under the supervision of **Dr. Nisheeth Srivastava**.

To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted elsewhere, in part or in full, for a degree. Further, due credit has been attributed to the relevant state-of-the-art and collaborations with appropriate citations and acknowledgments, in line with established norms and practices.

Revati Vijay Shivnekar
Roll No. 19128262
CGS Department
Indian Institute of Technology Kanpur

Abstract

Name of the student: **Revati Vijay Shivnekar** Roll No: **19128262**

Degree for which submitted: **PhD** Department: **CGS Department**

Thesis title: **Tracking Conflict in Reasoning**

Thesis supervisors: **Dr. Nisheeth Srivastava**

Month and year of thesis submission: **December 2024**

Reasoning is replete with conflict. When in a tough spot, we often go back-and-forth between alternatives as we consider contrasting arguments, entertain counterfactuals, and reassess the same information. Our preferences shift in tandem with our thoughts over time, but importantly, we are often keenly aware of this rich spectrum of experience when we reason. Despite this, existing measures of reasoning are often limited to post hoc assessments or risk distorting the reasoning process itself. In this thesis, we aim to track conflict in moral and logical reasoning tasks as it evolves alongside our thoughts, while using tools that minimize task interference. In Study 1, we show that employing mouse-tracking to capture conflict, although less intrusive, is suboptimal, partly because it is typically employed too late in the process. For conflict measures to be effective, they must be attuned to the fluid and evolving nature of reasoning. Preference reversals, for instance, are often conspicuous when we reason. We explored whether these vacillations or shifts in preference could serve as an indicator of conflict. To this end, we developed the Switch paradigm, which allowed participants to report their changing inclinations in real-time by pressing keys. Across three experiments in Study 2, we demonstrate that this measure correlates with both internal and external validity checks of conflict in moral and logical reasoning tasks, while also offering a more direct means of testing predictions from established decision-making models. In Study 3, we explore whether conflict is experienced more continually when deliberating about moral issues, as opposed to when solving a logical problem with learned strategies. We combined the Switch paradigm with eye-tracking metrics, using pupil dilation and fixation duration as indicators of cognitive conflict. Our preliminary results suggest that conflict in logical reasoning may be more localized, emerging especially when the reasoning strategy encounters difficulties, whereas conflict in moral reasoning appears to be more persistent throughout the deliberative process. Our investigation into conflict across three studies introduces new constraints on theoretical models of moral and logical reasoning. We discuss our findings in light of single- and dual-process models of decision-making.

Acknowledgements

I leave with a feeling that I always caught up with Nisheeth too late—if at all I did. But nonetheless, I am grateful that I saw someone living with honesty and conviction.

Traversing life's confusing times would have been incredibly difficult without Prof. Narayanan Srinivasan.

Roshan Jayarajan, thank you for being the rock I could rely on through thick and thin.

Finding Kaushik's friendship again was one of the best things to come out of this degree.

Contents

List of Figures

List of Tables

Abbreviations

AUC	Area under the curve
CI	Confidence interval
CrI	Credible interval
D	Deontological choice of inaction in sacrificial dilemmas
DPT	Dual-process theory
High-C	High-conflict personal dilemma
LME	Linear mixed-effects (model)
Low-C	Low-conflict personal dilemma
MAD	Maximum absolute deviation
SD, SE	Standard deviation, standard error
U	Utilitarian action in sacrificial dilemmas

List of Publications

Publications from Thesis

1. Paper 1 *Judgment and Decision Making.*

2. Paper 2 *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*.

Others

1. Paper 3 *Cognitive Processing*.

"A picture held us captive. And we couldn't get outside it, for it lay in our language, and language seemed only to repeat it to us inexorably."
- Ludwig Wittgenstein, Philosophical Investigations

Dedicated to Vandana and Vijay Shivnekar.

Chapter 1

Introduction

Reasoning is neither static nor uniform; it is a dynamic process that evolves over time which serves diverse functions. Thoughts can shift unpredictably, influenced by context, emotions, and immediate demands. They are often unreliable, reshaping our inclinations as new information emerges or as we reassess prior beliefs. The challenge of aligning our actions with our professed convictions is a well-known struggle. Even thorough deliberation does not always lead to consistent decisions, and at times, we may defer choices entirely, postponing reasoning for later. This fluidity highlights the internal struggles inherent in reasoning, revealing conflicts as we navigate competing perspectives and shifting motivations.

The adaptive nature of reasoning is evident in the inconsistency of our choices, regardless of the stakes involved. For instance, I may deliberate over the price of a shampoo bottle one day, yet on another occasion, I might simply grab my usual brand without hesitation. Consider a different scenario where someone must decide whether to provide intensive medical treatment to an injured animal. Although it may be possible to administer the treatment and attempt rehabilitation, the decision must account for whether it is worth subjecting it to pain, especially when its chances of survival are slim and medical resources scarce. The aspects of the context we prioritize before making a decision in such situations can differ significantly. We may take the time to reflect on certain issues while disregarding others. Some situations introduce complex considerations that demand deeper thought beyond the initial decision, whereas others allow for quick and straightforward resolutions without the need for second-guessing. As a result, we are often aware that our preferences are unstable and choices inconsistent.

Our idiosyncratic predilections and tendency to pay moody attention to the particulars of the decision context shape how we reason. Reasoning, being extended in time—much more so than quick perceptual decisions, like deciding whether the traffic light has turned red—often allows contrasting motives and details to sway us from making a straightforward choice. Thoughts come to us one-by-one but do not necessarily build a single coherent narrative and as a result, we are unmistakably aware of the mental conflict borne from our contemplation.

Yet, the dynamics of the ever-adapting and, on numerous occasions, fickle nature of the conscious thought remain largely absent from current theories of reasoning in psychology and cognitive science. These theories often overestimate the patterns in reasoning while also underestimating its operations. Deciding may sometimes be easy, with a judgment or an opinion seemingly popping out of nowhere in our minds. These quick decisions are thought to be devoid of reason. Any reasoning generated is a justification of the already committed choice. We are especially resistant to reasoning if it does not conform to our views and opinions. Some have also argued that the primary function of reasoning is not to arrive at a decision but simply to defend our actions and choices to others [113]. These tendencies are reflected in our continual susceptibility to biases that reveal the absurdities and inconsistencies in our choices when we "give in" to our intuitions.

While the intuition-based explanations of our choices undermine the diversity in the experience of reasoning, other accounts overestimate the predictability of thoughts. Thinking and reasoning keeps us in a state of flux. Conflict is a regular part of our daily decisions, whether the choices are trivial, such as deciding between tea or coffee, or more significant, like determining the best course of treatment for an ailing parent. When we deliberate such choices, we consciously consider multiple lines of arguments and counterfactuals before making a decision. Phenomenologically, these thoughts seem to appear to us sequentially, with different arguments leading us to tentatively prefer different options one after another. We fluctuate in preferences mentally, vacillating back and forth between options as different considerations reveal themselves as we reason.

This intricate account of consciously experienced conflict is absent from theories about reasoning. Take the deliberation mechanism proposed in moral reasoning, where people are thought to engage in reasoning when their intuitions are in conflict and they cannot reach a clear judgment. These theories often assume that intuitions precede reason and thus impose a strict pattern on how judgments are made. For instance, avoiding actions that cause harm should take precedence in moral judgments, but is that always the case?

Suppose a self-driving car faces with an unavoidable crash. The car's programming must decide whether to steer towards a group of pedestrians or a single individual to minimize overall harm. The default assumption might be that the car should prioritize minimizing total harm, thereby steering towards the fewer number of people. However, such a principle cannot be uniformly defended—and it is not [4]. Some might argue that the car should prioritize saving the single individual, perhaps because they view the individual's life as more valuable due to personal biases or specific circumstances. This example illustrates that the expected pattern of prioritizing harm minimization does not always align with actual reasoning processes.

In short, decision-making theories often attempt to fit reasoning in set patterns or discount rationalizations as largely ineffectual. But the very irregularity in our judgments and flexibility in considerations that sometimes pave the way for those judgments are indicative of the creativity in how we navigate the constant push and pull of slew of demands on our choices. Our decisions do not stand in isolation but are vetted by how confident we feel about them. We are vividly aware of how uncertain our judgments can be. Frequently vacillating in our convoluted reflections may make us less confident in our justifications and judgments. This intricate inner picture of our mental lives, as it unfolds over the time course of deliberating over decisions, is missing in the current theories of reasoning.

1.1 Gaps in capturing reasoning dynamics

Some models categorize decision-making into fast and slow processes, suggesting that the cognitive mechanisms driving extended deliberations are distinct from those behind rapid, intuitive judgments. Conflict is typically seen as the outcome of the interactions between these two systems, especially when these processes cue uncertainty in the ultimate judgment. Even though the process of reasoning is likened to one wild and one tamed horse pulling on Plato's chariot, with a sharp contrast proposed between intuitive and reflective reasoning, it fails to capture the free-flowing nature of our thoughts. Our daily experience of reasoning is filled with moments of conflict, and we are often keenly aware of them. When deciding becomes challenging, we weigh competing arguments and entertain counterfactuals. Various considerations pull us in different directions which shape and reshape our preferences, swaying our choices. This rich spectrum of experience is rarely reflected in decision-making and reasoning theories.

The primary explanandum for theories of reasoning in psychology has been to describe the tendency of producing certain judgments in specific contexts, as well as to lay out

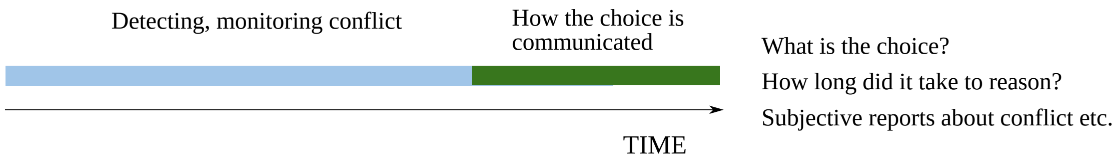

Detecting, monitoring conflict
How the choice is communicated
What is the choice?
How long did it take to reason?
Subjective reports about conflict etc.
TIME

FIGURE 1.1: Conflict in reasoning and its measures. The horizontal bar represents a deliberation which ends in a response. Response dynamic tools are measures of post-decisional conflict which is reflected in how a choice is communicated under controlled experimental settings. Trial-level summaries are collected at the end of the trial and are compared across different conflict conditions. These measures assume that conflict is reflected either in people's responses or in how those responses are generated *after* the conflict has been experienced.

mechanisms for reasoning and problem-solving. Naturally, then, these models hypothesize about conflict experienced during reasoning, proposing it either as a process or competition between choices, aspects of the decisional context, or processes that support them. But the measures or tools used to test these theories are highly impoverished, partially due to the extended nature of reasoning in time. Difficult contexts may demand significant cognitive resources to reason. Even after extended and intense deliberation, reaching a definitive conclusion is not always guaranteed. Consequently, methods used to assess reasoning often compel individuals to provide a response, regardless of their level of conviction or the depth of their contemplation. By obtaining an explicit judgment, researchers can analyze and compare the reasoning process that led to that decision across different dimensions. We refer to these measures which are employed at the end of the deliberation activity to infer underlying conflict as trial-level summaries. Trial-level summaries, like choices and their latencies (time taken to respond), are assumed to reflect a summary experience of reasoning. Measurements are aggregated across trials and then compared to understand broader patterns and insights into the reasoning process.

Figure 1.1 provides a schematic overview of this process. Imagine you are given a problem to solve. You may or may not immediately detect conflict or competition between motivations supporting contrasting alternatives as you work through the problem. Once you feel like you have reached a decision, you communicate your choice. The tools that are used to analyze how an individual got to a decision usually come into play at this final stage, after the choice has already been made. The experience of conflict is inferred by evaluating the choice reported, comparing it to some normative expectation, assessing time taken to respond etc. But these measures only capture the end of reasoning once the reasoner has settled on an answer.

Other measures like the response-tracking tools can also be employed at the end of the trial. These tools, which record judgments in controlled experimental settings, are frequently paired with tracking physical movements during the decision-making process. Researchers in moral and logical reasoning (which are the two types of decision contexts we focus on in this thesis) have utilized methods like hand-, mouse- or eye-tracking paradigms to map the underlying processes driving decisions [82, 134, 141]. In movement-tracking studies designed to capture conflict as it unfolds, the assumption is that motor movements during a given period reflect the cognitive processes occurring at that time [26, 98]. But the response dynamics methods, too, lack temporal granularity because they capture measurements only after the reasoner has potentially detected and resolved the conflict between alternatives. As a result, both trial-level measures and response dynamic tools offer indirect insights into the experience of conflict and fail to fully characterize its depth.

While we test predictions from the two-systems framework in reasoning, alternative approaches also exist. One such approach is accumulation-to-threshold models, which suggest that decisions are made by accumulating evidence for each option until a threshold is reached. These models have been widely applied in perceptual decision-making tasks such as deciding if the net movement of a group of moving dots in a display is rightward or leftward [70, 71]. Recently, evidence accumulation mechanisms have also been proposed as an explanation for reasoning tasks [82]. Unlike the two-systems view, these models argue that decision-making is driven by a single mechanism, irrespective of how long the decision takes to reach a conclusion. These models align more closely with our experience of reasoning. When we are in a tough spot that forces us to pause and think, we entertain arguments and justifications for different choices. Our preferences swing from one option to another as we deliberate more deeply. Under the evidence accumulation mechanism, each consideration can be viewed as evidence accumulated in favor of the corresponding choice (see [13, 108]). Although we do not test the single-process mechanism in this thesis, we return to this idea in the final chapter in light of the evidence presented.

1.2 Tracking conflict in reasoning

Deciding whether to binge-watch another episode or retire early for a more productive morning? Sometimes the battle takes its time, underscoring the importance of time-sensitive and pre-decisional measurements of conflict. These measurements are essential for drawing meaningful inferences about reasoning processes that unfold over time and for testing predictions from theories where conflict is a central component.

We sought to track conflict while keeping two broader considerations at the forefront, as suggested by others elsewhere [146]. First, we wanted to track conflict *in time*. As we have already touched upon above and explain in detail in the next chapter, trial-level summaries offer a poor temporal description of the psychological processes that support reasoning. For instance, response latencies can tell us which problems typically took longer to solve, but they cannot discern when, or if at all, people considered conflicting information in the problem. Response dynamics, like the mouse-tracking method, are also confined to the end of reasoning, when people are ready to communicate their choice. These methods have consistently worked in perceptual decision-making or categorization tasks, but because reasoning is inherently spread over time and entertains variability in functions and patterns, we believe the inferences drawn from trial-level summaries are limiting. Our second concern when selecting tools to track conflict was minimizing the interference they might pose to the task itself. Think-aloud protocols, in which participants verbalize their thoughts while solving or reasoning about a problem, may provide a finer temporal window into how we reason, but it is unclear how they may affect task performance itself. Physiological measures like pupillometry, electrodermal activity, and heart rate tracking fare better on this dimension.

With these tools, we tested specific predictions from well-established theories in reasoning, which are reviewed in the next chapter (Chapter 2). Over three studies and seven experiments, we developed and tested the efficacy of measures that tracked conflict increasingly closer in time. We began our investigation in Chapter 3 by tracking conflict at the response level using mouse-tracking metrics. Previous studies that have used mouse-tracking to study reasoning, especially moral reasoning, have had little success in replicating the predicted findings [82, 103]. Although the theories themselves might be lacking in their predictions, we argue that response dynamics are also inadequate because they are employed at the tail end of reasoning. We demonstrate that these tools exhibit far more diversity than anticipated, raising questions about their interpretability using conventional approaches.

Our conscious experience of conflict is still missing from this picture. Indecision, marked by constant shifts in our preferences, is common. We wanted to account for these vacillations between alternatives to build a more comprehensive picture of conflict. In Studies 2 and 3, we argue that these switches in preferences can also be used as indicators of internalized conflict. Frequent switches between alternatives may be indicative of an unstable choice. We simply asked participants to press a key associated with a preference while they were reasoning, such that consecutive dissimilar key presses highlighted that the participant had vacillated during that period (Chapter 4). The task was simple and did not require participants to verbalize their thoughts as they deliberated, making it less intrusive than other concurrent measures, such as think-aloud paradigms. Additionally, in Study 3, we used eye-tracking to test whether conflict, identified by switches in preference, showed a signature in eye-tracking measures, such as pupillometry and average fixation duration, that provided an even closer window into preference buildups and reversals (Chapter 5).

As with any new method for measuring a psychological phenomenon, it was crucial to establish both validity and reliability. We internally validated conflict measurements by correlating them with subjective ratings and externally through well-established and replicated indicators of conflict in reasoning. Since reasoning has been studied in diverse contexts, our goal was to capture the experience of conflict regardless of the specific context that triggers it. To achieve this, we validated our Switch paradigm using problems where participants either had prior strategies for solving or no formal training.

Moral dilemmas, for instance, represent the latter, where conflicting ethical principles often create tension, and individual differences in moral preferences, as well as specific problem details, can influence the decision. Here, there is often no clear right or wrong answer. On the other hand, logical problems like syllogisms can be solved using formal strategies with a definitive correct answer, offering a clear standard for comparison. This distinction was especially relevant in establishing the reliability of our paradigm in different contexts. In our final empirical study (Chapter 5), we explore how conflict signatures in eye-tracking measures might shed light on these differences—being more sustained in moral reasoning and more staggered in logical reasoning.

Chapter 2

Literature Review

Empirical study of reasoning has long been a central focus in cognitive psychology. Researchers have developed various theories and methodologies to understand how individuals navigate complex dilemmas, balance conflicting information, and arrive at judgments. Conflict in reasoning, thereby, has also received substantial attention from researchers. It has often been conceptualized as a competition between preferences and operations that underlie them. Among these, the dual-process theories have garnered significant attention, positing that human cognition operates through a mechanism which is a result of an interplay between the fast-intuitive and the slow-deliberative processes.

In this review, we will explore both the conceptual and empirical foundations of conflict in reasoning by examining key theoretical models and findings. We begin by discussing models proposed in moral and logical reasoning research, focusing on the testable predictions they offer. Next, we discuss the methodologies used to study conflict, including trial-level summaries and process tracing methods. These approaches offer insights into how conflict is measured and analyzed in reasoning tasks. By synthesizing the existing research, we set the stage for the empirical evidence presented in the next three chapters. We will explore existing and novel measurement tools that test predictions from established theories to further our understanding of this complex and dynamic process of reasoning.

2.1 Moral reasoning

Moral psychology underwent a significant transformation around the turn of the 20th century. Earlier research had primarily focused on describing the development of ethical concerns in children and young adults. Building on Piaget's cognitive theory of development, Kohlberg introduced a three-level model of moral development, which outlines distinct stages through which an individual's moral reasoning evolves over time. This model focuses on how moral character is shaped and refined as people progress through each stage [101]. The first stage, the *pre-conventional stage*, mainly centres on the self-interest of the actor, such as avoiding punishment and gaining rewards. The next stage, the *conventional stage*, involves concerns that extend beyond oneself, such as adhering to societal conventions and understanding their role in maintaining social order. Finally, the *post-conventional stage* is characterized by reasoning about morality in terms of universal principles. Kohlberg's influential theory, along with his use of moral dilemmas to study moral reasoning, left a lasting impact on both the subject matter studied under moral cognition and tools used for research.

By the end of the century, rather than attempting to classify individuals into rigid moral stages—a method that had proved difficult to reliably implement [151, 164]—researchers turned their attention to the cognitive mechanisms underlying moral decision-making [85, 86, 87]. This shift was marked by basing moral judgments in emotional states and considerations of consequences. Jonathan Haidt's social intuitionist model, for instance, placed intuition at the center of moral judgments [83]. He proposed that moral inclinations are largely instinctive and rarely reasoned. Any reasons one may generate in support of our moral motivations, beliefs, and choices are simply post hoc justifications of the intuitions that come to us without reasoning. Just as we instinctively respond to potential threats, moral intuitions arise swiftly and without conscious deliberation. For example, the emotion of disgust, which evolved to protect us from exposing ourselves to noxious smells and foods, is often linked to moral transgressions like incest, cannibalism, and betrayal (though evidence supporting this link remains debated—see Landy and Goodwin (2015) [109]). Any reasoning provided for these quick reflex-like judgments often only serves to rationalize the judgment, rather than preceding, and thereby informing, the judgment itself [83]. Hence, this model argues that people's judgments in moral situations are often not founded in conscious reasoning.

Around the same time, another influential theory brought moral cognition under the broader framework of decision-making by proposing that moral decisions are supported by both intuitive and deliberative processes. The general framework of dual-process theory (henceforth, DPT) posits that there is a qualitative difference in the decisions that are arrived at quickly and those which follow elaborate deliberations [18, 32, 48]. Intuitions, like in Haidt's theory of moral judgment, come to us without reasoning and are hard to justify. The underlying operations that support these quick, intuitive judgments are frequently referred to as the System 1. Over the years, System 1 has been described as an autonomous, reflexive system that is intuitive, emotion-based etc. Contrarily, the System 2 is more calculative. It employs deliberative strategies by using the decision context to elect a judgment. System 2 has been argued to support slow, capacity limited but conscious processing of information, that usually correlates with consequential decision making [52]. The DPT has been arguably the most diversely popular theory in decision-making. It has been proposed as an explanation of behaviors in wide variety of fields including but not limited to moral cognition [6, 81], logical reasoning [35, 46, 49, 50], social cognition [163], misinformation susceptibility [22, 132], social media use [181], cooperation [17, 90, 137], behavioral economics [8, 38, 72] etc.

Joshua Greene and his colleagues were the first ones to formally propose that moral decisions are supported by the interaction between the fast and slow systems. They also popularized particular kind of dilemmas called sacrificial dilemmas to study moral decision-making. These dilemmas pit ethical principles—usually deontological and utilitarian—against each other such that choosing one option rules out the reasoner endorsing ethical principles behind the options not chosen. The dilemma popularly known as the Trolley Problem is such a sacrificial dilemma. A runway trolley is going to hit and kill 5 workers unless it is intercepted. In the Switch version of the problem the trolley can be stopped by throwing a switch that diverts it onto another track. However, a worker on the diverted track is sure to be killed if it hits him. In the Footbridge version of the same problem, instead of throwing the switch, the trolley can be stopped by pushing a large person to his death onto the track from a footbridge hanging on top of the tracks (this alteration of Phillipa Foot's Switch version (1967) was proposed by Judith Jarvis Thompson in 1976 [55, 169]). Therefore, people have a choice with an action that maximizes the number of lives saved either by throwing a switch or pushing a person (the utilitarian principle) or not taking any action to adhere to the principle of "thou shalt not kill" (the deontological principle). The utilitarian principle entails a commitment to act, while deontology means omitting to endorse the same action. Note that the trade-off between lives saved and lost if the action were to be taken is exactly the same across both dilemmas with only the action that brings about these consequence varying between the two versions (see Figure 2.1 for a schematic depicting both versions). Yet, people are generally inconsistent in their responses [4]. Most people endorse the action in the Switch version sacrificing one to save five, but not in the Footbridge version. This curious inconsistency in people's judgments has led researchers to employ systematic variations in such dilemma structures to formulate theories about the factors influencing people's reasoning in specific contexts.

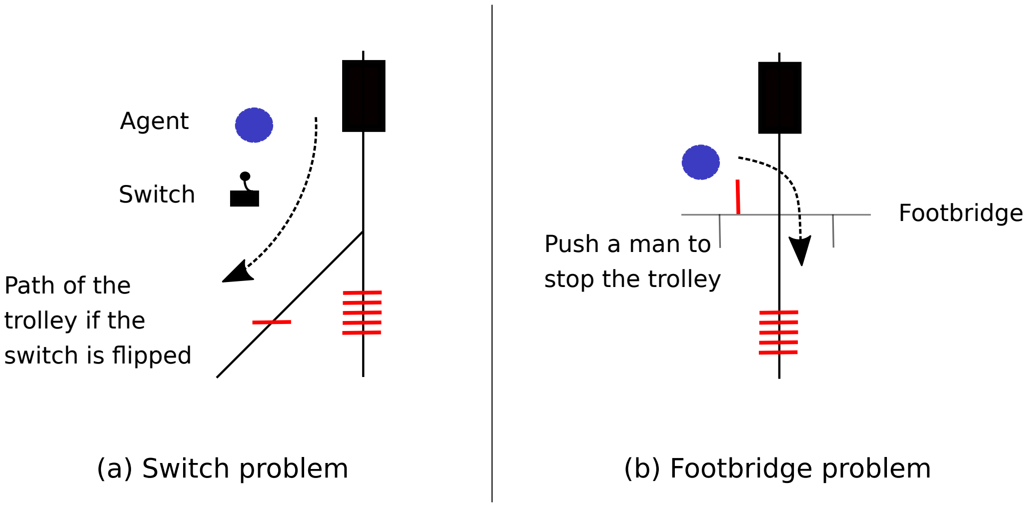


FIGURE 2.1: A schematic depicting the two popular versions of the trolley problem. Notice that both versions have the same five-to-one trade-off between lives saved and sacrificed with actions that bring about that outcome being different. Taken with permission from Shivnekar and Srinivasan (2024) [148].

These experimental paradigms align closely with the DPT framework of moral psychology [6, 28, 78, 80, 81]). Moral dilemmas like these ostensibly offer a much clearer interpretation of which principle is preferred in specific contexts and with particular set of actions. But this interpretation of the choice is conflated with rejecting the principle behind the unchosen alternative, an assumption that reportedly does not hold up under empirical testing [27].

Under the hood of DPT, different mechanisms have been proposed to explain how System 1 and System 2 interact in the reasoning process. According to the corrective or default-interventionist model of DPT, conflict is often attributed to the interplay of these two systems (Figure 2.3). When System 1's quick but strong, emotion-based response is to be overridden by a more calculative and resource demanding System 2, the conflict is likely to occur. To elaborate with an example pertinent to moral reasoning, in impersonal dilemmas where the action operates indirectly through mechanistic mediations, like in the Switch version, there is only a weak inclination for the deontological principle from System 1. This preference is easily overridden by System 2's strong preference for the utilitarian principle. This results in the majority of people choosing the utilitarian option in such dilemmas. On the other hand, in personal dilemmas, which involve an action that directly harms the victim through muscular force, like pushing and smothering, most individuals typically opt for the deontological alternative because it has a robust activation that System 2 cannot overcome [28, 78, 117]. Greene and colleagues reported that when individuals do choose the atypical utilitarian alternative in personal dilemmas supported by System 2, the conflict arising out of two systems cueing opposing judgments must be resolved first. This was hypothesized to add extra time to commit to the atypical response, resulting in longer response latencies [81]. It is worth noting that the personal-impersonal split of dilemmas simply based on the proximity of the action to the consequence does not consistently reflect in typical responses [13, 82, 111]. Therefore, the yardstick for replicating previous findings in this thesis was replicating the pattern in choice data under different operationalizations of conflict which did not rely on this distinction.

So far, conflict has been attributed to a mechanistic phenomenon arising from the competition between both systems for the control over the final judgment under the default-interventionist model of DPT. Both systems have a preferred response and they are employed sequentially. Specific predictions can be derived from this assumption. System 1 kicks in first to generate a quick deontological response, which—if enough cognitive resources are available—can get updated through System 2's operations to endorsing the utilitarian judgment. Hence, an individual's preference can be expected to either stay deontological, or get updated from deontological to the utilitarian judgment. But utilitarian to deontological shift is not anticipated. People's tendency of choosing one response over the other in specific moral dilemma types and the time they take to decide are also used as indicators of conflict in experimental studies of moral decision-making (see Baron and Gürçay (2017) for alternate perspective [13]) [81, 128]. Greene and colleagues' view is an influential perspective in moral reasoning research. We reconsider this proposal later in light of measurements and paradigms employed to test these predictions and their findings.

Another mechanism of the interaction between the fast and slow systems is proposed by De Neys [32]. Like the default-interventionist mechanism, De Neys' hybrid model, too, proposes a sequential architecture of the two systems, with System 1 operating on the information before System 2. Unlike the classical default-interventionist model, though, it does not assume responses are exclusively generated by a specific system. Instead, System 1 generates multiple intuitions at the same time. In context of moral reasoning, both intuition of action commission and action omission can come from System 1, albeit with different activation or associative strength between the stimulus and the generated intuition. The activation strengths of different intuitions are different and can change over time, too. System 1 is charged with the responsibility of keeping track of how strengths change in comparison to an uncertainty threshold beyond which System 2 is cued to step in. If a dilemma consistently sees people preferring inaction, then according to the hybrid

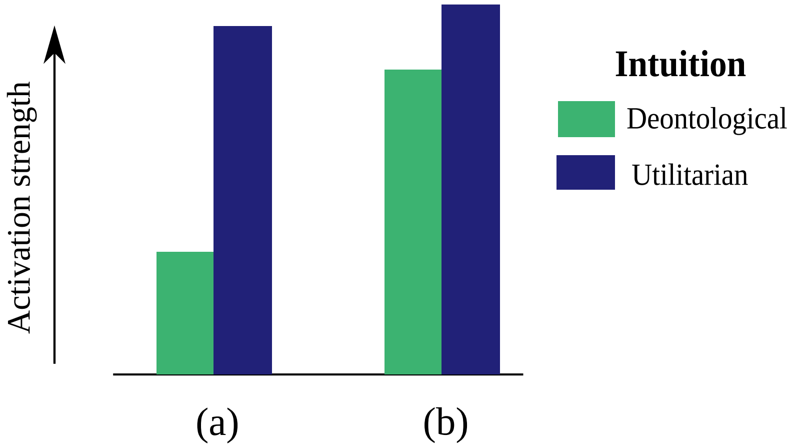


Figure 2.2: Hybrid DPT mechanism of moral reasoning. System 1 can cue both deontological and utilitarian judgments. If prompted early in the reasoning process, System 1 emits the response that has the largest activation strength is selected as a response. In both (a) and (b), System 1 would produce the utilitarian response. System 2 is engaged only if the difference between the strengths of the two responses is small, like in (b).

model, the activation strength is high for the deontological response. If two intuitions tied to distinct responses have comparable activation strengths then instead of System 1 electing a response, System 2 ensues deliberations to weigh in on the information. Reasoning processes such as deliberating, generating new responses, suppressing others, or seeking reasons to justify a response are supported solely by System 2, which operates on the activation strengths of intuitions, ultimately allowing System 1 to select one as the final response (see a schematic depicting how a response may be elected based on activation strengths in Figure 2.2).

Bago and De Neys proposed that moral decisions, too, can be explained with a version of hybrid model of DPT [6]. By demonstrating the tendency of participants to prefer utilitarian response when under time-pressure, the authors argued that responses are not exclusively constrained to a system. The two systems instead differ in how the activation strengths of the responses are compared. System 1 simply compares the absolute strength of all the intuitions produced. If participants are asked for a judgment at this point in reasoning then the alternative with the highest activation is chosen. On the other hand, System 2 compares the relative strengths of intuitions. Whether or not the initial choice gets updated is dependent on the activation of the competing alternative. In other words, conflict is defined as the comparison of the strength of each intuition.

Over the years, the two-process view of reasoning has remained popular in decision-making, with proposed mechanisms being continuously updated and altered. While some models within DPT make strict predictions, newer mechanisms have been adapted to accommodate new empirical data. For example, default-interventionist models propose a one-step updating mechanism of preferences, where System 1's elected response must be overridden before System 2's calculative judgment can become the final choice. We show across three studies that people's preferences do not necessarily update in a particular order. In Studies 2 and 3, we demonstrated that people switching between alternatives multiple times before settling on a choice is far too common in reasoning. While the hybrid model can theoretically be extended to account for multiple shifts in preferences, simpler models that do not posit a division in reasoning processes can also explain these shifts, avoiding the significant shortcomings of the two-process view. These issues are discussed in depth, in light of the empirical results from this thesis, in Chapter 6.

2.2 Belief-bias in logical reasoning

Just like moral dilemmas in moral decision-making research, logical reasoning has been studied widely by using syllogisms. Syllogisms have two premises and a conclusion linking three terms to each other with connectors called *moods* viz., "all", "no", "some", "some ... not". Figures of syllogisms dictate how the terms are arranged in a syllogism. Suppose a syllogism has 3 terms, P, Q, and R. If the conclusion is of the order P-R, then the premises can be arranged in four distinct figures (P-Q, Q-R; P-Q, R-Q; Q-P, Q-R; Q-P, R-Q). The task in a deductive reasoning study with syllogisms is to decide if the conclusion is logically valid, if the premises are assumed to be true. Following is an example of a valid syllogism:

All P are Q.
No Q are R.
Therefore, no R are P.

Often in syllogistic reasoning tasks context-relevant terms are used instead of abstract terms. People find it difficult to ignore their expectations tied associated with these terms when deciding the validity of a syllogism. Take the following example:

Some dogs are not pets.
Some animals are dogs.
Therefore, some animals are not pets.

Table 2.1: The results display a typical belief-bias effect, where participants are more likely to accept a syllogism as valid when its conclusion aligns with general expectations and is believable, as reported by Evans et al. (1983) [49].

	Believable	**Unbelievable**
Valid	89%	56%
Invalid	71%	10%

This syllogism is invalid because the conclusion does not necessarily follow from the premises [1]. People often inaccurately think that the syllogism above is valid. In fact, beliefs about the conclusions systematically interact with logical status of the syllogism such that people are more likely to judge a believable syllogisms valid than an unbelievable syllogism as valid. Particularly, there is also an interaction effect between the validity and believability such that the rates of conclusion endorsement differ in valid and invalid trials depending on the believability of the conclusion (see Table 2.1 for a representative result) [49, 92, 120, 125].

The belief bias effect is sometimes attributed to the conflict between believability of the conclusion and whether the conclusion logically follows from the premises. The conflict can be resolved accurately—and unlike in moral dilemmas, there is an accurate answer in syllogisms—if the response cued by prior beliefs is suppressed. In no-conflict problems where both these factors are congruent (valid-believable and invalid-unbelievable syllogisms), no such inhibition has to take place and, hence, people are generally highly accurate. According to the hybrid model of DPT of logical reasoning, System 1 cues intuitions about the validity that can be congruent with logic or beliefs, thereby detecting the disagreement between the two responses in conflict syllogisms (valid-unbelievable and invalid-believable). The resolution of this conflict is carried out by System 2's deliberations. To suppress belief-based responses, reasoners must identify the conflict between belief and logic, a process contingent upon their familiarity with the logical rules and the application of their own priors within the task [30].

Different theoretical frameworks, such as selective scrutiny, misinterpreted necessity, and the theory of mental models, present varying predictions regarding the sequence of preferences over time [15, 37, 45, 94]. The selective scrutiny model proposes a DPT explanation with a default-interventionist mechanism to explain the belief-bias effect by associating belief-based and logic-based judgments to System 1 and System 2, respectively. Reasoners

[1] To see how this syllogism is invalid, consider the following version of the premises:
Some *stray* dogs are not pets.
Some animals like *stray dogs* are dogs.
In this case, the animals that are dogs (stray dogs) are indeed not pets. But if the animals that are dogs were pets (like pet dogs), then the conclusion "some animals are not pets" would be false.

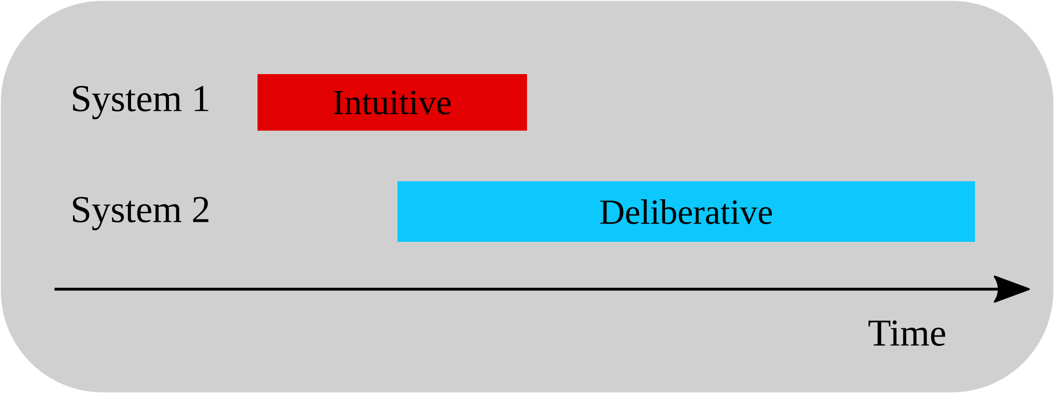


FIGURE 2.3: The model posits that each system has a characteristic response (e.g., deontological principles in moral dilemmas or the believability of a conclusion in syllogisms). Conflict arises when the responses from the two systems are mismatched.

purportedly begin by assessing the likelihood or believability of the conclusion. If the conclusion is deemed unlikely, only then do they scrutinize its logical status. Hence, the errors as seen in the Table 2.1 are more in the conflict syllogisms when belief priors clash with the logical status of the syllogism. In short, the selective scrutiny models predicts that judgments will be corrected more often in valid- and invalid-unbelievable syllogisms.

Conversely, the misinterpreted necessity model proposes that belief-bias results from individuals failing to understand the concept of syllogistic validity. In the example involving dogs and pets, the premises can be modeled in multiple ways. Such syllogisms are called multiple-model syllogisms and are valid *only if* the syllogism is valid in all arrangements of the premises. In other words, if a syllogism is valid in some models but invalid in others, it is logically invalid. According to the misinterpreted necessity model, people may start reasoning logically, but if the validity cannot be determined definitively, they rely on whether the conclusion is believable. Consequently, belief-bias is most pronounced in multiple-model syllogisms, where deciding the validity status of a syllogism is complex and ambiguous, and less so in single-model syllogisms, which can be assessed through a single consistent set of relationships among the terms [122].

Similarly, the mental model theory argues that individuals begin reasoning by constructing mental representations of the given premises and then evaluate the conclusion based on its logical validity. If the conclusion lacks logical alignment, it is rejected; only logically fitting conclusions prompt consideration of their believability, with unbelievable conclusions leading the reasoner to form alternate models from the premises. This reasoning process is possible only in syllogisms that permit the construction of alternate models, such as multiple-model syllogisms.

All four models described above provide specific predictions about how people navigate reasoning with syllogisms. The selective scrutiny and misinterpreted necessity models predict which type of response is given priority (belief-based or logic-based, respectively). Additionally, the hybrid, misinterpreted necessity, and mental models theories also predict specific kinds of reasoning errors. These predictions were examined in Studies 2 and 3.

2.3 Existing measures of conflict

Although conflict is a central concept in many decision-making and problem-solving theories, directly measuring this phenomenon remains challenging. How do we know an individual is conflicted while thinking? One way is to interrupt their deliberations in some set fashion to sample their experience at that point in time. Any external interruption, however, can become repetitive and interfere with the process under inspection. We can also ask participants to provide continual verbal account of their thoughts which can then be rated and categorized. Most such direct approaches risk distorting the task itself. Hence, researchers frequently rely on indirect measurements that are collected at the end of the trial. In this section, we explore both indirect and direct methods used to study the mechanisms of reasoning, organized into two subsections: "Trial-level summaries" and "dynamic measures of conflict." These subsections review various strategies employed by researchers to infer and examine the experience of conflict during reasoning. Since the focus of this thesis is on moral and logical reasoning, this review of methods is limited to measurements and paradigms used in these contexts. We also comment on the implications of employing these measurements to the theories of reasoning described in the section above.

2.3.1 Trial-level summaries

Conflict measurements are often relegated to the end of a trial. These include latency of response production [34, 74, 79, 81, 174], normative expectations of behavior within the given choice framework [50, 73], and subjective ratings [61, 115, 130]. We refer to these as trial-level summaries because they encompass the entire trial in a single measurement produced at the end. These measures are then aggregated and trials that are supposedly conflicting and non-conflicting are compared. While these measures may be less intrusive than others, they allow an indirect inference about conflict because they are not employed concurrently. Below, we review some of these measures and paradigms and their shortcomings in the context of reasoning studies.

Response times are commonly used to measure conflict, based on the assumption that experiencing conflict delays decision-making. If two alternatives are in competition, detecting, monitoring, and potentially resolving the conflict adds to decision-making time, making trials longer. The DPT proposes that the conflict is ensued when two intuitions or processes are in conflict. Given that the fast and slow operations operate on different timescales, relative difference in time taken to respond to stimuli is taken as evidence for conflict in decision-making.

The System 2 operations are hypothesized to be slow and calculative. If a problem produces conflict among alternatives, individuals would take longer to scrutinize the information before making a decision. In one of the earliest studies to hypothesize choice competition in reasoning, Greene's default-interventionist model of moral reasoning [81] was tested. Participants in their study gave judgments on a series of moral dilemmas, either personal or impersonal (see Section 2.1 above for the distinction between these two types). The authors argued that deontological responses were emotion-based because the perusal of the personal dilemmas, which typically produce such responses, co-occurred with activation in areas assumed to predominantly respond to affective information. Although mere concurrent activations do not guarantee affective processing, the authors still interpreted the longer trials with an atypical response to personal dilemmas as a sign of conflict detection and resolution in favor of the atypical utilitarian judgment.

To test conflict between systems more directly, some have also used time-pressure conditions to provide a window into early processing in reasoning tasks. These methodologies assume that only System 1 operates on the information when System 2 is overwhelmed by cognitively demanding tasks. Suter and Hertwig (2011) recruited participants to judge moral dilemmas either under time pressure (8 seconds) or with sufficient time to reason and decide (3 minutes) [165]. They proposed that participants who had to respond quickly would not have enough time to reconsider their initial preference (which, according to the default-interventionist model of moral reasoning, would be the deontological omission). Indeed, their data demonstrated that participants in the time-pressure condition were more likely to choose the deontological alternative than those who had enough time to think and decide. Evans and Curtis-Holmes (2005) employed a similar paradigm, demonstrating increased belief bias in syllogistic reasoning under time pressure, and attributed it to the inhibition of System 2 processing under time pressure [50].

Although response times have been a popular indicator of cognitive conflict, subsequent empirical investigations have shown that these findings are often not reliably reproduced, particularly in the domain of moral reasoning. For instance, a meta-analysis by Baron and Gürçay (2017) revealed that individual differences in preference and the complexity of the dilemma, rather than atypical responses, were the primary factors driving response times. Similarly, Bago and De Neys (2019) demonstrated that participants generated both deontological and utilitarian responses under extreme time pressure and cognitive load, designed to diminish System 2 operations. This evidence challenges the assumption of sequential conflict resolution as proposed by dual-process theories. Furthermore, the conclusions drawn from response times are often misleading. While response latencies indicate the end of the reasoning process, they are frequently used to infer earlier stages of preference formation, conflict detection, and resolution. This reverse inference overlooks that additional cognitive processes occurring alongside the hypothesized conflict could contribute to longer decision times. In essence, response times may confound the process of interest with other cognitive functions [107].

Researchers also compare people's preferences to a normative standard. The biases in reasoning are often projected as failures to act rationally. A belief bias, for instance, is often seen as a failure to reason logically. In tasks where deciding the normative standard is rife with controversy, such as moral reasoning, researchers substitute descriptive consensus for them. For instance, Koenigs et al. (2007) operationalized conflict by examining agreement on final judgments in moral dilemmas among individuals [100]. They categorized personal dilemmas into high- or low-conflict groups. A dilemma was considered low-conflict when nearly all participants agreed on choosing a particular alternative. In contrast, high-conflict dilemmas displayed no consistent pattern in judgments, featuring varying degrees of endorsement of an alternative at the cohort level (we tested this definition of conflict extensively in this thesis).

Another way to operationalize conflict is as a shift in preference. If people frequently change their judgment about an issue when asked at different times, it can be taken as an indicator of conflict. Bago and De Neys (2019) employed a two-response paradigm to compare judgments from System 1 and System 2 [6]. To get a judgment devoid of System 2's deliberations, its operations need to be 'disabled'. The stricter the limits on System 2's functions, judgments are assumed to be devoid of any deliberations. Hence, participants are often required to undertake multiple tasks to knock off any supposed involvement of the calculative System 2. In Bago and De Neys' experiment, participants first memorized locations of 4 dots randomly placed in a 3x3 grid displayed briefly before reading a moral dilemma. Soon after reading the dilemma, they had to generate a response under time pressure. This way, the authors argued, System 2 was acutely disabled by limiting time allowed to generate an early response while the additional memorization task impaired reasoning about the problem. After the initial response, participants reasoned about the problem and logged in their final judgment on the dilemma. Results revealed that reasoning patterns are not necessarily of a "corrective" nature. People who chose the utilitarian alternative had already selected the same option even under time pressure. This prediction is not in line with the classical default-interventionist model of DPT, which expects the quick response to be deontological in the case of moral dilemmas. With this, the authors argued that System 1 concurrently generates both deontological and utilitarian intuitions, and utilitarian responses are not necessarily corrective [20, 21] (also see [82, 149, 150]).

A two-step paradigm like this allows for dissecting the decision-making process but only to a limited extent. The paradigm, as employed by Bago and De Neys (2019), rests on the fundamental assumption that the fast and slow systems exist and can be at least partially dissociated. In their experiment, the researchers attempted to dissociate the two systems by disabling System 2 through the imposition of time pressure and additional cognitive load. By doing so, they supposedly isolated the intuitive judgments of System 1, independent of the deliberate, calculative operations of System 2. However, while this approach offers insights into how initial preferences are formed, it hinges on the assumption that System 1 and System 2 function either sequentially or independently, without directly testing this division in their operations. Furthermore, these methodologies are interrupting and have little ecological validity as they overlook the possibility that additional tasks or time constraints may influence reasoning in unexpected ways. Beyond being highly invasive, the paradigm fails to capture the full scope of the reasoning process that occurs after participants report their initial inclinations.

Trial-level summaries, hence, are limiting in important ways. The measurement taken at the trial-level offer little temporal resolution. Judgments, response times, and most other end-of-the-trial measurement offer limited insight into the mechanisms of reasoning, especially because it takes time to reason. Prodding these processes at different point periods necessitates additional assumptions about the nature of the mechanisms underlying them and their interactions. Therefore, direct measures that concurrently track conflict in reasoning are needed to hypothesize and test reasoning as it naturally occurs.

2.3.2 Dynamic measures of conflict

The classical information processing framework posits that behaviors result from a serial process: first, a stable internal representation of the external environment is constructed through perceptual processing, then cognitive mechanisms formulate a response, and finally, this response is executed through motor actions. In broad terms, this view assumes that, for building a plan and executing an action in response to environmental pressure, a unified and stable—though perhaps only partially so—representation of the environment is first constructed. This representation is then used to build and execute the necessary actions. Parallel processing models reject this view in favor of simultaneous interactive processing of functions, generally categorized under perception, cognition, and action. These models propose that behaviors result from a more continuous interaction between an organism and its environment. Consequently, multiple motor programs are simultaneously formed and ready to be deployed in response to environmental demands and the organism's goals. These programs compete for execution through selection processes, although disputes remain about how this selection is carried out. Nonetheless, according to this view, tracking motor programs at any point in the process can be informative about the current state of the system.

Evidence for such processing can be found in neurobiological and behavioral studies alike. Cisek and Kalaska (2010) argue that behavior is a continuous process involving the simultaneous specification and selection of multiple potential motor actions [26]. An organism's environment is constantly in flux, and the organism's attention frequently shifts based on intrinsic or extrinsic demands. As a result, motor control must adapt quickly. They argue that the brain has evolved to support such a mechanism by significantly sharing the neural correlates among perceptual, cognitive, and motor execution, making it difficult to modularize different brain regions based on their participation in either perception, cognition, or motor functions. They propose that the brain's architecture is designed to support the preparation of several actions in parallel. This capability allows organisms to rapidly switch between potential actions, a necessity for survival in dynamic environments. The competition among these programs is biased by the reciprocal connectivity of several different regions distributed over the cerebral cortex. The selected program then receives feedback internally and from the environment through a predictive feedback system, while the system prepares to build and select from competing motor programs once again.

An illustrative example of simultaneously formed motor plans with information acquisition in a perceptual task can be found in the work of Gold and Shadlen (2000, 2003) using a random dot motion (RDM) task [70, 71]. In a typical RDM task, a display of randomly moving dots is shown on the screen. Depending on the difficulty condition of the trial, a portion of these dots move coherently in one direction, either up or down (or left or right). The subject must decide the overall direction of motion. Gold and Shadlen trained a monkey to respond in this discrimination task by making a saccade in the direction of the perceived motion. In some trials, they stimulated the frontal eye field—an area assumed to be responsible for generating saccades—with a brief electric current. The current was sufficiently small to induce an involuntary micro-saccade in the rightward direction. Interestingly, the deviation of this saccade was also influenced by the choice the monkey would eventually make. For instance, if the motion display had coherence in the upward direction, the microstimulation would cause the saccade to deviate rightward and upward, even early in the decision process. Gold and Shadlen argue that this deviation in the direction of the eventual choice indicates that the evidence tracking continually informs motor plans.

This proposed correspondence between perception, cognition, and action has been argued to be useful in measuring often inaccessible mental processes. Conflict evolves in tandem with our thoughts that reveal contrasting arguments and choices to us. Given its dynamic nature, a more effective measurement tool for conflict needs to operate in real-time, tracking changes during deliberation while also being minimally intrusive to avoid substantial interference with the process being studied [146]. Researchers in moral and logical reasoning fields have used think-aloud paradigms, mouse-tracking, eye-tracking etc. to map these processes closely [5, 82, 134, 152, 166].

Response dynamics refer to tools that measure aspects of motor movements made during the communication of a response. Task demands dictate the types of movements typically made to commit to a choice, ensuring that response movements are comparable across trials within an experiment. Consequently, these measurements are captured at the end of each trial. Methods of response dynamics are various ranging from hand–tracking to tracking hand-held devices such as a computer mouse [58, 141, 153, 180]. In mouse-tracking studies designed to measure conflict, mouse movements are utilized to elucidate the temporal dynamics of a choice such as to evaluate whether a specific switching pattern is more predominant than others [82, 103]. However, mouse-tracking measures lack clarity regarding which part of the process is captured or whether the entirety of it is reflected in the response dynamics. Eye-tracking methods, on the other hand, are not necessarily limited to comparing how a response is produced. Some of the eye-tracking metrics operate on the assumption that the decision process is dynamic and gaze position and other aspects of eye movements reveal what information is being favored currently [66, 127]. Based on this assumption, eye tracking paradigms are employed to decipher the moment-to-moment updating of preferences as decisions are being constructed.

This thesis is a study in the measurement of conflict in reasoning, with two primary aims. One, we wanted to track conflict as closely as it unfolds over time. As reasoning is extended in time and often takes much longer than a perceptual decision-making task, the measures needed to be sensitive to the changing experience of conflict. Secondly, we aimed to explore measures that minimized the risk of interfering with reasoning. Imposing time pressure or introducing additional cognitive tasks can inadvertently distort reasoning processes. Given the extensive use of mouse- and eye-tracking techniques in Studies 1 and 3, the following sections offer a detailed review of these methods within the context of moral and logical reasoning research.

2.3.2.1 Mouse-tracking

Mouse-tracking has recently become a popular tool for studying cognition through response dynamics. Since its introduction, it has been used in a wide variety of tasks with few alterations made to the general paradigm (for a variety of tasks employing this paradigm, see [103, 116, 156, 162, 167]). This method was introduced by Spivey et al. in 2005 [157]. The task aimed to uncover whether phonological competition is resolved continually or in discrete intermittent steps. The connectionist model of the mind introduced above hypothesizes that motor movements can reveal continuous and gradual changes in largely inaccessible perceptual and cognitive representations. By tracking and measuring how a response is produced under controlled settings, we can uncover these concurrent processes. On the contrary, discrete models of cognition, such as the models under the DPT framework, propose that the readout from the psychological processes to motor representations is intermittent. Spivey et al. (2005) tested the viability of these two hypotheses in their experiment.

In a trial, Spivey et al. (2005) showed participants pictures of two distinct objects [157]. Each picture was displayed in one of the top corners of the screen, one of which was the target picture. Participants were required to click on the box at the bottom center of the screen to initiate the trial. Soon after doing so, participants heard a word over the headphones that identified the target picture. The task was to simply click on the picture that contained the object that was prompted. The authors manipulated the similarity of the two objects presented in a trial. In some trials, the names of the two objects were phonologically similar (like “candle” and “candy,” as opposed to “candle” and “jacket”). Similar-sounding objects were expected to create competition between their lexical representations when participants made a response. If cognition and action are tightly linked, as in connectionist models, it should be reflected in how the mouse is moved from the bottom edge of the screen to make the response. Specifically, the authors hypothesized that these response trajectories would curve toward the distractor picture before clicking on the appropriate choice. On the contrary, if the motor plans come into action only after the cognitive processes choose and commit to an answer, then the trajectories should be straight, joining the initial click and the click made to record a response in a direct path.

Spivey et al. (2005) demonstrated that their data supported the former hypothesis. They showed that, on an aggregate level, the cursor first moved straight upward before curving in toward the chosen alternative. In trials where the target and the distractor shared phonological properties, participants moved the mouse inward later than in non-conflict trials. This resulted in a larger area under the curve in the conflict trials, bound by a straight line joining the positions of the cursor when the mouse was clicked to initiate and conclude the response and the trajectory itself.

Experiments employing this methodology now have a variety of metrics to index response curvatures (some of them are depicted in Figure 2.4. See Wulf et al. (2019) for a review [180]). Web applications and packages are easily available for implementing this paradigm for a range of tasks and analyzing the data [58, 97, 129]. The mouse metrics generally index either the curvature, complexity, temporal dynamics, or a combination of these indices. For instance, area under the curve and maximum absolute deviation are indicators of curvature. Both of these metrics can be used to help infer the pull on the response by the distractor. However, area under the curve considers pull only from the competitor alternative’s side of the screen by subtracting the area where a convex trajectory may turn outward toward the edges of the screen. Such portions of trajectories cannot be said to be under the influence of the distractor. Maximum absolute deviation, on the other hand, is simply a distance metric indicating the largest “mode” of the trajectory.

Mouse trajectories are not all necessarily curved. They may be more abrupt, such that the cursor may move to one alternative before switching to the other. These movements can be captured by the number of times the direction of the trajectory changed or when the vertical center of the screen is crossed. Some researchers have also looked at the time taken to respond and acceleration profiles. The temporal dynamics of the trajectories can also reveal the sequence in which information is processed. For example, Freeman et al. (2013) studied cultural context effects on race judgments. American and Chinese

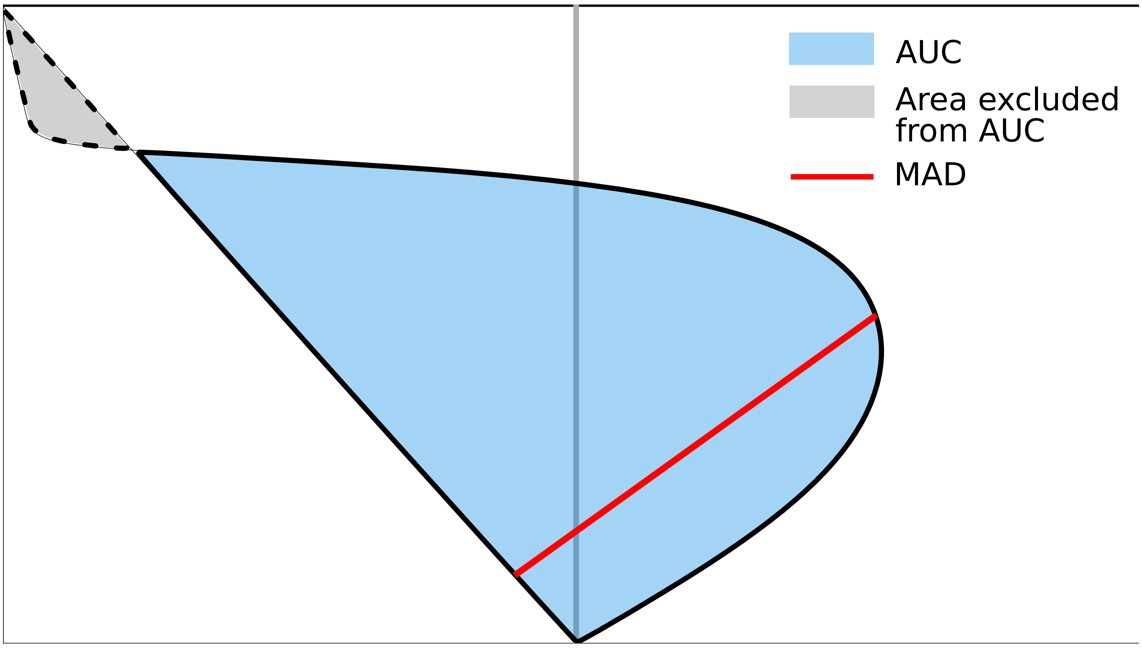


Figure 2.4: Schematic of mouse metrics used in this thesis. Response trajectories originate at the bottom-center of the screen and move toward the top-left or top-right corners where the alternatives are located. The Area Under the Curve (AUC) is represented by the blue shaded region between the observed trajectory and the direct path, excluding the grey area beneath the direct path. Maximum Absolute Deviation (MAD) is indicated by the red perpendicular line drawn from the point of highest deviation to the direct path. Reversals along the X-axis are the number of times the cursor crosses the Y-axis, denoted by the grey vertical line at the center of the screen. This trajectory shows 2 reversals.

participants were tasked with categorizing white and Asian faces based on race. The faces were embedded in background scenes typically associated with either white or Asian neighborhoods. As hypothesized by the authors, curvatures of the trajectories were larger when the face and the background were incongruent. However, the response trajectories of native Chinese participants curved earlier during the response than those of Americans, lending support to the authors' hypothesis that contextual effects influence people from individualistic and collectivistic cultures differently [59].

Mouse-tracking paradigm has also been used in describing the conflict in reasoning tasks. These tasks usually take longer than perceptual and cognitive tasks of shorter duration. Travers et al. (2016) used this methodology to investigate how participants resolve conflict in cognitive reflection test [173]. Koop (2013) and Gürçay and Baron (2015) employed the same method to study how conflict between deontological and utilitarian choices in moral reasoning is resolved [82, 103]. Participants in these studies read sacrificial dilemmas that were presented centrally on the screen, with the action commission and omission options in top-left and top-right corners. Once participants read the problem, they clicked on the initiating button at the centre of the bottom edge of the screen and the corner holding their choice. Curvier trajectories were expected when participants respond atypically in a trial (e.g. endorsing the utilitarian action in personal dilemmas), reflecting the stronger pull from the stereotypical judgment on the response trajectory of the mouse. However, both of these studies failed to demonstrate this hypothesized effect, concluding that DPT fails to capture moral reasoning because preference updates do not necessarily occur in the predicted way. In Study 1 of this thesis, too, we did not observe curvier trajectories for atypical responses. We contend that although moral reasoning may not occur in the fashion anticipated by the DPT view as Koop (2013) and Gürçay and Baron (2015) argue, mouse-tracking may also be a flawed tool to employ in reasoning tasks. Mouse-tracking is essentially a late- or even post-conflict measure that is believed to capture the competition between responses after people are ostensibly done deliberating and ready to answer. In categorization tasks like in Spivey et al. (2005) and Freeman et al. (2013) described above, the time window to reason and respond is short. Therefore, the temporal distance between experiencing conflict and recording a response is reduced, allowing the possibility that some competition between alternatives can still be reflected in trajectories.

Contrarily, participants can read a moral dilemma, detect and resolve conflict between alternatives before they move their mouse to record their response. Because reasoning often takes considerably longer to complete, conflict may arise well before it is captured by measurement, leaving enough time for it to no longer influence the movement of the mouse. This delay could jeopardize the effectiveness of mouse-tracking as a reliable indicator of conflict. Additionally, the other two studies using mouse-tracking in moral dilemmas were either focused on testing the personal-impersonal distinction between dilemmas or hypotheses assuming the distinction is informative. In our study, we tested the level of conflict expected to occur while reasoning in moral dilemmas based on a literature-established operationalization of conflict. Even then we failed to produce support for the pattern in mouse movements reflecting increased competition between alternatives in atypical trials.

In summary, mouse-tracking has been used to understand cognitive processes by assuming a strong coupling between cognition and motor actions in real-time decision-making. The method's ability to measure continuous and gradual shifts in mental representations offers a powerful tool for testing theories of cognition. By capturing the subtle dynamics of response movements, this technique can reveal how cognitive processes unfold, particularly in situations of conflict or competition.

2.3.2.2 Eye-tracking

Eye-tracking is a widely available, non-invasive tool that provides a plethora of metrics to trace cognitive processes. These metrics capture different aspects of eye-related events—like saccades and fixations—and link them to other observable behaviors. Saccades are rapid, largely ballistic movements of the eyes between two locations. Fixations, on the other hand, are stable periods when the eyes focus on an object at the fovea. Most modern eye-trackers can not only determine the location of eye fixations but also provide additional data, such as fixation duration, saccade amplitude, velocity, acceleration and pupil size.

Broadly, these measures can be categorized based on whether they inform us about the location of gaze or its temporal properties. The underlying assumption for location-based metrics is that when an object is fixated on, it is processed on priority. This is because limited information is acquired while the eyes are in a saccade (the eye-mind hypothesis by Just and Carpenter (1976) [95]). Therefore, the location of fixation is often used to dynamically trace attention allocation in a visual display. For instance, the sequence of fixations can infer the prioritization of objects in the environment. In a study investigating belief bias using eye-tracking, Ball et al. (2006) demonstrated that premises were viewed longer in conflict syllogisms (invalid-believable and valid-unbelievable) after the conclusion was read, thus establishing its belief status [7]. This does not align with either the selective scrutiny or the mental models theory of logical reasoning. Both theories would predict that premises in only unbelievable syllogisms should result in longer inspection times, as participants would need to reconstruct these syllogisms. However, the observed sequence of eye movements suggests that individuals detect the conflict between the syllogism's validity and its believability, rather than merely focusing on the syllogism's belief status.

There are also location-independent measures such as fixation duration and pupil size change. Although fixation duration is often used in conjunction with where eyes are fixated, change in pupil size is compared over a period of time as a result of change in the stimulus or task load. In the interest of keeping the literature review brief and relevant, below we review some seminal studies that have employed pupil size and fixation duration in logical and moral reasoning tasks.

Pupil dilation and constriction are often linked to cognitive control and are hypothesized to be modulated by the locus coeruleus-norepinephrine (LC-NE) system. In turn, pupillometry has been proposed as an indirect and non-invasive measure of LC-NE activity. The adaptive gain theory, proposed by Aston-Jones and Cohen, argues that the activity of the LC is indicative of exploratory or exploitative strategies in a task [3]. LC activity is often divided into the phasic and tonic modes. The phasic mode is reactive specifically to task-relevant stimuli and adapts quickly to changes. It responds to the utility of the task by increasing a phasic release of NE in synchrony with task-relevant events. The tonic release of NE, on the other hand, is often linked to attention disengagement. It responds to task-relevant and irrelevant stimuli invariably. When the LC releases NE in a tonic or more sustained manner, it facilitates strategy change by engaging in exploratory behaviors. Transitions between these two phases are controlled by the afferent input from the orbitofrontal and anterior cingulate cortex, areas supposedly linked to task outcomes and effort allocation.

LC activation has also been correlated with changes in pupil size, although the anatomical pathways are debated [25, 39, 69, 93, 121]. Exploratory strategies are inferred from an increase in baseline pupil size and relatively smaller dilation activity during the task, assumed to follow from a tonic release of NE [69]. Jepma and Nieuwenhuis (2010) investigated pupil size as participants made utility judgments on a four-arm bandit task in which the utility of each slot varied over the period of the experiment [93]. Pupil size was calculated continuously in the experiment. Participants selected a slot machine and received points. A trial consisted of one such choice. Averaged pupil size before participants made a new choice was considered a baseline to which pupil responses were compared. Exploratory choices, in which participants switched away from the previous choice of slot machine, were preceded by a larger baseline pupil diameter than when they stayed with the choice. Moreover, the baseline pupil size was also indicative of the changes in the utility of the strategy as participants switched from exploration to exploitation.

Exploration and exploitation strategies have also been studied in other cognitive tasks. Using pupillometry in tasks where an individual trial can span tens of seconds to a minute or more can become rife with holes in interpretation, particularly because any change in pupil size is difficult to map onto the underlying reasoning or thinking processes. Hayes and Petrov (2015) conducted one of the first, and still few, studies of pupillometry with an analogical reasoning task [89]. They measured pupil size as participants solved Raven's Advanced Progressive Matrices task (APM) [140]. The APM tests geometrical reasoning. The authors gave participants 14 individual problems. A problem in the APM consists of a grid of 8 figures in a 3x3 display with the cell in the last row and column empty. The task is to choose a correct figure for the last cell from 8 choices presented on the same screen below the grid. A task like this can be incredibly useful for detecting exploratory or exploitative strategies in reasoning. For instance, a potential pattern can be tested out by applying it to all figures of the grid. If it fits, then the alternative fitting that pattern can be chosen. This phase of solving is an exploitative strategy. On the other hand, when a pattern seems to be a misfit, a new strategy needs to be explored. The authors predicted that when participants switch strategies, the percent increase in pupil size will be more than when they stick to a current pattern. They used think-aloud verbal protocols to identify whether the participant was currently applying an exploratory or exploitative strategy. As expected, the pupil diameter was larger during exploration than exploitation.

In a broader perspective, pupil dilation has been used as an indicator of cognitive effort employed to resolve conflict in tasks. Increases in task demands like switching between tasks, inhibiting certain aspects of information, or monitoring and updating task-related information is followed by pupil dilation (for a review, see [177]). Recently, Purcell et al. (2023) tested the claim of the hybrid model of the DPT in the context of belief bias in syllogistic reasoning with eye-tracking (for a brief review of syllogistic reasoning tasks, see Section 2.2) [135]. The traditional two-process model, like the selective scrutiny model, argues that people must overcome prepotent belief priors to respond in a manner consistent with logical inference. Contrarily, hybrid models propose that both kinds of responses, supported by logical and belief-based inference, are available to the reasoner from the get-go. Hence, if these responses are in disagreement, then physiological markers such as pupil dilation should reflect this conflict. The authors tested their hypothesis under two sets of instructions. One set of participants was asked to judge the validity of the syllogisms while the other set was asked whether the conclusion was believable. The authors hypothesized that if people are inherently sensitive to the conflict between logic and intuitive beliefs, then it should result in dilated pupils regardless of which instructions were received. Participants solved a series of syllogisms with the believability of the conclusion and validity of the syllogism fully crossed. Pupil sizes were averaged over the period of reasoning. They reported that under both instruction sets, participants' pupils were dilated more in conflict than non-conflict syllogisms, lending support to the hybrid model of DPT.

Along with pupil size, average fixation duration also correlates with cognitive effort and information acquisition. When eyes are in a saccade, information uptake is limited. Therefore, visual information is mainly acquired when eyes fixate on an object (although information processing likely continues through successive saccades and fixations). The longer eyes fixate on an object, the more in depth they are assumed to be processed. Therefore, fixation duration has been widely employed as an index of subjective utility of objects in consumer research. A seminal study on brand choice by Pieters and Warlop (1999) gave participants six brands of products like rice, shampoo, canned soup, and salad dressing [133]. Under different time pressures and task motivation (whether or not the participants were offered their choice of brand as a reward), the duration of fixations on the ultimately chosen option was longer than on the non-chosen option.

Pärnamets et al. (2015) extended the use of fixation duration as indicators of processing information to directly modulate choice in a moral decision-making study [127]. Participants heard statements spoken over headphones like “Murder is sometimes justifiable,” with two alternatives like “sometimes justifiable” and “never justifiable” displayed on the screen for a short time. The period of this display was varied in two experiments. In the first experiment, if one of the alternatives was fixated on for longer than 750 ms and the other for at least 250 ms, the trial ended, and participants were prompted for a response. As expected, the alternative that was viewed for longer was likely to be chosen as the ultimate response. More interestingly, in the second experiment, the authors manipulated the trial duration such that trials were terminated if one of the randomly selected alternatives (the target) was fixated on for at least 750 ms and the other for 250 ms. Participants in this experiment chose the target alternative more than 58% of the time. The authors claim that they were successfully able to bias participants’ decisions on moral issues by essentially controlling information uptake from fixating on an alternative. While these results are intriguing, they may not necessarily capture how we reason in real time. Moral issues like deciding whether murder is justifiable brings a host of reasons, arguments and counterfactuals to our mind (e.g. e.g., Was it in self-defense? Did the victim of murder harm a loved one or was guilty of another crime?). These considerations often require time and extended deliberation to sort through, if we ever manage to make up our minds about such issues with certainty. In a short timescale, like in Pärnamets et al.’s study above, these contrasting motivations may be revealed in and affected by fixation times but the decision that is eventually reached may remain unstable. We will return to these interpretation themes later in the concluding chapter of the thesis, but for now, the above study suffices to serve as an example of how fixation duration (whether averaged or accumulated) has been used as an indicator of how deeply information is processed.

2.4 Tracking conflict in time

Think-aloud protocols offer a close approximation for tracking the reasons and arguments produced by a reasoner during task deliberation. Identifying the reasons that drive individuals to make decisions is a complex task. A substantial body of research suggests that post-decisional accounts are often unreliable. Participants do not necessarily provide an accurate reflection of how they reasoned and considered information during the decision-making process. Instead, their reports tend to align more with post hoc justifications or reconstructed narratives, rather than an accurate description of how the decision was made, as we have already mentioned above [83, 86, 113, 123].

This prompted researchers to investigate how we think as we engage in a task in terms of vocalizations and verbal data. Introspective methods for investigating cognitive processes have been central to early psychological research [29, 60, 170]. After behaviorism, verbal data once again became a key method in the 1970s for studying cognitive processes such as problem-solving and thinking, and this approach has been refined in subsequent decades. The vocalizations and verbalizations produced during task performance are assumed to reflect the underlying cognitive processes at work. These methods are rooted in information processing theory, which posits that cognitive processes unfold in a series of stages, with information being stored in memory units of varying capacities and accessibility. During reasoning, individuals move through a series of states held in short-term memory which is easily accessible but has a limited capacity, and any observable behaviors occurring at these stages are viewed as indicative of the cognitive processes involved in problem-solving [42].

Simon and Ericsson (1980, 1993) proposed a close correspondence between the intermediary steps in cognitive processing and the verbalizations produced. Specifically, the order in which information is transformed (i.e., the temporal order of cognitive states) mirrors the order of verbalizations [42, 43]. Participants are prompted to report their thoughts or explain their actions concurrently while completing a task. They suggested that each cognitive step corresponds to information that is attended to and/or held in short-term memory. Although asking participants to think aloud may extend the duration of the task, Ericsson and Simon (2003) argued that this delay reflects the time it takes to convert silent thoughts into spoken words, while maintaining the same sequence of cognitive states [44].

The verbal reports method has, thus, been proposed as a more accurate description of cognitive processes than post-decisional accounts, which are often retrospective justifications for the decisions made. After the decision is communicated, the cognitive processes that led to it may no longer be accessible in short-term memory, making it difficult to accurately recall the reasoning involved. In contrast, verbal reports made during the decision-making process have been argued to capture cognitive steps as they unfold in real time and in the order of occurrence.

But while thinking aloud during reasoning can provide a finer temporal resolution than trial-level summaries, the method has the potential to interfere with the reasoning itself. Indeed, when participants are asked to explain their actions while performing a task, it can alter their performance on the task [64]. For instance, in a study of consumer decision-making, it has been shown that verbalizing thoughts while deciding reduces confidence in the chosen alternative [179]. Additionally, verbal data analyses, such as componential analysis, demand meticulous attention to identifying the purpose and units of analysis, as well as establishing coding systems in advance [178].

Another issue with verbal data, particularly relevant to the present context, is the challenge of inferring conflict from verbalized or listed thoughts. For example, Zhao, Richie, and Bhatia (2022) explored how decisions informed by memory are processed [183]. Participants were presented with a problem designed to evoke no strong preference for an answer—essentially, a conflicting question. While reasoning to make a decision, participants wrote down the thoughts they were entertaining. The authors employed language models to identify thought clusters and used these clusters to infer whether a thought favored one of the available choices or suggested continued sampling of information due to similar weight given to both alternatives. However, the reliability of such data depends heavily on participants' ability to identify and articulate their thoughts as discrete units. When deliberating complex or contentious issues, individuals may struggle to pinpoint their reasons clearly, let alone verbalize them in a structured, point-by-point manner. At any given moment, individuals may not be fully aware of which specific reason they are considering; instead, they may only be cognizant of their readiness—or lack thereof—to make a choice. Although this discussion of Zhao et al.'s work focuses on their experimental method rather than the computational model, it underscores the limitations of verbalization or thought-listing methods in accurately tracking preferences and conflict as experienced by participants.

But beyond vocalizations or verbalizations, we are often also keenly aware of how we are navigating the context of the problem in our daily reasoning. Some decisions are obvious, while other necessitate deliberations. At times, we also recognize conflict as discrepancies between competing alternatives or their attributes. For instance, when faced with two equally appealing options, we may hesitate or oscillate mentally between them, reflecting an ongoing effort to reconcile conflicting preferences. In our Switch method, we track these momentarily shifting preferences in deliberations to see if they can be used as indicators of internalized conflict in reasoning.

In summary, this thesis had three-fold aims in investigating the dynamic nature of conflict: First, we sought to closely track this process while keeping task interference from the method to a minimum. Second, we also wanted to validate these metrics externally and internally by comparing the measurement to the existing operationalizations of conflict and the subjective ratings given by participants, respectively. Lastly, we wanted to establish the reliability of these measures by applying them to different types of reasoning problems. This was especially critical given that in Studies 2 and 3, we propose new methods for tracking conflict.

Chapter 3

Measures of Mouse Movements

We begin our investigation into measurement of conflict using an easily accessible and increasingly popular method of mouse-tracking. The assumption behind this method is that the curvature of response trajectories reflects real-time competition between alternatives, as motor plans are updated concurrently with the decision conflict. Over the past two decades since its first introduction, mouse-tracking has been extensively employed to study competition between competing choices [103, 116, 156, 162, 167]. As a response dynamics tool, it operates at the end of the trial. For tasks with short trials spanning only a few seconds, such as categorization tasks, mouse-tracking may still reflect competition between alternatives in the response trajectories because conflict or its resolution is likely experienced when the response is being made. In longer tasks, such as when deliberating over moral issues, the conflict is extended over time and takes longer to resolve. Simply put, the time spent communicating a response is likely to be proportionally greater in categorization tasks compared to deliberation tasks. This makes mouse-tracking metrics more effective in the former, as the measurement is taken closer to when the conflict occurs.

Perhaps, as a result, mouse trajectories in moral reasoning studies have been shown to be less optimal [82, 103]. The primary goal for these studies was to test whether preferences are updated in the order anticipated by the classical default-interventionist model. For instance, according to Greene's earlier versions of the model, personal sacrificial dilemmas with actions that more directly bring about the utilitarian consequence (e.g. smothering or pushing someone to their death) are processed preferentially by the fast and automatic System 1 [73, 80, 81]. People are expected to be deontological in these dilemmas as it is the associated response with System 1. However, if a person overrides this initial deontological reaction and opts for a utilitarian alternative, it is taken as evidence that the conflict between the two systems (System 1 and System 2) has been resolved in favor of the slower, more deliberate reasoning of System 2.

Koop (2013) and Gürçay and Baron (2017) hypothesized that atypical responses (such as choosing the utilitarian option in a personal dilemma) would be reflected in curvier trajectories demonstrating competition experienced by the reasoner in choosing an irregular alternative [82, 103]. However, their studies found no such patterns in the trajectories, leading them to challenge the validity of the default-interventionist model. Despite these conclusions, there is a possibility that the failure to detect conflict may not indicate a flaw in the model's mechanism but rather the limitations of mouse-tracking as a tool in extended decision-making tasks. The response trajectories may be engaged too late, potentially *after* the conflict is detected and resolved, diminishing their sensitivity as indicators of conflict. Furthermore, the trajectories themselves often exhibit substantial heterogeneity, pointing to sudden changes in mind not necessarily captured in smoother trajectories that are linked to continuous competition between options.

Previous studies rely heavily on the personal-impersonal distinction in moral dilemmas, a categorization that has been criticized before [13, 111]. In contrast, we moved away from this dichotomy in our experiments presented below. Instead, we conceptualized conflict in moral dilemmas based on group-level judgments: dilemmas that consistently elicited unanimous judgments were classified as low-conflict, while those that generated divided responses were categorized as high-conflict.

3.1 Experiment 1

In Experiment 1 of this study, conflict was defined based on cohort-level disagreements observed in dilemmas documented in established literature. Koenigs et al. (2007) categorized a personal dilemma as producing less conflict if the majority of participants generated the same judgment, resulting in uniform responses [100]. Similarly, Haidt et al. (1993, 2000) suggested that certain actions elicit a strong aversive reaction from most people, leading to a consensus where these actions are uniformly judged as highly inappropriate [85, 86]. We employed this conceptualization of conflict in moral reasoning to evaluate the effectiveness of three mouse-tracking metrics: AUC, MAD, and reversals along the X-axis. Specifically, we aimed to determine whether the experience of conflict while reasoning on moral dilemmas, as reflected by the consistency in judgments, could be captured through post-decisional response dynamics of the mouse trajectories.

3.1.1 Method

Participants

An email invitation was sent to students at the Indian Institute of Technology, Kanpur, encouraging participation in the experiment. A total of 70 participants took part in the experiment. After applying the qualifying criteria (explained in the subsection Preprocessing below), data from 67 participants (13 females; mean age = 22.38) were analyzed. All participants provided informed consent before starting the experiment and were compensated with Rs. 100/-. The experiment design and materials was approved by the Institute Ethics Committee (IEC).

Materials

To test the efficacy of the mouse-tracking method in capturing conflict in the reasoning process, we utilized a set of stimuli that have been extensively used in literature over the past two decades. This approach aimed to make our findings more comparable to existing research.

For this experiment, we selected 25 problems from literature. All 25 problems were presented in text format, featuring a third-person actor, X. The actor X had a choice between taking a proposed action or abstaining from it to solve the problem described in the text. X chooses to carry out the action in each problem and participants were asked to judge the appropriateness of X's choice.

Out of the 25 problems presented, 12 had actions which were presumed to be devoid of any moral connotations such as resolving scheduling conflicts, choosing between investment plans, and managing ingredients for food preparation. We refer to these problems as non-moral. The rest 13 problems were moral dilemmas of type low-conflict personal, high-conflict personal (henceforth, Low-C and High-C, respectively), impersonal, and harmless-offensive. Non moral, Low-C, High-C, and impersonal problems were taken from Koenigs et al. (2007) [100]. According to the authors, moral dilemmas are categorized as personal if the action within them is direct and evokes an emotional response. Contrarily, impersonal dilemmas involve indirect harm, typically through a series of mediating mechanisms. In this experiment, the actions in impersonal dilemmas did not involve physical proximity, such as bribing a judge, keeping money from someone's wallet, and withholding someone's property. In contrast, actions in personal dilemmas were more direct such as smothering, pushing, or performing a medical procedure. Although this categorization is debated, it was not central to our experiment. We were more interested in investigating mouse trajectories in supposedly conflicting and non-conflicting moral dilemmas. Koenigs et al. (2007) operationalized conflict as the disagreement or dissimilarity among participants' judgments [100]. That is, a dilemma was considered low on conflict if most people endorsed the same choice leading to a close to 100% agreement in their responses. Only Low-C dilemmas garnered such a cohesive response from participants in the original study with most of them choosing not to endorse the action in such dilemmas. On the other hand, High-C and impersonal moral dilemmas prompted variable endorsement rates in final judgments.

We also included problems commonly used in social-intuitionist theory literature. These problems are expected to elicit strong preference for rejecting the stated action within the dilemma, thereby allowing them to be categorized as low on conflict based on the current operationalization. These actions are harmless but highly offensive because they violate social standards such as having sex with a sibling when it cannot result in pregnancy, eating one's dead pet dog, cannibalism, and not honoring a dead parent's dying wish. These harmless-offensive problems were sourced from Haidt et al. (1993, 2000) [85, 86].

The text of each non moral problem and moral dilemma was divided into three paragraphs: the first paragraph described the broader context of the decision, the second detailed the action, its consequences, and that the actor in the problem, X, had endorsed the action. Third paragraph asked, "Is it appropriate for X to do that?" To maintain consistency, we modified the original dilemmas so that actor X always took the action. Below is an example of a harmless-offensive problem:

> "X works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body to science for research.
>
> One night X sees a body that is going to be discarded the next day. She knows the cadaver is thoroughly disinfected and hence is perfectly edible. X decides to take a piece of it home, cook it and eat it.
>
> Is it appropriate for X to do that?"

All problems used in this experiment can be found in Appendix.

Procedure

The experiment was conducted online. Each trial presented the problem text in white font in the middle of the screen against a gray background. Three boxes were displayed: two response boxes and one initiation box. The response boxes, colored blue, were placed in the top-right and top-left corners, labeled YES and NO, respectively. The position of the response boxes did not change throughout the experiment. The initiation box, colored green, was located at the bottom center of the screen with START written on it. To record a response, participants had to click the START box first, which enabled the response boxes, and then click their chosen response box. The response boxes were disabled until START was clicked. Once START was clicked, the initiation box turned gray and displayed CHOOSE to indicate that participants could record their choice now.

Participants completed two practice sessions before the main experiment. In the first practice session, participants saw a prompt (YES or NO) in the middle of the screen with the START, YES, and NO buttons in their usual places. Participants had to click on START and then click the response box corresponding to the prompt. No feedback was given. A picture of the screen layout during the practice session (with NO as the prompt) was also shown to participants for easy comprehension, along with the following instructions:

> "This experiment requires you to use the mouse precisely to record a response. To get used to it, we will start with a practice session. You will see three rectangles on the screen: A start, green rectangle on the bottom edge; 2 blue choice rectangles in top corners.
>
> This is what you need to do.
>
> 1. Read the prompt in the middle (in the picture NO).
> 2. Click START on the bottom edge.
> 3. Locate where NO is in the top two corners.
> 4. NO will always be in the top-left and YES in the top-right corner.
> 5. Then click on the blue rectangle which matches the prompt (here, NO). You need to click PRECISELY on the box.
>
> Click CONTINUE to start the practice."

In the second practice session, participants saw 16 YES-NO questions ('Does circle have corners?', 'Do you know how to swim' etc.) in the middle of the screen. The positions of the START, YES, and NO buttons remained unchanged. The following instructions were displayed on the screen before starting the second practice session:

> "Now that you understand the sequence of events, you will go through a practice session with simple questions instead of prompts.
>
> REMEMBER:
>
> 1. Read the question.
> 2. Click START.
> 3. Locate your answer in top corners.
> 4. Click the corner of your choice.
>
> Click CONTINUE when you are ready."

After completing the practice sessions, participants proceeded to the main experiment with these instructions:

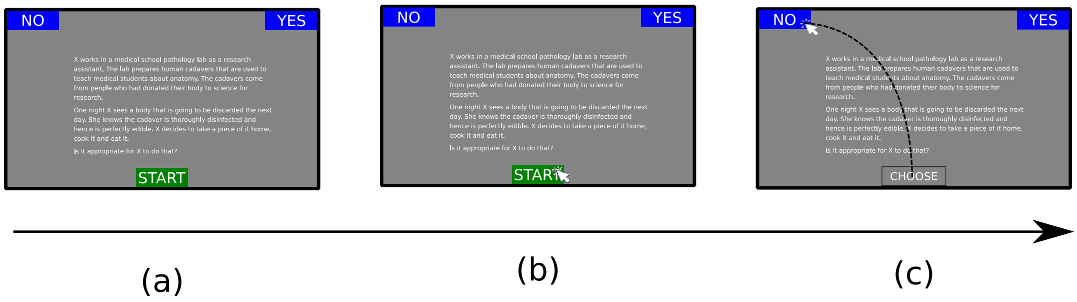
> "For the main experiment, you will be reading some stories with X as the main actor in them. At the end of the situation, you will be asked if X's action in the story was appropriate according to you. You have to click YES or NO to indicate your judgment.
>
> For instance, imagine X is taking a stroll in a park when she notices an ice-cream truck. X gets herself an ice-cream, despite knowing that if she catches a cold, she will have to quarantine for COVID-19 and miss work. If you think X's decision was appropriate, then you would click YES in one of the top corners. If not, then you would click NO.
>
> There may not be a right or wrong answer! Just answer according to what you think is appropriate. No pressure!
>
> Click "CONTINUE" when you are ready."

All trials in the main experiment were presented in random order. The experiment took 20 to 30 minutes to complete.

Preprocessing

The primary variables of interest in this experiment were the curvature metrics of cursor trajectories. Before calculating them, it was important to filter for the quality of the observed data. The trajectories are meaningful only if participants complete their responses by moving the mouse towards the response boxes immediately after clicking START. Although participants follow this way of recording a response on most trials, some response trajectories are uninterpretable due to their unpredictable path. We expected participants to read the problem text displayed at the centre of the screen before initiating the response by clicking START. In some cases, though, participants initiated the response too early, leading them to abandon moving the cursor toward the response boxes. Instead, the cursor paused in its tracks for too long, moved in ostensibly random patterns or hovered over the text presumably following the words the participants were reading at the time. Visually, such trajectories look tangled in the middle of the screen creating what we call 'messy middles'. We plotted all trajectories first and then filtered out those with unpredictable path like in Figure 3.2 (a). We also excluded three participants who selected the same alternative on all trials. In total, 6.2% of trials were removed from the final analysis.

For every trial, we recorded the mouse positions when the participant initiated a response on a trial by clicking on the START box until one of the response boxes was clicked (see Figure 3.1 (c)). Since the experiment was conducted online, the screen sizes could not be controlled across participants. Anticipating this, we coded the experiment in 'height' units using PsychoPy so that the stimuli positions were displayed and cursor positions were tracked relative to the screen dimensions [129]. Under this unit of distance, the limits of the


X works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body to science for research.

One night X sees a body that is going to be discarded the next day. She knows the cadaver is thoroughly disinfected and hence is perfectly edible. X decides to take a piece of it home, cook it and eat it.

Is it appropriate for X to do that?

FIGURE 3.1: Trial structure for mouse-tracking experiments. (a) The text of the problem was displayed in the centre of the screen with response and start boxes visible. (b) When ready, participant clicked the start box. (c) Participant located the answer in one of boxes in the top corners and clicked on it. The mouse trajectory mapping the start and end of the response is recorded.

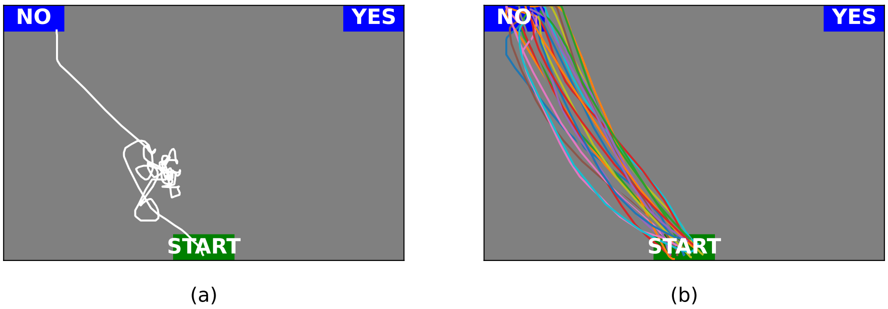


FIGURE 3.2: Illustrations of trials removed before analysis in mouse-tracking experiments. (a) Messy middles; (b) Participant recording the same response on all trials.

Y-coordinates are constant at (-0.5, 0.5), but the limits of the X-coordinates are dependent on the aspect ratio of the screen. For a standard 4:3 aspect ratio, the coordinate of the bottom-left and top-right corners are (-0.6667, -0.5) and (+0.6667, +0.5), respectively. For a widescreen 16:10 aspect ratio, the bottom-left is (-0.8, -0.5) and the top-right corner is (+0.8, +0.5). We scaled the X-coordinates of the tracked cursor positions to range between (-1, 1) by extracting the largest absolute X-coordinate recorded for a participant and then scaling all the tracked positions along the X axis to fit between (-1, 1). Y-coordinates were also scaled to range between (-1, 1).

The frame rate, or the rate at which the cursor position is sampled, also could not be controlled in this experiment. It varied among participants depending on the the frame rate of their systems. Additionally, the number of samples in a trajectory across trials is not the same as it would change, too, based on the time taken to finish recording the response. This often complicates aggregating and statistically comparing trajectories. Therefore, we time-normalized the trajectories using a package, written in programming language R, called 'Mousetrap' [97, 136]. Time-normalization interpolates trajectories by chunking them in equal number of positions (101 in this experiment following the standard protocol from Spivey et al. (2005) [157]) separated by constant time epochs. The package can also be used to calculate various mouse metrics, including those related to the curvature of the trajectory, the complexity (flips along the X or Y axes, known as x-flips and y-flips, respectively), and time-related information (total time without movement across the trial, time until the response was initiated, etc). We expected that if competing responses exert a pull on the cursor movements then it would show mainly in the curve of the trajectory of the cursor. Hence, we analyzed the area under the curve and the maximum absolute deviation of the trajectory (henceforth, AUC and MAD, respectively). AUC is the area bounded by the trajectory and the direct path joining the initiating click and the click signalling a response. AUC includes the area above the straight path joining the start and end of the

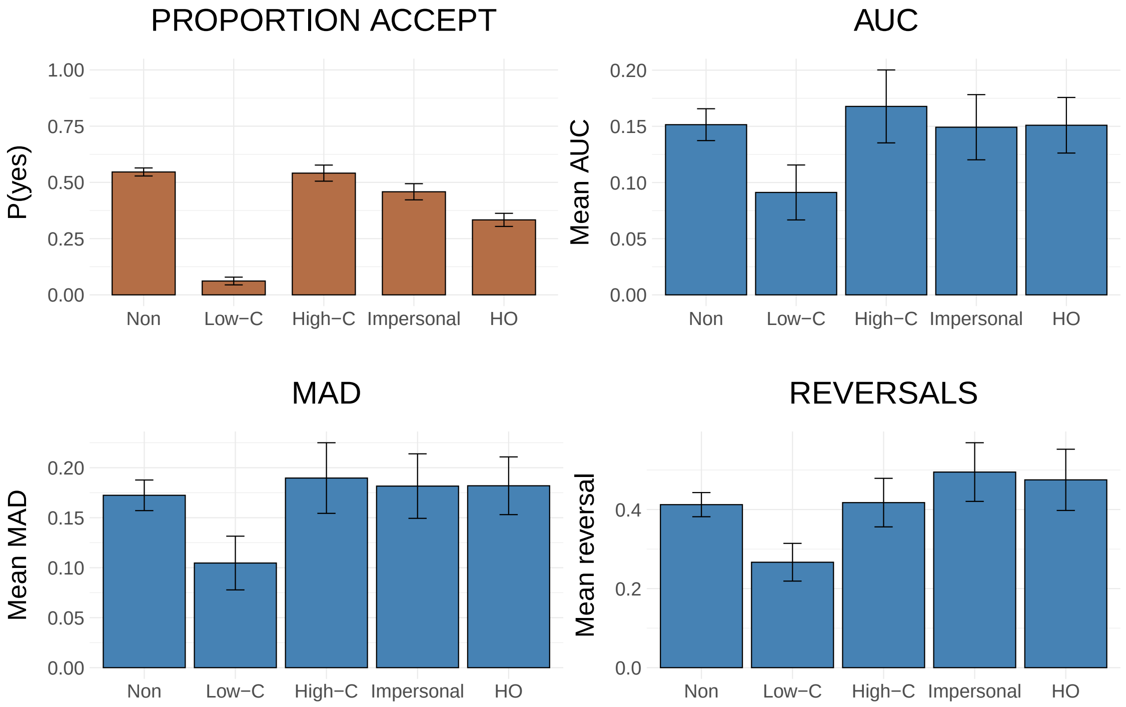


FIGURE 3.3: Mean acceptance rates, AUC, MAD, and reversals in all problem-types (Non and HO indicate non-moral and harmless-offensive trials, respectively) in Experiment 1 of Study 1. Error bars indicate standard error.

trajectory while subtracting the area under this straight path. This way, AUC captures the *pull* from the competing but unchosen alternative. On the other hand, MAD considers the maximum deviation regardless of it being under or over the straight path. Hence, although these two measures are expected to be correlated, AUC calculates only the competition from the competing alternative. In addition, changes of the mind could also be *sharper* such that participant moves in straight direction toward one option before switching over to the other. These movements can be calculated by looking at how many times the cursor crossed the Y axis or the vertical centre of the screen while making a response (see Figure 2.4 for a schematic of the mouse metrics used in Study 1). We calculated X-axis reversals for each trial, resulting in three mouse-tracking measures: AUC, MAD, and X-axis reversals. These measures were used to investigate whether conflict, operationalized as cohort-level disagreement, is reflected in post-decisional trajectories.

Table 3.1: Mean and 95% confidence intervals (CIs) for acceptance rates (P(accept)), AUC, MAD, and X-axis reversals in Experiment 1 of Study 1. CIs were obtained through 1000 bootstrap samples.

	P(accept)	**AUC**	**MAD**	**Reversals**
Non-moral	.55 [.51, .58]	0.1515 [0.13, 0.18]	0.1724 [0.14, 0.2]	0.4124 [0.35, 0.48]
Moral	.35 [.31, .38]	0.1406 [0.11, 0.17]	0.1658 [0.13, 0.19]	0.4181 [0.35, 0.5]
Low-C	.06 [.03, .1]	0.0912 [0.04, 0.14]	0.1047 [0.05, 0.16]	0.2667 [0.17, 0.36]
High-C	.54 [.47, .62]	0.1677 [0.11, 0.23]	0.1897 [0.12, 0.26]	0.4175 [0.3, 0.55]
Impersonal	.46 [.39, .54]	0.1492 [0.09, 0.21]	0.1816 [0.12, 0.25]	0.4948 [0.36, 0.66]
Harmless-offensive	.33 [.28, .39]	0.1510 [0.1, 0.2]	0.1820 [0.13, 0.24]	0.4751 [0.34, 0.65]

3.1.2 Results and discussion

In this experiment, participants assessed appropriateness of actions in moral and non-moral dilemmas. Actions in non-moral problems were deemed appropriate in approximately 55% trials. Actions in in moral dilemmas were judged appropriate less often (35%; see Table 3.1 for all summary statistics). In a generalized mixed-effects logistic regression model with participant as the random effect, this difference was significant ($\beta_{Non-moral} = 0.1901, SE = 0.08, z = 2.38, p = .02$; $\beta_{Moral} = -0.6434, SE = 0.10, z = 8.06, p < .001$; Random effect: $Var_{participant} = 0.0703$). However, the choices in both moral and non-moral problems in our study did not exhibit the level of cohesion that Koenigs et al. (2007) used to categorize a problem as low conflict. It is possible that the non-moral problems were also challenging for participants to resolve. Additionally, we did not collect subjective ratings of conflict for these problems, which could have provided a basis for comparing the mouse measures. Due to these limitations, we did not hypothesize any difference in mouse trajectories between moral and non-moral problems. We conducted exploratory linear mixed-effects (henceforth, LME) models predicting AUC, MAD, and reversals based on problem type (moral vs. non-moral). None of these measures showed significant differences in moral problems (see Table 3.2).

Our primary objective was to use Koenigs et al. (2007)'s categorization of moral dilemmas into low and high conflict subtypes to test efficacy of mouse measures. In Experiment 1, we used moral dilemmas from four conditions based on the type of action endorsed:

Table 3.2: Results of LME models comparing non-moral and moral problems on (a) AUC, (b) MAD, and (c) reversals in Experiment 1 of Study 1. Participants are treated as the random effect. Predictor is dummy coded with non-moral as the reference level.

(a) AUC ∼ Problem type

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.1517	0.02	136.72	8.44 ***	Participant	0.0085
Moral	-0.0109	0.02	1552.24	0.56	Residual	0.15

(b) MAD ∼ Problem type

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.1729	0.02	141.3	8.89 ***	Participant	0.0094
Moral	-0.0068	0.02	1552	0.32	Residual	0.1831

(c) Reversal ∼ Problem type

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.4161	0.05	110.5	8.85 ***	Participant	0.0783
Moral	0.0035	0.04	1551	0.08	Residual	0.7937

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Low-C personal, High-C personal, impersonal, and harmless-offensive. Low-C trials were labeled as such because they produced consistent responses, with nearly all participants rejecting the proposed action, as reported by Koenigs et al. (2007) [100]. In our experiment, the endorsement rate for Low-C actions was similarly low, at 6.15%, validating the original study's post hoc categorization (refer to the top-left panel of Figure 3.3 and the probabilities of accepting the action as appropriate in Table 3.1). However, the harmless-offensive dilemmas did not yield uniform judgments among participants, contrary to our expectation. We address this discrepancy later in this section. Next, we compared the Low-C acceptance rates to other conditions in moral dilemmas by conducting a generalized mixed-effects logistic regression model with participant as a random effect. The results, presented in Table 3.3 (a), show that all other conditions had significantly higher acceptance rates than Low-C dilemmas, confirming Koenigs et al. (2007)'s Low-C and High-C categorization.

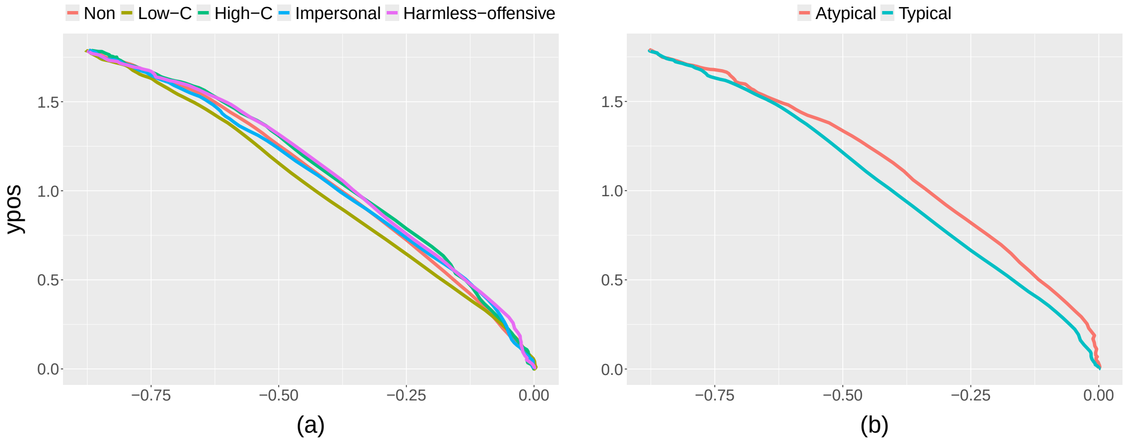


FIGURE 3.4: Aggregate trajectories of (a) all problem types and (b) typical and atypical trials in Experiment 1 of Study 1.

After confirming that our choice data replicated the original results of Koenigs et al. (2007), we modeled the mouse trajectories of these dilemmas after time-normalizing the data (see Preprosessing subsection for more details). We expected Low-C trials to have straighter trajectories reflecting lack of competition from the non-chosen alternative. Indeed, these dilemmas recorded lower AUC, MAD, and fewer reversals (see the descriptive statistics for these measures in Table 3.1). Despite this directional trend, the difference between Low-C and other dilemmas was not significant on AUC and MAD (Table 3.3 (b) and (c)). Although Low-C recorded significantly fewer reversals than impersonal and harmless-offensive dilemmas, the difference in reversals recorded in High-C and Low-C dilemmas was not significant.

According to the default-interventionist model of dual-process theory (DPT), System 2 engages after System 1 has processed information, but this engagement is not always guaranteed. In certain dilemmas, such as personal and harmless-offensive ones, a deontological response, supported by System 1, is typically dominant, while a utilitarian response is less common [85, 86, 100]. When a utilitarian response does occur, it is expected to take longer time due to the conflict arising from suppressing the strong deontological impulse, leading to extended response times. This prediction can be tested using mouse-tracking measures by comparing the trajectories of typical and atypical responses in low-conflict and harmless-offensive dilemmas. We selected these dilemmas because DPT predicts that most people will opt for the deontological alternative, judging the proposed actions as inappropriate and responding with a “NO.” Trials where participants responded with a “YES” were categorized as atypical.

Table 3.3: Results of models comparing Low-C dilemmas to other moral dilemmas in Experiment 1 of Study 1 for (a) final choice (generalized mixed-effects model), (b) AUC, (c) MAD, and (d) reversals along the X-axis (LME models). All models include participants as a random effect, with Low-C dilemmas as the reference level.

(a) Choice ∼ Dilemma type

Fixed effects				**Random effects**	
	Estimate	SE	z value		Variance
Intercept	-2.8724	0.31	9.15 ***	Participant	0.3437
High-C	3.0605	34	8.96 ***		
Impersonal	2.6928	0.34	7.92 ***		
Harmless-offensive	2.1229	0.33	6.4 ***		

(b) AUC ∼ Dilemma type

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.0917	0.03	352.74	2.97 **	Participant	0.0132
High-C	0.0755	0.04	774.78	1.94 .	Residual	0.1473
Impersonal	0.0582	0.04	775.04	1.49		
Harmless-offensive	0.0590	0.04	774.20	1.62		

(c) MAD ∼ Dilemma type

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.1054	0.04	345.15	3.03 ***	Participant	0.0173
High-C	0.0839	0.04	774.66	1.93 .	Residual	0.1843
Impersonal	0.0768	0.04	774.92	1.76 .		
Harmless-offensive	0.0762	0.04	774.1	1.88 .		

(d) Reversal ∼ Dilemma type

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.2683	0.08	338.9	3.47 ***	Participant	0.0861
High-C	0.1487	0.1	773.13	1.53	Residual	0.9137
Impersonal	0.2307	0.1	773.39	2.37 *		
Harmless-offensive	0.2052	0.1	772.56	2.27 *		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Out of 456 trials that were either the Low-C or harmless-offensive type, 99 recorded an atypical response. These atypical trials had larger mean AUC, MAD, and more reversals than typical trials (Table 3.4). We then examined whether this difference was statistically significant (see trajectories in Figure 3.4 (b)). We applied a Bonferroni correction for these post hoc tests, setting the Type I error rate (α) to 0.02, and conducted one-sided unequal variance Welch's t-tests on these three metrics. The difference was significant only in reversals (Table 3.4).

Lastly, we inspected the harmless-offensive dilemmas further. Actions in the harmless-offensive dilemmas were designed to break norms and provoke a strong sense of repulsion [85, 86]. Although we expected nearly all inappropriate judgments on these problems, participants deemed these actions inappropriate in two-thirds of the trials. By Koenigs et al. (2007)'s definition of conflict, who also categorized dilemmas post hoc, the harmless-offensive dilemmas in our experiment would not be considered low on conflict. In line with this, the mouse tracking measures as well did not reflect straight trajectories. The average AUC, MAD, and reversals for harmless-offensive actions were similar to those for non-moral problems (see Table 3.1). This discrepancy may be attributed to item-specific variability; for instance, Figure 3.5 shows that most atypical responses occurred in one particular harmless-offensive dilemma. However, the mouse metrics were larger on the first and the third problems. This implies that mouse-tracking metrics might be more reflective of participants' internal experience of conflict rather than their general response tendencies.

Overall, AUC and MAD were not effective in indexing conflict in our moral dilemmas. Koop (2013) previously tested the default-interventionist model of DPT using mouse-tracking and found similar results [103]. He used the stimulus from the same set as the

TABLE 3.4: Mean and standard deviations (M [SD]) for atypical and typical trials in Experiment 1 of Study 1, along with Welch two-sample t-tests comparing typical and atypical trials on AUC, MAD, and reversals.

		AUC		**MAD**		**Reversals**	
	N	M [SD]	t (df)	M [SD]	t (df)	M [SD]	t (df)
Atypical	99	0.1991 [0.44]	1.97 (136.17) .	0.2166 [0.48]	1.62 (140.39)	0.697 [1.67]	2.29 (109.71) *
Typical	357	0.1050 [0.36]		0.1301 [0.36]		0.2997 [0.77]	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

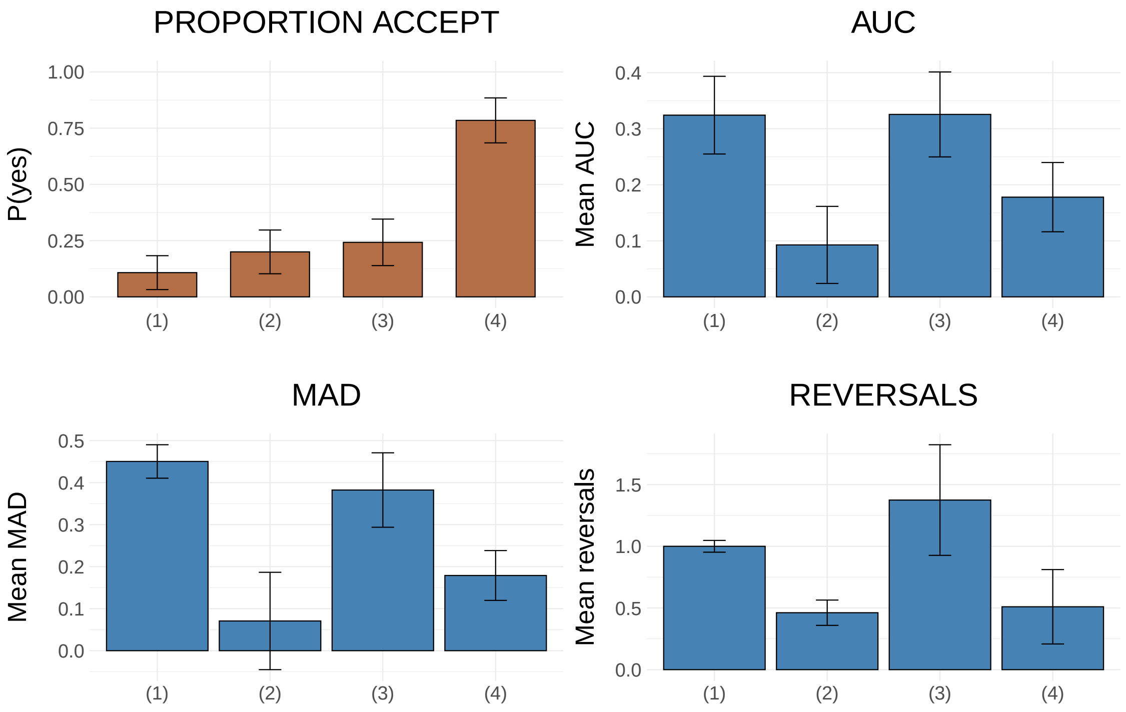


FIGURE 3.5: Item-wise dissection of harmless-offensive trials from Experiment 1 of Study 1. Each bar represents an item with a specific harmless-offensive action: (1) Cannibalism, (2) Incest, (3) Eating a dead pet dog, and (4) Breaking a promise made to a dying parent.

current experiment to assess how well mouse measures detected conflict, hypothesized as competition between System 1 and System 2 processes. Like in the current experiment, he replicated the choice patterns reported by Koenigs et al. (2007). Koop identified trials where participants instead provided a utilitarian response, predicting that these atypical responses should result in curvier trajectories due to the competitive pull from the typical response. However, the results did not support this hypothesis. He concluded that the conflict arising from concurrent activation of responses, as proposed by the default-interventionist model, did not adequately explain the data. However, it is possible that the assumption that mouse measures capture concurrent conflict may be inappropriate for reasoning problems that usually take longer to resolve than simple categorization tasks. Response dynamics are restricted to the concluding portion of reasoning, likely kicking in after evidence has been weighed in and perhaps resolved. These measures may not be sensitive to the conflict that preceded them.

Comparatively, reversals along X axis where participants were pulled to the other option enough to cross the vertical midline of the screen, were slightly more informative. For instance, reversals were lower in Low-C than impersonal dilemmas and in typical than atypical trials, as expected. However, more frequent reversals do not necessarily mean curvier trajectories. Participants may move the cursor straight to an option before switching their answer producing more angular and sharper trajectories. This type mouse movement should result in higher MAD but not necessarily AUC, given that mouse could move between options multiple times. Hence, reversals could be indicating abrupt changes of mind that do not necessarily inform the movements continually to produce smoother, curved movements. We revisit this issue in light of data from both Experiment 1 and 2 in the Discussion section of Study 1.

Although categorizing dilemmas as conflicting based on irregularity in judgments was largely replicated in our experiment, we could not infer about the conflict experienced by participants while reasoning. A group of participants may not choose the same alternative, but an individual may still be sure of what alternative she wants to choose. Consider, for example, a scenario where a group is tasked with choosing between tea and coffee. The group may be evenly divided, implying that the decision on beverage choice is potentially high-conflict. However, each individual member within the group may have encountered no internal conflict in making their personal beverage selection. Hence, internally validating mouse measures as indicators of conflict by comparing it to the actual subjective ratings is essential. We addressed this in Experiment 2.

3.2 Experiment 2

Research in the past twenty years has widely used sacrificial dilemmas, like some of the moral stimuli used in Experiment 1 above, to infer about human moral cognition. These dilemmas often give a choice between saving a larger number of people by sacrificing a few. Take the example of the switch version of the trolley problem. a runaway trolley threatens to kill five people working on a track. The decision-maker must choose whether to intervene by diverting the trolley to another track, where one person is working. This choice is framed as a conflict between the deontological principle of not killing and utilitarian calculations aimed at maximizing the number of lives saved. Participants are generally expected to overlook the realism of such scenarios and the legality of their potential actions.

Bauman et al. (2014) and subsequent researchers have critiqued these moral dilemmas for their artificiality, lack of realism, and limited external validity [16, 96, 145]. Such dilemmas are often decontextualized, intentionally omitting information about the actors and victims beyond the factors under investigation. While this decontextualization facilitates the comparison of specific moral principles at a more abstract level, it also undermines the external validity of the stimuli, potentially leading to poor predictability of choices in real-world dilemmas. To counter this, Schein (2020) has proposed the use of more realistic and contextualized moral problems to better test moral theories [145].

Hence, for Experiment 2 in this study, we incorporated problems that reflect moral conflicts typical of everyday interactions. These problems did not just focus on endorsing typically immoral actions based on their closeness to the outcome, as suggested by the personal-impersonal distinction. Instead, the actions in the problems used in Experiment 2 were more contextualized to convey the pull of contrasting motivations on the choice. While doing this, we also sought to operationalize conflict as cohort-level disagreement in judgments. To achieve that, moral problems for this experiment were chosen from a popular social networking site. Users of these communities share personal stories involving moral or ethical dilemmas they have encountered in their lives, while other community members judge the appropriateness of the actions. We particularly selected the problems that had split the opinion of the online community.

3.2.1 Method

Participants

For Experiment 2, we recruited participants by advertising through emails to the students of Indian Institute of Technology, Kanpur. Total number of participants who completed the experiment was 69. After applying the exclusion criteria which is explained in the subsection Preprocessing below, final analyses were performed on data from the remaining 65 participants (14 females; mean age = 22.35). Participants were compensated with Rs. 100/- for their time. The experiment was approved by the IEC.

Materials

For this experiment, we prepared a stimulus set comprising 18 problems, categorized as either non-moral or moral. We included 10 non-moral problems from Koenigs et al. (2007) [100]. Like in Experiment 1 of this study, we wanted to test the efficacy of the mouse measures in detecting conflict in moral dilemmas operationalized at the cohort-level. For the current experiment, we selected dilemmas from the social networking site, Reddit. Reddit is organized into subreddits, each of which focuses on a particular topic, theme, or interest. Subreddits are identified by the prefix “r/”, followed by the name of the community. To choose dilemmas for Experiment 2, we shortlisted two subreddits called r/moraldilemmas and r/AmItheAsshole. These subreddits are popular online communities where individuals seek judgments on actions or decisions in various situations. Users post personal anecdotes or dilemmas, asking fellow Redditors to evaluate whether they behaved appropriately or if they were at fault. The community votes and comments, often providing detailed moral assessments with their responses. Although these communities have been criticized for producing uniform judgments across a variety of morally charged situations, there are some dilemmas that users find difficult to judge and convey them to be so.

We focused on finding such dilemmas by filtering posts tagged as controversial by the Reddit algorithm, indicating a divided opinion among users. Content on these subreddits can span various topics, such as politics, religion, social issues, or personal anecdotes. We specifically filtered for posts that had gained significant attention and excluded those strictly involving current political scenarios, focusing instead on interpersonal conflicts. The selected dilemmas included scenarios such as a father deciding to cut ties with his drug-addict son to safeguard his other child, a family giving up their adopted child with mental health issues after realizing they could not provide adequate care, a brother refusing to take responsibility for his irresponsible younger sister etc.

These dilemmas were rewritten so that they were not too long compared to non-moral problems (word count ranges of non-moral problems = [53, 124] and moral dilemmas = [79, 107]). Like Experiment 1, all dilemmas had a third-person actor, X, who always took the action stated within the dilemma. Participants were asked to judge if the said action was appropriate. All moral and non-moral problems were broken down in three paragraphs, like in Experiment 1. First paragraph described the problem while the second paragraph described the alternatives their consequences and the actor’s decision to carry out the action. Participants were asked if it is appropriate for X to endorse the action. An example of a moral dilemma from Experiment 2 is below:

> X is at the top of his chemistry class. She has been consistently scoring good grades. Once she just could not concentrate for an upcoming class test. She still appeared for it but could not answer most of the questions. She got frustrated and crumpled her answer sheet, stuffed it in the trash and left the hall.

Table 3.5: Mean and 95% confidence intervals (CIs) for acceptance rates (P(accept)), response times (RT), difficulty ratings (rating), AUC, MAD, and X-axis reversals in Experiment 2 of Study 1. CIs were obtained through 1000 bootstrap resamples.

	P(accept)	**RT**	**Rating**	**AUC**	**MAD**	**Reversals**
Non-moral	0.5988 [0.56, 0.64]	40.6677 [38.36, 43.27]	2.4863 [2.4, 2.57]	0.3038 [0.26, 0.35]	0.3381 [0.29, 0.39]	0.7128 [0.63, 0.8]
Moral	0.4783 [0.43, 0.52]	37.8876 [33.41, 43.73]	2.7505 [2.66, 2.85]	0.3225 [0.27, 0.38]	0.3772 [0.32, 0.43]	0.7788 [0.67, 0.9]

> Next class, the professor asks her to stay behind. He apologizes for losing her answer sheet and averages her previous exam scores. X feels bad but decides not to correct him.
>
> Is it appropriate for X to do that

All the dilemmas are detailed in the Appendix.

Procedure

The procedure for this experiment was identical to Experiment 1 above. It was conducted online and advertised via email to the students and staff of the Indian Institute of Technology, Kanpur. Similar to Experiment 1, Experiment 2 included two practice sessions to familiarize participants with using the mouse to record a response, followed by the main experiment session. The prompts in the first and second practice sessions remained unchanged. In the main experiment, the trial structure remained the same, except for an addition of rating screen at the end of each trial. We also collected response time data for each trial. Response times were measured from the moment the problem was presented until the response was initiated by clicking on START. After every trial, participants were asked 'How difficult was the last question to answer?' with a 5-point rating scale displayed on the screen ("Not at all", "A little", "Neutral", "Somewhat", "Very much").

The body of text in the experiment was presented centrally in white font on a gray screen background. The START button was displayed at the center of the bottom edge of the screen within a blue box. The response could be initiated by clicking on this box, which then disabled the START box and activated the green response boxes. The alternative YES was always in the top-left corner and NO in the top-right corners. Position of the

Table 3.6: Mixed-effects models comparing non-moral and moral problems in Experiment 2 on (a) alternative chosen (generalized mixed-effects model), (b) response times, and (c) difficulty ratings (LME models). Participants are treated as the random effect. Predictor is dummy coded with non-moral as the reference level.

(a) Choice $\sim$ problem type

Fixed effects				**Random effects**	
	Estimate	SE	z value		Variance
Intercept	0.4012	0.08	4.98 ***	Participant	0.0085
Moral	-0.4885	0.12	4.135 ***		

(b) Response time $\sim$ problem type

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	41.281	3.26	77.518	12.652 ***	Participant	534.1
Moral	-2.844	2.44	1114.15	1.164	Residual	1750.5

(c) Rating $\sim$ problem type

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	2.4880	0.08	80.48	31.68 ***	Participant	0.3262
Moral	0.2670	0.05	1119	4.96 ***	Residual	0.8490

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

response boxes did not change throughout the experiment. The instructions were not altered. All trials in the main experiment were presented in random order which took participants about 20 to 25 minutes to complete. Participants were compensated with Rs. 100/- for their participation.

Preprocessing

Before analyzing the data, we excluded trials that produced 'messy middles' (see Figure 3.2) and participants who gave the same response to all trials (total excluded participants = 2), as in Experiment 1. As an extra precaution against initiating responses too early, we also excluded data from trials where participants took less than 10 seconds to initiate the response by clicking START. In total 16.15% trials were excluded from final analysis. The experiment was coded in 'norm' units of PsychoPy, version 2021.1.4, and hence, X

Table 3.7: Results of LME models with predictors problem type and ratings modeling (a) AUC, (b) MAD, and (c) reversals in Experiment 2 of Study 1. All models include participants as a random effect, with non-moral problems as the reference level.

(a) AUC ∼ dilemma type

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.19629	0.07	670.66	2.91 *	Participant	0.0259
Moral	0.1373	0.1	1157.36	1.44	Residual	0.3789
Rating	0.0432	0.02	1080.01	1.8 .		
Moral:Rating	-0.0476	0.03	1164.36	1.42		

(b) MAD ∼ dilemma type

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.2099	0.07	633.31	3.06 **	Participant	0.0412
Moral	0.0593	0.09	1148.65	0.63	Residual	0.3614
Rating	0.0518	0.02	1144.82	2.16 *		
Moral:Rating	-0.0127	0.03	1154.85	0.39		

(c) Reversals ∼ dilemma type

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.5008	0.1308	532.36	3.83 *	Participant	0.231
Moral	0.1140	0.17	1138.96	0.68	Residual	1.173
Rating	0.0885	0.04	1178.96	2.02 *		
Moral:Rating	-0.0278	0.06	1143.91	0.47		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

and Y coordinates were within range (-1, 1) by design [129]. 'Norm' units distort the objects on the screen across different monitor sizes. However, for all horizontal screens like laptop screens and monitors, the relative size of the stimuli on the screen is the same. Only vertical screens such as mobile phones render the stimuli improperly. Hence, it was important that people use wide-screens for appropriate rendering of the stimuli. We ensured this requirement by asking participants to press SPACE key on their keyboards while agreeing to the consent form, assuming most people do not typically use keyboards while using phone.

Raw data were first time-normalized into 101 chunks of constant time intervals and AUC, MAD, and reversals for each trajectory were calculated using the 'Mousetrap' package [98].

3.2.2 Results and discussion

Stimuli used in Experiment 2 were problems of either non-moral kind, like in Experiment 1, or moral interpersonal conflicts that had caused a split in judgments on an online community. Since non-moral problems in our stimuli also do not usually produce cohesive responses (see the Results section in Experiment 1 of Study 1 and Koenigs et al. (2007) [100]), we did not predict that there will be systematic differences in the choices or response times. Overall, there were fewer trials where the action in a moral problem was deemed as appropriate. Moral problem were also rated slightly higher than non-moral problems on the difficulty scale. However, response times were shorter on moral problems than non-moral problems (for the descriptive statistics, see Table 3.5). We modeled the choice, response times and ratings data in mixed-effects models displayed in Table 3.6. There was a significant difference in the choice data and ratings, but response times were not significantly different on moral and non-moral problems.

We constructed LME models with problem type (moral or non-moral) and difficulty ratings as predictors to model the mouse metrics. We did not specifically hypothesize the main effect of problem type on the dependent variables, because the choice data could not be used to categorize either of these as low-conflict problem type. Further, the response times, which are also generally used as indicators of conflict in reasoning, were comparable in both conditions. However, we could use the ratings given by participants as an indicator of experienced conflict while reasoning. Indeed, ratings alone were better predictors of MAD and reversals, with AUC showing the expected trend but not reaching the significance criterion, as shown in Table 3.7. In other words, higher difficulty ratings were correlated with larger MAD and more frequent reversals but did not have the same effect on AUCs.

3.3 Discussion

In Study 1, we explored the potential of using mouse-tracking to monitor conflict during moral reasoning. Our findings offer mixed support for this methodology. Conflict was operationalized differently across both experiments, but in each case, it was defined by the level of similarity in group judgments. Experiment 1 employed well-established moral dilemmas commonly used in the literature, while Experiment 2 used novel stimuli which, although not standardized, were more realistic and had produced dissimilar judgments on Reddit. The categorization based on choice data aligned with our expectations in both experiments. However, despite these consistent patterns in trial-level summaries, mouse-tracking measures did not reliably capture conflict across both experiments.

Such lack of consistency in effectiveness of mouse-tracking in indexing conflict may be partly due to it being used too late in the reasoning process. This criticism goes beyond the mouse measures and is applicable to some popular tools of response dynamics. To use these tools as indicators of conflict, one must assume that the conflict is actively experienced when a response is being made. However, reasoning problems that span more than just a few seconds seconds, such as problems in this study, this assumption may not necessarily hold true. Participants may have already experienced conflict while reading the text of the problem, long before recording the response. For response dynamics to be meaningful in such tasks, one must further assume that motor plans are altered by the experience of conflict in a detectable way, suggesting that conflict impacts subsequent cognitive activity. Even then, participants are not required to respond immediately after experiencing conflict, which diminishes the ability of response dynamics tools to capture the competition among alternatives that occurs earlier in the reasoning process. In any case, response dynamics cannot explicate the temporal dynamics of tasks that take longer to resolve.

Another reason why the mouse metrics did not match with the operationalization of conflict in Experiment 1 of this study could be the absence of any indicators of conflict given by the participants themselves. In Experiment 1, although choice data lined up with the original data reported by Koenigs et al. (2007), we still do not know if the problems which supposedly produced greater levels of conflict were really conflicting. This piece of information was useful in Experiment 2. Regardless of whether the dilemma was moral or non-moral, participants' trajectories were curved toward the unchosen option more especially when the problem was rated as highest on the difficulty rating (see Figure 3.6). This suggests that subjective experience of reasoning may be more informative in modeling reasoning processes than supposed by the extant models of reasoning processes.

When introducing mouse-tracking as a tool to track competition between responses, Spivey et al. (2005) compared two hypotheses about how motor movements can be used to infer underlying cognition [157]. According to the continuity of mind hypothesis, motor movements are continually and directly updated along with the changing competition between alternatives [155]. In this view, readouts of psychological processes are reflected

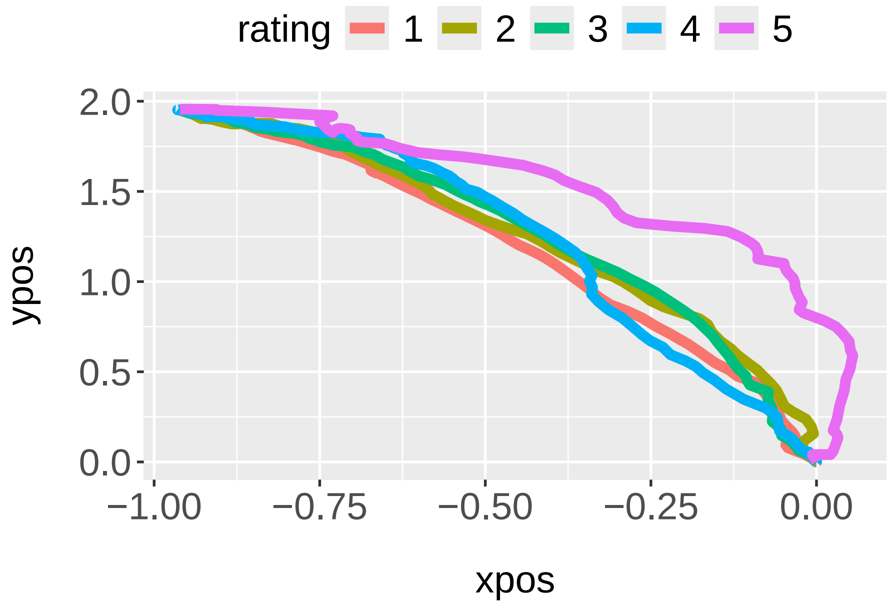


FIGURE 3.6: Normalized trajectories in Experiment 2 of Study 1 by difficulty ratings.

in constantly changing motor plans. Therefore, by continuously tracking movements while a response is being deliberated—which can be made readily available by controlling the way a response is given as part of the task demand—we can infer the underlying representations that guide the response. Such continually updated motor plans are expected to result in trajectories that show a degree of curvature, depending on how the competition between two alternatives is being resolved. The curvier the trajectories toward the unchosen option, the greater the pull from that alternative. In other words, trajectories can simply be indexed by curvature metrics such as AUC and MAD, and together they should form a unimodal probability distribution of these indices.

However, not all trajectories are of the same type. Often, participants move directly toward one choice before switching to the other alternative, resulting in more angular response trajectories. See the graph on the right in Figure 3.7. We combined data from both experiments to identify trajectory types. Using the ‘Mousetrap’ package, we matched trajectories in our data to five distinct prototypes based on a distance metric. Each panel in this graph shows a distinct type of mouse movement. For the ease of interpretation, the final choice made in all trials is mapped onto the top-left corner. The first three panels show trajectories that gradually curve toward the competing alternative in the top-right corner. However, these were not the only trajectory types in our dataset. Trajectories under the ‘Switch-1’ response first moved straight to the competing alternative. Only then did the response trajectories reverse, ultimately moving to the corner with their final answer. In

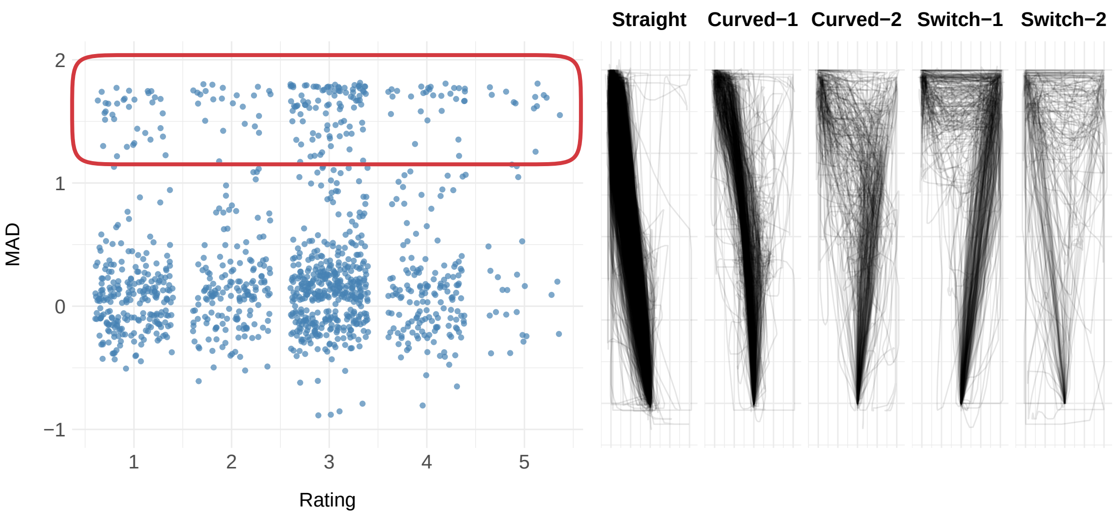


Figure 3.7: These graphs illustrate the variability in trajectories across participants. The left graph plots the MAD of all trajectories in Experiment 2 against participant ratings, with highlighted sections indicating the presence of outliers in the distribution. The paneled graph on the right categorizes different trajectory types in both experiments using prototype matching.

'Switch-2', participants switched answers twice while responding: first, they moved to the option they would ultimately choose, then switched to the competing alternative before switching back. 'Switch' trajectories like these can also be detected in the distributions of curvature metrics. Reversals can potentially identify these angular trajectories, as frequent switches increase the number of reversals along the X-axis. However, the number of switches detected in reversals may also be conflated with curvier trajectories that cross the midline of the screen before moving to the corner representing the choice. A distribution of MAD (maximum absolute deviation) of trajectories can more clearly demonstrate the presence of these angular trajectories. They produce extreme deviations in MAD, resulting in two peaks in the distribution, as seen in the graph on the left in Figure 3.7.

The hypothesis of continuous updation of motor plans cannot accommodate such heterogeneity in trajectory types. People change their mind only after reaching an alternative. Such abrupt shifts in preference are inferred from the switch in the trajectories, but not the curvatures themselves. Since these trajectories are not curved, they provide limited insight into how the competition between alternatives was resolved. To accommodate these data, the discrete model argues that instead of continually updating motor plans, they are intermittently altered based on the underlying cognitive processes. This results in often observed motor movements that are largely ballistic followed by more controlled adjustments. Of course, heterogeneity in trajectory types does not render the mouse-tracking methodology less useful. Although only a fraction of trajectories are angular like observed in our experiments, these are not limited to tasks like ours which takes longer than the typical categorization or perceptual tasks. They have been reported in perceptual and cognitive tasks as well [57, 62, 180].

Mouse-tracking has been applied to a wide range of tasks, including attention [63, 116], memory [1, 104], social cognition [57, 62], human-computer interaction [40], and decision-making [82, 103, 162, 167]. The decision-making tasks using this methodology range from brief decisions about gambles to extended tasks, such as moral reasoning, that span over an extended period of time. We argue that the advantage of mouse-tracking as a real-time cognition tracker diminishes in the longer tasks. For example, to determine whether System 1 and System 2 processes operate in series or in parallel, a measurement tool must closely track these processes in time. Only then can hypotheses about the sequence of responses be accurately tested. Koop (2013) and Gürçay and Baron (2015) used methods similar to ours to investigate the dual-process account in moral cognition. They expected that atypical responses, such as choosing the utilitarian alternative in personal dilemmas, would reflect internal conflict in mouse trajectories, either through more extreme curvatures or increased switches between choices, similar to the 'Switch' cases in Figure 3.7 [82, 103]. However, both studies failed to demonstrate these effects. While it is possible that the default-interventionist model of DPT does not fully capture the complexity of human moral cognition, we believe that mouse-tracking may have limited utility in testing hypotheses about cognitive processes. Our data in Study 1 also fails to provide a reliable test for process models of moral cognition.

3.4 Study 1 in review

Mouse-tracking has gained popularity as a tool for inferring internal conflict between alternatives in categorization and reasoning tasks. However, in reasoning tasks—especially moral ones—that unfold over several seconds, the response trajectories of the mouse do not necessarily reflect the conflicting moral decisions being made. In this study, we argue that while mouse-tracking can be useful for shorter tasks, it is inadequate for capturing conflict detection and resolution over extended periods. This limitation arises because the method is inherently post hoc, typically applied only at the end of the reasoning process. As a result, it may fail to capture the full phenomenological experience of reasoning. Our data demonstrate that conflict during deliberation is not always reflected in curvier trajectories; instead, we observe more abrupt shifts in preferences. These shifts can occur throughout the deliberation process and are likely integral to it. As we weigh alternatives, reconsider options, and experience conflict as a tug-of-war between competing motivations, these shifts may manifest as momentary changes in preference. In the next chapter, we track these shifts in real-time as they unfold during reasoning.

Chapter 4

Vacillations As Indicators of Conflict

When we struggle to commit to a choice, we often experience shifts in our preferences. These vacillations are a frequent aspect of our reasoning process. Although theories of reasoning frequently predict the order in which choice updates in the pre-decisional period, direct tests of these predictions are rare. Hence, measuring such mental vacillations within choice trials may prove to be consequential for a realistic assessment of cognitive conflict and the differentiation of plausible theories of the reasoning process. In Study 1, we demonstrated that response dynamics such as mouse-tracking methods are insufficiently granular to detect moment-to-moment variations in preferences. Methods like thinking-aloud paradigm may offer more insights into how we reason but they are overly intrusive to adequately detect changes of mind. Here, we introduce a novel method for measuring vacillations that minimizes interceptions with the reasoning task and has more temporal resolution than response dynamics methods. Our Switch paradigm captures participants' instantaneous preferences during reasoning unobtrusively.

In the experiments described below, participants were instructed to report the direction their thoughts were leaning while deliberating on a problem with two possible choice alternatives. They were encouraged to express their preferences whenever they felt them building and as frequently as desired. In a trial, the decision process was divided into reasoning and committing to a final decision. The problem was presented at the center of the screen, with each of the two alternatives identified by either the right or left arrow keys. While reasoning, participants pressed the arrow keys whenever they felt they were strongly considering the corresponding choice. This allowed for multiple presses of the same key

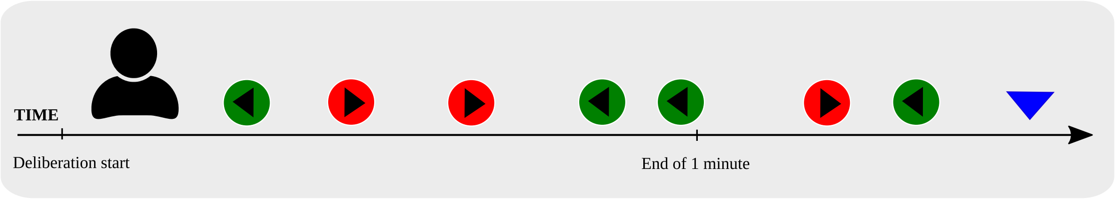


Figure 4.1: From Shivnekar and Srivastava (2023) [150]. The figure depicts representative key-presses during the deliberation phase of a trial. After reading the problem, participants pressed these keys whenever they wished to record an interim preference. Green and red symbols are the LEFT and RIGHT key presses, respectively. Blue triangle indicates participant ending the deliberation to record the final judgment which they could do so only after 1 minute was over.

and the freedom to press them in any order. Participants were explicitly informed that the key presses they made during reasoning did not necessarily have to align with their final decision. This approach was implemented to mitigate the potential bias associated with feeling compelled to reason in line with the normative choice.

The primary dependent variable in our paradigm was the switches in preference. When a participant pressed the right key after pressing the left key, or vice versa, we inferred that during that period, the participant changed their mind. We hypothesized that participants would exhibit more switches while deliberating over conflicting problems compared to those with a straightforward choice (see Methods and Figure 4.2 for detailed description of the paradigm).

Our objective in employing this paradigm was to establish both the internal and external validity for our measurement tool as a gauge of cognitive conflict. To establish internal validity, we aimed to demonstrate that people vacillate more when they subjectively feel conflicted during a choice. For external validity, we wanted to see how vacillations map onto previously proposed measurements of conflict in the literature. Experiment 1 and 2 of the current study tested two operationalizations of conflict in moral reasoning proposed by Koenigs et al. (2007) and Bago and De Neys (2019), respectively [6, 100]. Next, we wanted to test the generalizability of the paradigm when the rules of reasoning are familiar. We used categorical syllogisms in Experiment 3 in place of moral dilemmas, while keeping other details unchanged. Our results demonstrate that directly measuring vacillations in reasoning can help differentiate theories of both moral and logical decision-making.

4.1 Experiment 1

In Experiment 1, we were interested in testing the efficacy of our new method of tracking vacillations and its consistency against the operationalization of conflict proposed by Koenigs et al. (2007)[100]. We designed a paradigm in which participants could report interim preferences as they deliberated on moral problems. The number of shifts in the reported preferences was used as an indicator of conflict.

We tested this measure in trials where we expected fewer (Low-C personal dilemmas) and more frequent vacillations (High-C personal dilemmas). Koenigs et al. (2007) had categorized dilemmas post-hoc as low or high on conflict based on the observed consensus among participants' choices. These provided a robust test to validate our new method of measuring vacillations as indicators of internalized conflict. Moreover, the same stimuli set had yielded similar judgments with the sample in Experiment 1 of Study 1, establishing reliability of the categorization and allowing for an easy comparison between our method and the existing literature. Harmless-offensive trials have been reported to produce strong and quick responses that reject the proposed action in the dilemma. These kinds of problems are also resistant to argumentation, often leading people to simply rationalize their judgment [85, 86]. Despite this, in Experiment 1 of Study 1, the harmless-offensive trials did not yield uniform responses in our sample (see treatment of this issue in the Discussion section of Study 1). Therefore, these trials were excluded from the current experiment to ensure focus on more consistent dilemmas.

The stimuli set was constructed based on theoretically relevant variables presumed to influence choices. With these stimuli, we could compare the predictions from the proposed models of moral cognition such as the default-interventionist and hybrid models of DPT. The default-interventionist model posits that preferences are updated sequentially if the dominant response from System 1 can be overridden by expending resources [52, 75]. The hybrid model, on the other hand, suggests that System 1 supports intutions about both deontological and utilitarian alternatives [6]. Thus, preference updating can potentially occur when the strength of the intuitions are comparable.

4.1.1 Method

Participants

Twenty-five participants were recruited for this experiment (13 females; mean age = 25.3 years). The sample size was derived from a pilot study, where the effect size of the difference between High-C personal and Low-C personal dilemmas was 0.67 (Cohen's d; for details about the dilemma types, see below), achieving a power of 0.8 with a significance level of $\alpha = 0.05$. The sample size was calculated with G*Power software [54]. The experiment design was approved by the IEC. Participants were compensated with Rs. 100/-.

Materials

We selected 16 problems from Koenigs et al. (2007)'s paper which were divided in four conditions: non-moral, low-conflict personal (Low-C), and high-conflict personal (High-C), and impersonal [100]. The authors had provided mean emotionality ratings for all moral dilemmas (Low-C, High-C, and impersonal) in their stimulus set. We ranked them based on their rating and selected four moral problems for each condition, taking into consideration anticipated familiarity of participants. No specific criteria were needed for selecting non-moral problems, as their contexts were fairly neutral, like scheduling appointments, choosing between routes (two problems featured this action), and deciding to purchase product A instead of B. The non-moral scenarios were expected not to invoke any moral principles.

Stimuli for this experiment resembled the problems from Experiment 1 of Study 1 as both of them were taken from the same set. Participants were tasked with making a two-alternative forced choice between performing an action and refraining from it. The stimuli contained the context of the problem in which the alternatives (action and inaction) were made clear along with their consequences. Specifically, actions in the Low-C and High-C conditions involved saving a larger group at the expense of injuring or killing a smaller number of people. These actions were considered personal, as they directly caused harm to individuals or groups such as breaking someone's arm, smothering a baby etc. (for a more detailed discussion of 'personal' actions in this context, refer to Greene et al. (2001, 2004) [81, 80] or Chapter 2 of this thesis). Six out of eight of these trials involved scenarios where death was a possible outcome.

To reiterate, the authors of the original paper to use these dilemmas had classified a personal problem as “low-conflict” post-hoc when almost all participants in their study disagreed with endorsing the utilitarian action and “high-conflict” when varying degrees of disagreement were observed in their sample pool. Impersonal dilemmas did not include any problems in which the victim died directly from carrying out the action which benefited the actor’s welfare, e.g., stealing cash from a wallet on the ground, bribing to win a case, etc.

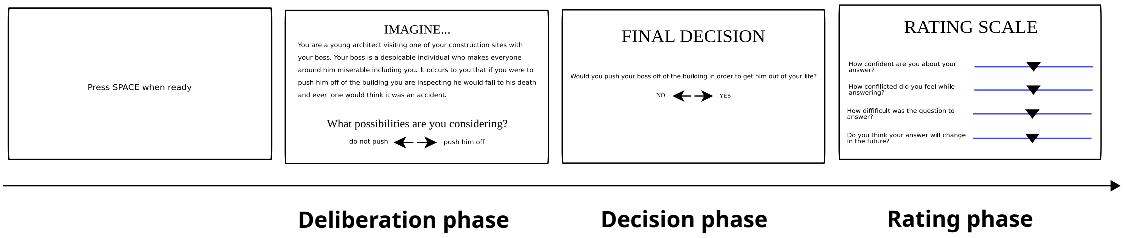
For Experiment 1, we wanted participants to deliberate actively. We retained the original moral dilemmas, modifying only elements for relevance such as changing currency units and names of the city. To make the dilemmas more immersive, the actor in the dilemmas was always the reader unlike in stimuli from Experiment 1 of Study 1 where the actor was a third-person X. In each problem (moral or non-moral), participants were asked if they would take the action proposed in the dilemma.

All 16 dilemmas can be found in the Appendix.

Procedure

All trials were self-paced and consisted of three phases: deliberation, decision, and rating (see Figure 4.2 for trial structure). During the deliberation phase, participants read a problem which clearly outlined their choice options for that trial. The question at the end of the original dilemmas was omitted. Instead, participants were presented with the prompt, “What possibilities are you considering?” at the bottom of the screen, with each choice linked to either the LEFT or RIGHT arrow key. To reinforce the association between the keys and the choices, the screen displayed each alternative alongside a corresponding arrow picture. The LEFT arrow key represented the deontological option where applicable, and the right key represented the utilitarian option. Henceforth, we refer to these alternatives as D and U, respectively.

Participants were instructed to pay attention to their thoughts and indicate their preferred choice at any given moment by pressing the corresponding arrow key in the deliberation phase. This setup allowed participants to actively and continuously express their preferences as they deliberated on the problem. They could report their preferences multiple times (but at least once) at any point during this phase. Trials where no key was pressed were excluded. The deliberation phase lasted a minimum of one minute, although participants could take longer if needed.



Press SPACE when ready

IMAGINE...

You are a young architect visiting one of your construction sites with your boss. Your boss is a despicable individual who makes everyone around him miserable including you. It occurs to you that if you were to push him off of the building you are inspecting he would fall to his death and ever one would think it was an accident.

What possibilities are you considering?

FINAL DECISION

Would you push your boss off of the building in order to get him out of your life?

RATING SCALE

How confident are you about your answer?

How conflicted did you feel while answering?

How diffificult was the question to answer?

Do you think your answer will change in the future?

Figure 4.2: Trial structure in Study 2 experiments had the text of the problem displayed centrally on the screen during the deliberation phase. A moral dilemma is displayed in this schematic. In Experiment 3 of Study 2, participants saw two premises and a conclusion of a syllogism on separate lines. Participants record their final decisions in the decision phase on a separate screen any time after a mandatory period of one minute is over. Every trial concludes with rating the reasoning experience on subjective measures.

After the deliberation phase, participants proceeded to the next screen to make their final decision using the arrow keys, which corresponded to the same options as before. Lastly, participants rated their experience of reasoning on the following four 5-point scales: (a) How conflicted did you feel while answering? (b) How confident do you feel about your answer? (c) How difficult was the question to answer? and (d) Do you think you will change your mind about your answer?

Participants completed the experiment alone in a laboratory room after filling in the consent form and reading the instructions. Experimenter remained in the room to clarify any doubts about the task. The initial set of instructions that participants read is as follows:

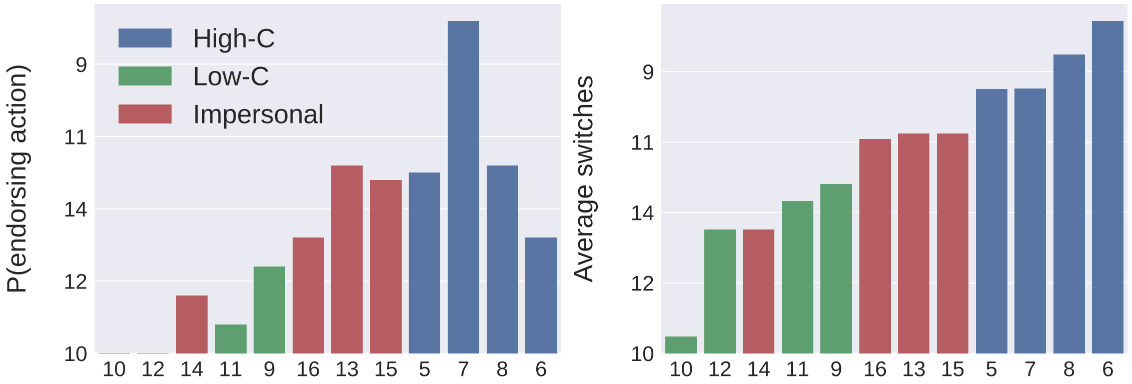
> "Welcome to the experiment!
>
> This experiment will take 30 minutes to complete. The aim of this experiment is to understand how individuals arrive at their choices. Let's take an example to understand this:
>
> Imagine being in an unfamiliar restaurant, faced with a tempting menu that offers you options like farm-fresh pasta, pizza, and garlic bread with spread. As you contemplate your choice, various arguments may cross your mind, such as the comfort of pizza or the lighter option of bread and spread when not very hungry. Your task is to pay close attention to these arguments, categorize your thoughts based on the preferred option, and indicate your choice at the end of each scenario.
>
> In this experiment, you will read a few stories and will be asked to think and make a choice at the end. When you are deciding you have to categorize your

> thoughts based on which choice they indicate. Hence, the task for you is to be attentive to your thoughts and indicate which option you prefer currently while reasoning. Whenever you find your thoughts are leaning towards one of the options, you can report your preference by pressing the corresponding key on the keyboard. Feel free to press these keys multiple times and in any order.
>
> Now, let's walk you through an example trial within the experiment."

After these preliminary instructions, we showed participants screenshots of a dummy trial with an non-moral problem to help navigate the task. Each phase of the experiment carried specific instructions. The experimenter read the following instructions out loud while displaying the screenshots from the dummy trial to the participant.

> "All trials are self-paced. A trial is made of 4 screens. Press SPACE to continue.
>
> This indicates start of a new trial. You can rest on this screen between trials. [First screen from the left in Figure 4.2]
>
> This screen contains the context of the scenario. At the end of each scenario, you will be asked to report the possibilities you are considering. LEFT and RIGHT arrow keys will indicate different kinds of considerations. When you catch yourself thinking about one of them, press the respective key. After about 1 minute, you can press SPACE to go to next page. You can take longer if you have not decided by then. [Deliberation phase. Second screen from the left in Figure 4.2]
>
> This screen indicates you have to report your final decision by single key press of LEFT or RIGHT arrow key. [Final decision phase. Third screen from the left in Figure 4.2]
>
> Finally, you must indicate how it felt to answer the question. There will be four scales: (a) How conflicted did you feel while answering? (b) How confident do you feel about your answer? (c) How difficult was the question to answer? and (d) Do you think you will change your mind about your answer?" [Rating phase. Last screen in Figure 4.2].

Following this, participants completed two practice trials—one with a non-moral problem and one with a Low-C problem. Any difficulties encountered during these trials could be addressed by asking the experimenter for clarifications. The experimenter exited the room once the experiment commenced.



(a) (b)

FIGURE 4.3: Results of Experiment 1 of Study 2 across all moral items are depicted in the figure, where each bar represents a stimulus item (moral dilemma) and is color-coded by the dilemma type. (a) displays the proportion of trials in which the given action (usually U) was endorsed in the final decision, and (b) presents the average number of switches observed in each dilemma.

4.1.2 Results and discussion

The stimuli were categorized as Low-C and High-C post hoc by Koenigs et al. (2007) [100]. Experiment 1 of the current Study replicated this pattern of cohesiveness of judgments at the cohort-level, much like Experiment 1 from Study 1. Low-C dilemmas demonstrated fewer commitments to the U action, with no participant agreeing to take the action in two out of four of them. There was also a greater variability in endorsing the action in High-C dilemmas (Figure 4.3 (a)).

Our primary focus was connecting the cohort-level conceptualization of conflict to an internalized experience of it. Participants' momentary preferences, which are frequently subject to modification during deliberation, were identified with shifts in key presses during the deliberation phase. If on a trial a participant presses dissimilar keys one after the other then it was counted as a switch. All four High-C dilemmas showed frequent switching in preferences compared to Low-C (Figure 4.3 (b)). This observation aligns with the prediction that for these stimuli, cohort-level disagreements may indicate internal conflict within the individual. To account for effects of both participant and item on switching, we ran an LME model treating participants and items as random intercepts, revealing that Low-C dilemmas, indeed, recorded fewer switches than high-conflict (Table 4.1).

Following that, we investigated the pattern in which D and U inclinations are considered. Bago and De Neys (2019) employed the two-step paradigm to discern the temporal order in inclinations by requiring a quick response at the beginning of the trial, followed by a

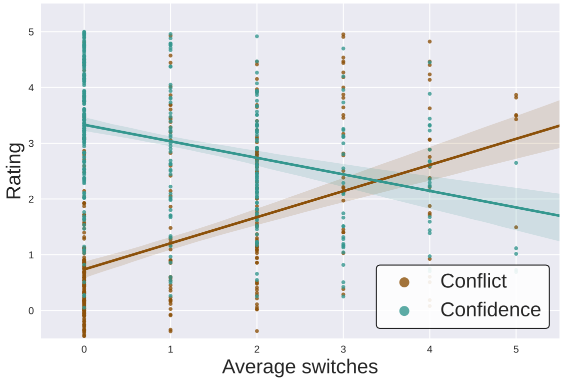


FIGURE 4.4: Regression plot between conflict and confidence ratings and switches recorded during the trial in Experiment 1 of Study 2.

reasoned response when participants had their final judgment ready [6]. While this method allowed for the dissection of the process to a certain extent, constraining the investigation to specific time windows excludes a significant portion of the reasoning process that follows the initial inclination. To demonstrate this limitation, we created key-press pairs of the first and the last keys pressed on a trial during the deliberation phase, resulting in four possible pairs: DD, DU, UD, and UU. Here, the first letter in each couplet denotes the first key, and the second letter signifies the last key pressed during this period.

Notably, all four response change types were reported in moral dilemmas (see Table 4.3 for how frequently these pairs were observed in moral trials). We aimed to determine if there is a predictable order in which these inclinations come to the reasoners' minds. According to the corrective default-interventionist model of DPT, reasoners should be inclined toward the D alternative at the beginning of the trial and switch over to U if mental resources

TABLE 4.1: An LME model of switches recorded in Experiment 1 of Study 2 by conditions with participants and items as random effects. Conditions are dummy coded with Low-C as the reference level.

Switches ∼ condition

Fixed effects					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	0.8200	0.24	20.03	3.416 **	Participant	0.60
High-C	1.2357	0.26	9.04	4.76 **	Item	0.05
Impersonal	0.5600	0.26	8.98	2.16 .	Residual	2.09

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.2: Results of Experiment 1 of Study 2: LME models of (a) conflict, (b) confidence, (c) difficulty, and (e) changes of mind ratings by switches. Participants and items as random effects.

(a) Conflict rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	1.1614	0.2	19.95	5.82 ***	Participant	0.21
Switches	0.3605	0.04	293.27	9.07 ***	Item	0.3
					Residual	0.98

(b) Confidence rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	3.0634	0.16	20.45	18.60 ***	Participant	0.15
Switches	-0.2246	0.03	293.84	-6.94 ***	Item	0.2
					Residual	0.656

(c) Difficulty rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	1.1920	0.22	17.57	5.46 ***	Participant	0.21
Switches	0.3009	0.04	291.85	8.42 ***	Item	0.41
					Residual	0.78

(d) Change of mind rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	0.6612	0.14	25.28	4.57 ***	Participant	0.17
Switches	0.2599	0.03	296.45	7.96 ***	Item	0.12
					Residual	0.66

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

permit which should make the DU transitions more prevalent than UD [81, 128]. However, in all three moral dilemmas, the frequency of DU was not significantly more than UD (see Table 4.3 for frequencies of key-press pairs; Impersonal: $\chi^2(1) = 12.78, p < .001$; Low-control: $\chi^2(1) = 3.33, p = .07$; High-control: $\chi^2(1) = 1.93, p = .17$). Furthermore, although DD and UU were the most frequently observed key-press pairs, between the first and the last keys pressed, participants had switched at least twice on 58% and 33% of trials, respectively, to land on the same key they started with. Such oscillations in reasoning are challenging to explain under the sequentiality assumption made by some models of DPT.

Finally, we examined participants' subjective ratings recorded at the end of each trial. While participants rated trials on four scales, our primary focus was on conflict and confidence ratings, commonly used indicators of conflict in decision-making [61, 115, 131]. Overall, more switches were associated with increased reported conflict and decreased confidence in the final answer (Figure 4.4). This trend is broadly reflected at the item-level, where conflict is positively correlated and confidence is negatively correlated with vacillations. However, given the item-wise variability in these associations, we modelled the subjective ratings by switches accounting for participant- and item-level random effects. Trials with more switches were correlated with increased level of conflict, reporting the problem as difficult and subjective sense that their mind about the answer might change in the future (Table 4.2 (a), (c) and (d), respectively). On the other hand, confidence ratings dropped with more vacillations (Table 4.2 (b)).

In summary, like in Koenigs et al. (2007) and Experiment 1 from Study 1, results from the current experiment also suggest that for the stimuli under inspection, less conflicting moral dilemmas observed more unanimous judgments with most individuals not endorsing the action in Low-C scenarios, while responses were mixed in High-C as well as impersonal dilemmas [100]. Vacillations which are internalized within the reasoner mapped reasonably well with the overall cohort-level disagreements in final decisions as well as subjective feeling of conflict. Furthermore, we employed vacillations as an analytical tool to examine the reasoning process and scrutinize models of moral reasoning by outlining how preferences evolve during deliberation. These patterns revealed that infrequent transitions that were not anticipated under certain models of DPT were common in reasoning. Frequent shifts between alternatives are challenging to explain under certain theoretical models.

Table 4.3: Response changes during deliberation in Experiment 1 and 2 in Study 2.

Experiment 1					**Experiment 2**				
	DD	**DU**	**UD**	**UU**		**DD**	**DU**	**UD**	**UU**
Impersonal	41	4	22	33	**Non-conflict**	4	8	4	50
Low-C	80	4	12	4	**Conflict**	8	5	7	46
High-C	32	19	11	37					

4.2 Experiment 2

The results of Experiment 1 above show that our measurement of mental vacillations when summed across a trial correlates well with the expectation of people experiencing conflict during reasoning, as operationalized via cohort-level disagreement by Koenigs et al. (2007) [100]. In Experiment 2, we sought to validate our findings from Experiment 1 with a more recent definition of conflict. Bago and De Neys in 2019 manipulated conflict in moral decisions in terms of convergence of deontological and utilitarian principles on a choice [6]. They propose that when these two principles contradict each other, people feel conflicted. We tested this definition of conflict in moral decisions in a pre-registered study below.

4.2.1 Method

Participants

Sample size and hypotheses for Experiment 2 were pre-registered (see here). We collected data from 27 participants and each participant was compensated with Rs. 100/- for their time. Four participants' data did not record reliably due to a technical issue in the software and one participant failed to qualify our inclusion criterion (ie., did not record any key press during the deliberation phase in any of the trials). We analyzed data of 22 participants (8 females; Mean age = 20.3 years). The experiment design was approved by the IEC.

Materials

In Experiment 2, we utilized a moral stimuli set from Bago and De Neys (2019) and non-moral stimuli from Koenigs et al. (2007) [6, 100]. By Greene et al.(2001, 2004)'s categorization, all moral dilemmas could be considered impersonal with a choice between an action and its omission [80, 81]. Consequence of each action within a dilemma was such

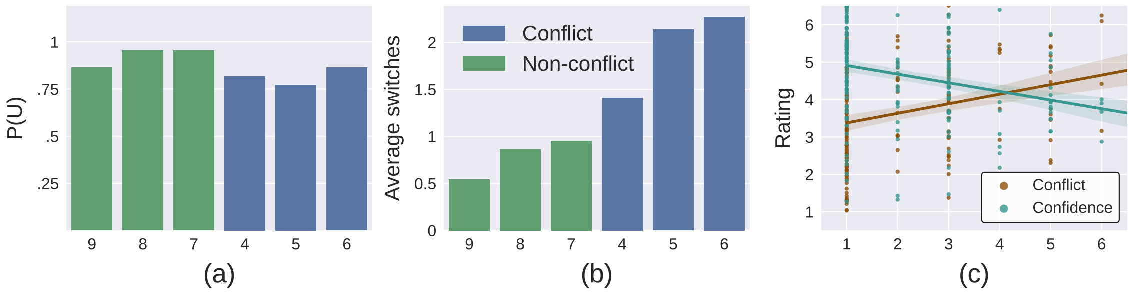


FIGURE 4.5: Results of Experiment 2 of Study 2 across all trials are depicted in the figure, where each bar represents an item and is color-coded by condition. (a) and (b) represent item-wise proportion of U responses as the final decision and switches, respectively. (c) shows Spearman correlations of switches within an item to the subjective ratings of conflict and confidence reported at the end. Bars are color coded based on the problem type (conflict or non conflict).

that it killed a group of people as a side-effect of saving another group. Also note that the consequence of impersonal actions here is different from the stimuli used in Experiment 1 of Study 1 and Experiment 1 of the current study). Bago and De Neys operationalized conflict in moral dilemmas based on the convergence of deontological and utilitarian principles. In conflict moral problems, the consequence of the action was such that it saved a larger group of people at the cost of harming or killing a smaller group. These trade-offs mirror the typical trolley problem setup and their variations such that the choice is between U and D [77, 169]. In non-conflict moral dilemmas, the action led to the death of a larger group to save a smaller group. Following is an example of choice alternatives and their trade-offs in a non-conflict dilemma:

> "...If you activate the emergency circuit to transfer the oxygen, these 11 miners will be killed, but the 3 miners will be saved. Would you activate the emergency circuit to divert the oxygen in the shaft?"

The authors call these dilemmas non-conflict because both deontological and utilitarian principles converge on the choice of not endorsing the action. The deontological principle in this context is to minimize harm. We refer to this converging choice as U in non-conflict trials for the ease of discussion. In conflict trials, the U choice refers to the action that saves many by killing a few (see Bago and De Neys (2019) and Conway and Gawronski (2019) for a detailed description of these problems [6, 27]). Following is an excerpt from a conflict dilemma. It demonstrates that in these dilemmas the choice is between a utilitarian action and a deontological inaction:

> "...If you push the button and divert the fire into the sideline, this building will explode and kill the 4 people in it, but the 12 in the building above the main line will be saved. Would you push the button to divert the fire explosion?"

Table 4.4: Results from Experiment 2 of Study 2: LME model of (a) switches by conditions with participants and items as random effects. Conditions are dummy coded with non-conflict dilemma-type as the reference level. (b) and (c) present results of LME models of conflict and confidence by switches, respectively, with both participants and items as random effects.

(a) Switches ∼ condition

Fixed effects					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	0.7879	0.31	13.01	2.53 *	Participant	1.18
Conflict	1.1515	0.3	4.0	3.9 **	Item	0.02
					Residual	2.44

(b) Conflict rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	2.8049	0.21	22.6	13.09 ***	Participant	0.77
Switches	0.1730	0.05	90.38	3.50 ***	Item	0.0001
					Residual	0.83

(c) Confidence rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	3.5449	0.19	22.03	18.94 ***	Participant	0.49
Switches	-0.1507	0.04	118.85	-3.35 **	Item	0.03
					Residual	0.65

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

All the stimuli can be found in the Appendix.

Procedure

Experiment 2 retained the trial structure from Experiment 1 above with a minor adjustment in the rating phase by including only confidence and conflict scales. The instructions remained unchanged from those provided in Experiment 1.

4.2.2 Results and discussion

In Experiment 2, we expected people to be highly cohesive in their final answers at the group level on both conflict and non-conflict items based on response pattern reported in the original paper. Although this was the case in both conflict and non-conflict cases, the vacillations reveal that the reasoning process was indeed distinct between these two problem types. Participants switched more frequently on conflict than non-conflict items (Figure 4.5 (b)). An LME model of switches with random effects of the participants and items corroborated this observation (Table 4.4 (a)). Vacillations also correlated with subjective measures of conflict such that trials on which participants switched more often, they reported greater conflict in reasoning and lesser confidence in the final judgment (Table 4.4 (b) and (c)).

However, a more nuanced narrative emerged when examining response transitions. Bago and De Neys's hybrid model of the DPT predicts that individuals who provide a U response both rapidly and after careful consideration (ie., UU transitions) would not need to alter their preference during deliberation [6]. Contrary to this prediction, our findings suggests that even if they start and end with the same response, participants may not necessarily adhere to that choice throughout the deliberative process. Approximately 35.42% of all UU trials exhibited at least 2 switches in between. Hence, once again, vacillations in preferences offer a more informative metric than simple response transitions, capturing the dynamic nature of the reasoning process between the initial and final decision points.

4.3 Experiment 3

Experiment 3 in Study 2 uses syllogisms as stimuli. One of the primary motivations for employing a different type of stimulus was to investigate the generalizability of our Switch paradigm to another form of reasoning. Additionally, theories related to syllogistic reasoning provide hypotheses specifically about the underlying cognitive processes involved in problem-solving that can be tested with the paradigm. Particularly, we investigate belief bias in single- and multiple-model syllogisms. Our motivation was to gain insights into syllogistic reasoning with minimal interference, ultimately contributing to a more nuanced understanding of the cognitive mechanisms at play.

4.3.1 Method

Participants

Based on a pilot study, we calculated the sample size of 25 (for a detailed pre-registration report, see here). Expecting some data loss, we collected data from 30 participants (7 females; mean age = 21.7 years). Participants were compensated with Rs. 100/- after completing the experiment. The experiment design was approved by the IEC.

Materials

Participants solved 8 categorical syllogistic reasoning problems. We employed a within subject 2x2 design with two factors: validity (whether the conclusion follows logically from the premises) and believability (whether the conclusion is believable). To manipulate the believability of the conclusions, we used the materials from Robison and Unsworth (2017) and Evans and Barston (1983) [49, 142] (see Table 4.5 for all 8 conclusions).

Our stimuli were also divided in one of the two forms, single- or multiple-model [122]. Single-model syllogisms were taken from Robison and Barston (2017) [142] and had the form:

All A are B.
All B are C.
Therefore, all A are C. (valid)

OR

Therefore, all C are A. (invalid)

Multiple-model syllogisms are those for which more than one conceptually distinct model can be constructed, and to necessarily conclude about the validity of the conclusion, all models need to be considered. These were of the following form, taken from Evans and Barston (1983) [49]:

Table 4.5: Conclusions in 8 syllogisms used in Experiment 3 of Study 2. Conclusions with universal quantifier "all" (1, 2, 5, and 6) are taken from Robison and Unsworth (2017) [142]. Rest are from Evans and Barston (1983) [49].

	Conclusions	Believable?
1	All lollipops are made out of sugar.	Yes
2	All animals that are able to swim spend majority of their lives in the water.	Yes
3	Some highly trained dogs are not police dogs.	Yes
4	Some addictive things are not cigarettes.	Yes
5	All college professors have medical degrees.	No
6	All objects with sides of equal area are objects with six sides.	No
7	Some priests are not religious people.	No
8	Some deep sea divers are not good swimmers.	No

Some A are B.

No B are C.

Therefore, some A are not C. (valid)

OR

No A are B.

Some B are C.

Therefore, some A are not C. (invalid)

For this experiment with syllogism, conflict trials were those in which there was a conflict between logic and believability (invalid-believable and valid-unbelievable), while in non-conflict trials, logic and believability both coincided on the same choice (valid-believable, invalid-unbelievable). List of the stimuli used are in the Appendix.

Procedure

The overall trial structure remained the same as in Experiment 1 and 2 of the current study (see Figure 4.2). Participants saw two premises and the conclusion in the middle of the screen with each statement on a separate line during the deliberation phase. The prompt in the deliberation phase was slightly altered to read "Which option are you considering?" for better interpretability. The RIGHT and LEFT arrow keys corresponded to the conclusion being logically TRUE and FALSE, respectively, in both deliberation and decision phases of the trial. On the last screen, participants reported on a 5 point scale how conflicted they felt while solving the syllogism and how confident they were with their final answer.

Since we used syllogisms which has a correct and an incorrect answer, the instructions were changed to include a dummy syllogism to explain the task. The preliminary instructions for Experiment 3 read as follows:

> "In this experiment, we are investigating how people solve a particular kind of problem. Let me explain it with an example.
>
> Imagine that you are trying to solve a multiple-choice question with only two options: A and B. Sometimes you know the answer right away. However, some other times, you may think A could be right before correcting yourself and saying B is right. Perhaps, you switch again to A and record it as your final answer.
>
> In this experiment, the task is to pay close attention to these thoughts before you settle on your final answer. As you notice yourself leaning toward one of the options, you have to indicate it by pressing a key on the keyboard.
>
> Any questions so far? Please ask now."

After clarifying any doubts participants had so far, they were explained the task by using a dummy syllogism.

> "You will see 3 statements. The first two statements will give you some information about the third one. You have to say if the third statement following 'Therefore. . . ' is a logically valid statement assuming that the first two statements are true.
>
> For example:
>
> All mammals are zephrodytes.
> All zephrodytes fly.
> Therefore, all mammals fly.
>
> Is the third statement 'TRUE' or 'FALSE'?
>
> You will solve 8 such problems. For each problem, you will have at least 1 minute to solve, but you can take longer if needed. Remember: while you are

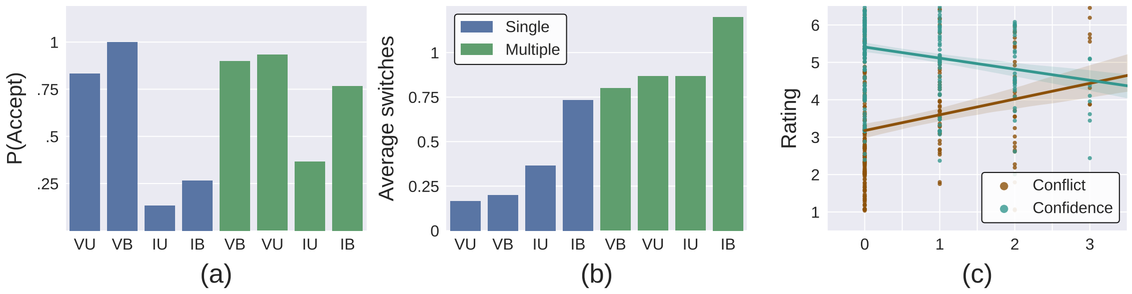


FIGURE 4.6: Results of Experiment 3 of Study 2 across all syllogisms. Each bar is a syllogism which is either valid (V) or invalid (I) and has either a believable (B) or unbelievable conclusion (U). All bars are color-coded by the model type. (a) and (b) depict the proportion of trials when the conclusion was accepted as valid and the average number of times participants switched between options while deliberating, respectively. (c) shows regression plots of conflict and confidence ratings onto the number switches recorded in the trial.

> thinking about this problem you also have to indicate which answer you are considering.
>
> Any questions so far? Please ask now."

After reading these instructions, participants saw screenshots of a dummy trial (with the same example used above) along with description of the trial structure. Instructions were followed by the main block with 8 syllogisms. At the end of the experiment, they filled out the actively-open minded thinking questionnaire (AOT) [10]. We also asked them if they were familiar with syllogisms.

4.3.2 Results and discussion

Twenty-six out of 30 participants were acquainted with syllogisms and had prior experience solving similar problems; however, none of the participants were excluded from the analyses. Despite their familiarity, participants still showed belief bias effect in their judgments. We operationalized participants' final decisions recorded following the deliberation phase in terms of accepting or rejecting the conclusion. We conducted 2x2 repeated measures ANOVAs with believability and validity as predictors for both single- and multiple-model syllogisms [1]. According to the belief bias studies, the difference in acceptance rates between valid and invalid trials is larger when the conclusion is unbelievable than when believable. However, this particular pattern of interaction remained predictive of acceptance rates only in multiple-model syllogisms (Table 4.6). Reasoners were highly accurate

[1] Although running an ANOVA on binomial data is ill-advised, we followed the convention to compare results with the previous literature. A more appropriate logistic model is reported in Appendix.

in judging the validity of single-model syllogisms which contained universal premises and conclusions which is in line with representative results from literature in belief bias studies [122].

Although people show belief bias in their judgments produced at the end of a trial, the reasoning process itself is much more textured. Consistent with the results from Experiment 1 and 2 of this study, trials with syllogisms in which participants shifted between their preferences often were likely to be rated higher on conflict and lower on confidence (Table 4.7). Multiple-model syllogisms which allow premises to be modeled in more than one way saw more switches in preferences than when they could be arranged only in one way like in single-model syllogisms, consistent with the mental models theory and misinterpreted necessity models (Paired t test: $t(29) = 3.43, p = .002, CI = [0.23, 0.90]$). Models of logical thinking discussed before also purport a temporal order in which arguments are considered. When believability and logical validity converge on the same answer (valid-believable and invalid-unbelievable syllogisms), participants were likely to consider the convergent choice first more than chance (Proportion = .78; One proportion $z(1) = 7.23, p < .01$). Contrastingly, when these two factors contradict, like in invalid-believable and valid-unbelievable syllogisms, participants' first preferences were more influenced by the logical validity than the conclusion's believability (Proportion = .69; One proportion $z(1) = 4.47, p < .01$). This aligns with the mental models and misinterpreted necessity theories because both theories predict that individuals initially assess validity of conclusions, but contrasts with the selective scrutiny model, which expects that individuals will examine a syllogism's believability before checking its logical validity.

According to the mental models theory, people should vacillate more while deliberating on valid-unbelievable syllogisms because people entertain alternate models only in such cases. However, participants switched the most on the invalid-believable syllogism (last bar in Figure 4.6 (b)). Switches recorded on invalid-believable trials were significantly

Table 4.6: Results of the analysis of variance for single-model and multiple-model syllogisms, from Experiment 3 in Study 2. '*' denotes the effect was significant. η_p^2 are partial eta^2 for the effect.

Single-model					**Multiple-model**			
Effect	**DF**	**F**	**p**	η_p^2	**DF**	**F**	**p**	η_p^2
Validity	(1, 29)	102.9	$< .001*$	.78	(1, 29)	19.12	$< .001*$	.4
Believability	(1, 29)	6.37	.02 *	.18	(1, 29)	13.05	.001 *	.31
Interaction	(1, 29)	0.14	.71	.01	(1, 29)	10.63	.003 *	.27

Table 4.7: Results of Experiment 3 of Study 2: LME models of switches in syllogistic reasoning predicting (a) conflict and (c) confidence ratings with participants as a random effect.

(a) Conflict rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	2.1138	0.15	34.18	14.29 ***	Participant	0.44
Switches	0.5172	0.07	232.74	7.29 ***	Residual	1.24

(b) Confidence rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	4.4087	0.09	38.88	50.74 ***	Participant	0.1
Switches	-0.2954	0.05	237.90	-5.55 **	Residual	0.74

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

more than switches on valid-unbelievable ($t(29) = 2.76, p = .01, CI = [0.12, 0.78]$) and invalid-unbelievable syllogisms ($t(29) = 2.57, p = .015, CI = [0.07, 0.63]$). While the misinterpreted necessity model anticipates reasoners to vacillate when making judgments on invalid multiple-model syllogisms, in our experiment, they only do so in invalid-believable syllogisms. Switching between preferences is comparable in valid syllogisms (both valid and invalid) and invalid-unbelievable cases. Therefore, the observed pattern of vacillations was not entirely consistent with either the mental model theory or the misinterpreted necessity model. However, our stimuli were limited to one problem each in each condition within single- and multiple-model syllogisms. It remains to be seen whether these predictions extend beyond the examples we have used in this experiment.

Finally, the AOT scores did not correlate with switches ($r_{switches}$ $= 0.25, p = .17$) as we had expected. AOT is indicative of a tendency of an individual to engage in in evaluative thinking [159]. While there is a trend in the anticipated direction for switches, suggesting that a reasoner who considers more diverse information is more likely to score higher on AOT, it remains unclear whether the lack of effect is due to a small sample size.

4.4 Discussion

Experiencing conflict during reasoning is a common phenomenon, yet its empirical measurement poses challenges. Researchers often rely on trial-level metrics of conflict such as final judgments or choices produced after reasoning and RTs, which limit our insight into how the conflict evolves. In Study 1, we incorporated mouse-tracking to this end. However, pinpointing when and how the adjustments in preferences occur during a given trial remains challenging. With real-time process-tracing methods such as think-aloud protocols, there is some level of distortion risk, meaning that the measurement of the phenomenon could itself hinder it.

Our conflict measurement method, in contrast, allows experimenters to observe the complete time-course of a respondent's decision, thereby facilitating more nuanced analyses. We believe it is also less intrusive than the think-aloud method because reasoners just have to press a key to report their inclination without having to verbalize or give an explanation for their preference. Additionally, our main dependent variable, switches in preference, can be readily analyzed without the need for inter-rater reliability checks or complex coding techniques. This opens avenues for exploring finer details, including the potential incorporation of gaze or neurophysiological markers to measure the experience of conflict in future research—a thread we explore in the next study of this thesis. In three experiments, we demonstrate a reasonable correlation between our measurement and participants' subjective experience of conflict in two different fields of reasoning. On problems that people vacillated while reasoning were also those which they rated as conflicting. Their confidence in their final answer was also predicted by how often they vacillated while reasoning.

Measurement of mental vacillations also offer constraints for theories of reasoning and decision-making. Our results from Experiment 3 in Study 2 illustrate that while the final acceptance rates aligned well with predictions from the mental models theory and misinterpreted necessity models, the pattern of vacillations between choices did not support either of them completely. Similarly, although Bago and De Neys (2019)'s two-step method might lend support to a hybrid model of dual-process moral reasoning, our results from Experiment 2 in this study unveil a less straightforward narrative as individuals frequently revisited alternatives considered before [6].

Our findings, while not determinative, are inconsistent with the two-system account that overlay a slow deliberative process on a faster heuristic process, and consistent with simpler race-to-threshold accounts of preference formation, with duality emerging as a property of the set of hypotheses under consideration. As an example of such a single process account, Srivastava and Vul [158] present an interesting demonstration that the characteristic signatures of both System 1 and System 2 decisions can be elicited from a single race-to-threshold model simply by changing the set of options an observer is selecting between, with fewer choices yielding System 1-like behavior, and more choices yielding System 2-like behavior.

Complementarily, Gürçay and Baron [82] propose that the choice structure of some problems, particularly moral dilemmas, is such that it evokes feelings of conflict. They conceptualize the reasoning process as a competition among the alternatives to control the final decision such that the reasoner may favor any of these alternatives, in no particular order, before settling on a choice. This idea is consistent with preference switching or vacillations in reasoning. Single-process models like these offer a simplified yet effective framework for understanding decision-making [24, 158]. These models propose that reasoning follows a singular, unified process that operates under a single algorithm, irrespective of the speed or duration of reasoning. In other words, the model does not differentiate between fast or slow decisions but instead posits that the decision-making algorithm works consistently across different time scales and contexts. The model can be refined to account for individual differences and the specific contexts in which decisions are made. For example, consider a situation where an individual is faced with a dilemma that pits self-interest against collective welfare. If a person has a strong motivation to prioritize the well-being of others, the model could be adjusted to reflect this by lowering the threshold required to choose a response that favors the collective good. In contrast, if an individual tends to act based on personal benefit, the model might set a higher threshold for responses that involve sacrificing personal gain for others. By incorporating such parameters, these models can better capture the complexity of decision-making processes, taking into account not just the mechanics of reasoning but also the personal values, biases, and environmental factors that may shape choices in real-life situations. Such single process models are re-examined as potential explanations for results in moral and logical decision-making in the concluding chapter (Chapter 6) of the thesis.

In this study, we introduced a new method to measure and employ vacillations as an indicator of conflict. Although our paradigm allows a more granular insight into the process of reasoning, there is room for enhancement to extract even more information about the intricacies in this process. To continuously monitor preferences, joysticks can be employed, with the direction of movement indicating the current preference and the distance displaced reflecting the extent of certainty associated with that preference simultaneously. Future

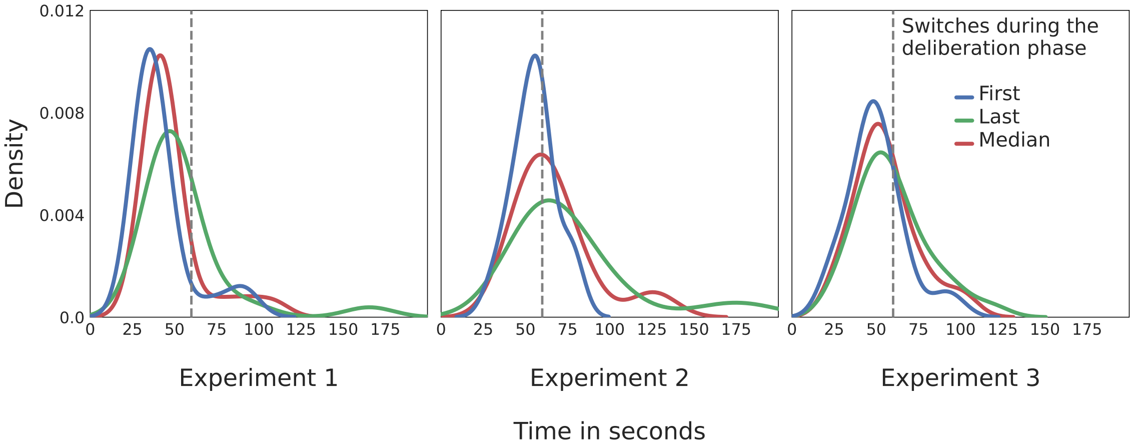


FIGURE 4.7: Density plots of response times of the first, median and last switches for each participant in Study 2. X axis is time in seconds. The deliberation phase starts with the presentation of the problem in text format to participants at X = 0. The dashed vertical lines are at X = 60 seconds after which participants could report their final decision.

studies may also explore modifying instructions to align with specific task demands such as tracking changes in confidence instead of preferences during reasoning.

Response keys were mapped to the action-inaction alternatives or the valid-invalid conclusions in the moral and logical reasoning experiments. The LEFT key represented inaction or the invalid option, while the RIGHT arrow key was associated with endorsing the action or considering it valid. This consistent mapping aimed to minimize confusion in participants' key assignments. However, this approach could potentially introduce confounds in participants' responses. Further, we had imposed a one-minute interval for deliberation to prevent inattentive responding and ensure active reasoning. However, such enforced temporal constraints on reasoning might introduce confounding factors. The actual time participants take to decide is unclear as a decision may have been reached well before the one-minute mark. A participant may continue to engage in reasoning even after reaching a decision, likely increasing the frequency of switching. A point to note here though is that this additional reasoning, forced or not, may not be ineffectual as it might alter the confidence in the judgment with more information considered.

The imposition of forced deliberation conditions may also impact the timing of key presses, potentially complicating the interpretation of switches as indicators of internalized preference shifts. Alerting participants that the one-minute period has ended could influence their key-pressing behavior. This is more problematic for our measurement if it pushes participants to switch when close to the prompt. We plotted the timing of switches during the deliberation period in Figure 4.7 to examine when these switches occur. In Experiment

1 of this study, participants' initial switches appear well before the one-minute mark (illustrated by grey lines in all three panels). However, in Experiments 2 and 3, switches seem to align with this mandated deliberation time. In future research, a more comprehensive understanding of the effectiveness of vacillations could be achieved by eliminating the imposed minimal time for responding. Without the time restriction, it would be possible to investigate the timing of vacillations, how it interacts with contextual factors and whether there are any individual-level differences. Furthermore, drift-diffusion models (DDMs) could provide additional insights into these variables by using this primary data to model preference shifts within decision-making as an evidence accumulation process.

4.5 Study 2 in Review

In summary, we present a new experimental paradigm for measuring how individuals vacillate between choices while deliberating. Empirical evidence obtained across three experiments indicates that this measurement holds the potential for greater insights than can be gleaned from trial summary statistics such as response times or cohort-disagreement levels. Our results demonstrate that prevailing theoretical accounts of reasoning struggle to adequately explain the sequence of vacillations seen in peoples' judgments. We anticipate that the enhanced visibility into the deliberation process afforded by our paradigm will contribute to refining and improving these theoretical models.

Chapter 5

Detecting Conflict with Eye-Tracking

From Study 1, which inspected conflict at the response level, we transitioned to Study 2 to test the actual shifting preferences within a trial. Vacillations, as measures of conflict, were both internally and externally validated across two very different kinds of problem-solving tasks viz., moral dilemmas and categorical syllogisms. While there is a correct answer in the latter, judgments on moral dilemmas cannot be categorized as right or wrong as each choice can be defended—at least on the surface—with an ethical principles. In fact, we explicitly stated in our instructions that there may be no right or wrong answers for these problems. In syllogistic tasks, this comparison to a set standard is present, as the task involves judging the logical validity of the syllogism. People also know how to solve syllogisms using strategies like logic tables and Venn diagrams. Although vacillations elucidated on how the choice evolved, they did not capture the difference between these two types of reasoning. In Study 3, we test whether eye metrics can help address this gap.

We selected fixation duration and pupil size to investigate signature of conflict in reasoning. Since most visual information uptake happens during fixations, the average fixation duration is used as a proxy for depth of processing [144]. Pupil size is a physiological measure which has been employed as an indicator of cognitive control [177]. Our main aim was to explore if people re-engage in the problem after stating an initial preference differently when they are reasoning about moral or logical problems. In moral dilemmas, we expected the re-engagement to be similar irrespective of whether it is followed by a preference congruent or incongruent with the preceding choice, indicative of a continually experienced conflict due to a lack of clear predilection for a single ethical principle. Therefore, we expected fixation duration to be longer and pupils to be dilated more regardless of whether there is a shift in the preference in moral dilemmas. On the other hand, when participants solve a syllogism with strategies they have formally learned, conflict should only arise when their current strategy seems to lead to an incongruent answer. Hence, we hypothesized that indicators of conflict like longer duration and larger pupil size will be observed preceding a switch in their choice.

5.1 Experiment 1

So far in this thesis, conflict in moral dilemmas has been investigated at the response dynamics level and at the more phenomenological level by measuring vacillation. We have used the stimulus set from literature to validate our findings against the current models of moral reasoning. In both experiments (Experiment 1 of Study 1 and Study 2), this set of moral dilemmas fared well in replicating the original findings as well as the pattern of vacillations in different conditions. Therefore, as we test a new measure of conflict now, we employed the same dilemma set to test the efficacy of fixation duration and pupil dilation as indicators of internalized conflict.

5.1.1 Method

Participants

An email advertising participation was circulated to the student community at Indian Institute of Technology, Kanpur. Twenty-seven participants completed the experiment. We had to exclude two participants from analyses as their data files were corrupted while collecting data. Final analyses were performed on data from 25 participants (3 females; Mean age = 22 years). All participants were compensated with Rs. 100/- for their time. The IEC approved the study design.

Materials

Participants solved 16 problems which were either of the type non-moral or moral. Moral problems were of three conditions viz., low-conflict personal (Low-C), high-conflict personal (High-C), and impersonal. Each problem was written in the second person. All the stimuli used in this experiment were from Koenigs et al. (2007) and exactly the same as Experiment 1 of Study 2 (see Appendix) [100].

Procedure

The experiment was coded in the programming language Python using the package PsychoPy (version 2022.2.2) for displaying the stimuli [129]. We employed the Switch paradigm in this study, too, with few changes made to accommodate tracking eyes. Instructions for the eye-tracking part were given before the experiment began (mentioned below). Instructions for the task remained unchanged.

Participants began each trial when they were ready. Before seeing the stimuli, participants saw a small circle at the centre of the screen. Participants were asked to look at the centre of this circle till it disappeared. This fixation dot served to correct for drift in eye movements. The experimenter verified that the drift was within acceptable limit (3°) before allowing the trial to proceed. The deliberation phase began with an "Imagine" screen which displayed the text of the problem (see Figure 4.2). This text contained the general context of the problem and a prompt that read "What possibilities are you considering?". Each problem had two alternatives, either endorsing an action and not. These options were displayed with their corresponding arrow keys. Like Experiment 1 and 2 from Study 2, the characteristically deontological response (D), or withholding of the action, was associated with the LEFT arrow key. The utilitarian action (U) was mapped to the RIGHT arrow key wherever the alternatives could be categorized as such. Participants could press these keys any number of times and in any order during the deliberation phase. This screen remained at least for 1 minute before participants could record their final answer on the next screen. After recording their final answer, participants were presented with two rating scales to assess their subjective experience of reasoning on the problem they had just encountered: conflict experienced while deliberating and their confidence in the final judgment.

To ensure proper eye tracking, the experimenter remained in the room while participants completed the experiment, seated behind a curtain out of the participants' view. Participants were informed that the experimenter could neither see their screen nor their keyboard during the experiment. This setup was designed to minimize any potential influence of the experimenter's presence on the participants' responses.

Participants eye movements were recorded using the eye-tracker Eyelink 1000 plus (SR Research Ltd.). The officially reported accuracy of this model of eye-tracker is between

0.25° to 0.5°. Participants were seated at a distance of 65 cm from the display screen (1920 x 1080 pixels) on which the stimuli were presented. A chin-rest was used to stabilize and limit excessive head movements for the stretch of the experiment. The room was darkened to avoid unwanted reflection from the ambient light sources. We captured monocular movements at 1000 Hz frame rate. Most participants' left eye was recorded unless it could not be calibrated properly.

Participants who typically wore glasses for reading did so during the experiment. Eyes were calibrated before receiving any task instructions with a 9-point calibration routine. Participants were asked to look at the centre of a circle as it appeared at the centre, midpoints of the four edges of the screen, and four corners of the screen one-by-one. Calibration was followed by the validation task to make sure the tolerance levels were below 3°. If participants did not meet this tolerance, then the calibration and validation tasks were repeated for the right eye. If this recording, too, was not satisfactory, then we cancelled their participation as the data from poorly calibrated eye could not be reliably used. After a successful calibration, participants were allowed to make any adjustments to the setup (chin-rest and the chair) before commencing the experiment. Participants were told to limit moving their head and chair once the experiment began. Again a set of calibration and validation task was performed with the selected eye before participants read the instructions and commenced the experiment. Before each trial, participants eyes were corrected for drift. If the error was beyond 3°, then calibration and validation tasks were repeated. If the error persisted, then the experiment was halted. The experiment was stopped midway for two participants due to poor calibration, but data up to the last successfully calibrated trial were included (12 and 15 trials were included for these two participants).

Preprocessing

Since our aim was to explore patterns in eye movements while people reason on moral dilemmas, we focused our analyses on data from the deliberation phase. This phase begins when participants are first presented with the problem in text format and continues until they press a key to end the trial, which can be done anytime after the mandatory one-minute period. While they deliberate, they record their preferences by pressing either LEFT or RIGHT arrow keys. To capture active reasoning, we defined the interested period starting from the moment participants were first presented with the problem till they pressed the last key while still in the delibeation phase. We excluded the fixation and pupil data recorded between the last key pressed in a trial and the key that signalled the

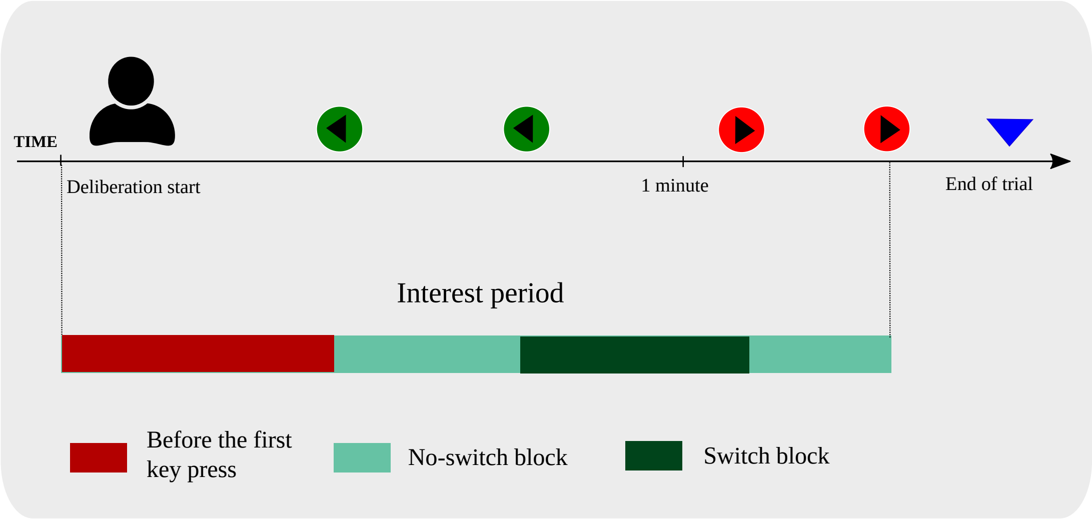


FIGURE 5.1: Schematic describing the interest period in Study 3 (Experiment 1 and 2).

end of the deliberation phase as we could not be sure if participants were thinking about the problem or simply waiting for the one-minute mandatory period to get over.

Next, we blocked the data based on whether it preceded a switch in preference. First, the data from the interest period were chunked between successive key presses. In some trials, participants pressed keys multiple times in quick succession. To avoid fragmenting the data further, we dropped keys which were pressed within 1500 ms of the succeeding key. For the remaining keys, data between two similar consecutive key presses (RIGHT-RIGHT or LEFT-LEFT) were categorized as a no-switch block. On the other hand, data from two dissimilar key presses (LEFT-RIGHT or RIGHT-LEFT) identified a switch in preference and hence, were categorized as switch block. Data before any key press were not included in these blocks and were treated as a separate subcategory (see Figure 5.1 for a schematic of the interest period).

The EyeLink software output contains samples produced at the frequency 1000 Hz. Each sample has x and y positions of the recorded eye, pupil size measured in terms of area, and events categorizing the current sample as a part of a saccade, fixation or a blink. We used the software's parsing algorithm to detect these events. A saccades is detected if the velocity of the recorded eye crosses 22°/s and if the acceleration is beyond $3800°/s^2$. If the saccade threshold is set off but the pupil data is missing for three or more samples in sequence then the sample is categorized as a blink. Fixations are periods that cannot be parsed as either a saccade or a blink. For the current study, we used the fixation duration and pupil size data output by the EyeLink software.

After segmenting the data from the interest period into blocks, we computed a rolling average of fixation durations using a window of 10 observations, where each observation represented a fixation duration. We used these averaged fixation duration data for analysis after removing any missing values. For pupil size analysis, we processed the samples produced by the EyeLink software. First, we removed any data that was recorded off the display screen. We also excluded all samples that were categorized as blinks by the software. However, the parsing algorithm often records incorrect pupil sizes around blink events. When the eyes start closing before a blink or opening after a blink, pupil sizes are recorded although these pupil size estimates are distorted by eyelids covering the pupils. Hence, we removed samples 100 ms before and after a blink was detected. Sometimes, during a blink, the eye-tracker mistakenly detects part of the closed eye as the cornea and records a pupil size. These often show as stray pupil sizes recorded among mostly missing pupil size data. We removed any such individual samples containing a series of at most 5 pupil sizes recorded in succession and surrounded by missing pupil data. After cleaning for blinks, we interpolated for the missing data using a cubic spline. Finally, we downsampled all pupil size data to 25 Hz.

Next, we wanted to standardize the pupil data to compare it across participants. Although pupil data are measured in terms of area, the output has arbitrary units (usually number of pixels covered by the pupil in a sample). As baseline pupil sizes are not equal across participants, pupil data in their arbitrary units cannot be compared directly. Typically baseline corrections are applied to circumvent this issue. Data from a baseline period, which is defined a priori, is averaged. The dilation or constriction of pupil is calculated relative to this baseline period. Such a period could not be easily defined in our experiment. One potential baseline could be the data collected before the first key press. But during this phase participants typically engage in reading the text for the first time and reasoning concurrently. Instead of defining a specific baseline period, we normalized pupil data for each participant. This approach allowed us to compare the relative change in pupil size against the participant's average pupil size (for a similar preprocessing example, see Purcell et al. (2023) [135]).

5.1.2 Results and discussion

Choice data and vacillations Overall, the utilitarian action (actions which either saved many by killing a few or were motivated by self interest) were chosen 42% of the times. Comparatively, non-moral actions (which included actions like resolving scheduling

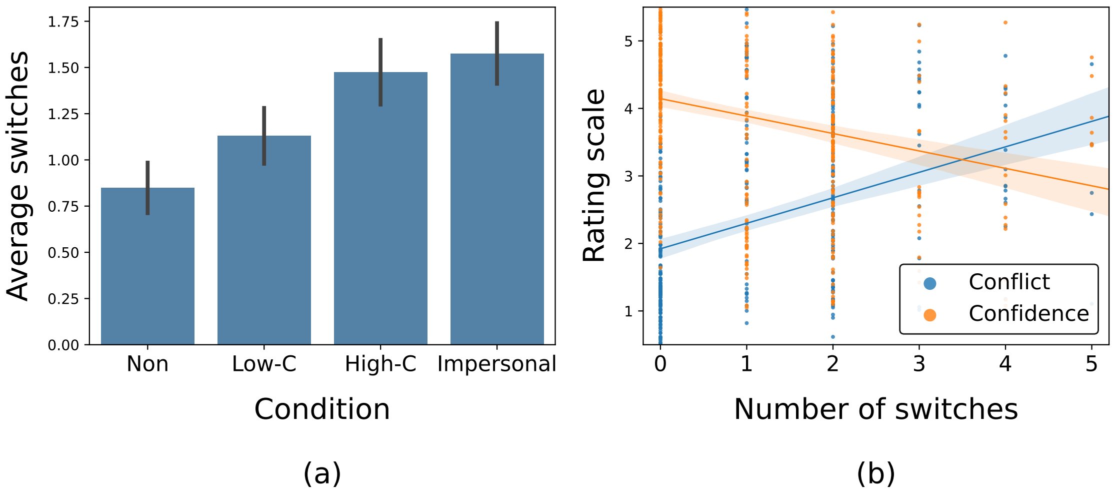


FIGURE 5.2: (a) Average switches in each condition with standard error bars and (b) correlation plot of rating scales (conflict and confidence) by number of switches in Experiment 1 of Study 3. Data points in the scatter-plot are jittered along the Y axis.

conflicts and measuring ingredients for cooking) recorded more consensus among participants (see Table 5.1). Participants also switched significantly more often on moral than non-moral trials (LME regression with participants and items treated as random effects: $\beta = 0.8500, SE = 0.28, z = 3.08$; $\beta_{Moral} = 0.5284, SE = 0.23, z = 2.3$; $Variance_{Participant} = 0.9124, Variance_{Item} = 0.099, Variance_{Resiadual} = 1.4786$). While we were not focused on the statistical difference between moral and non-moral trials, it is worth noting that non-moral trials showed a close to uniform consensus at the cohort-level in final judgments, whereas moral problems did not. In light of this, the difference in switches between the two types of problems aligns with the earlier definition of conflict given by Koenigs et al. (2007) as variability among individuals' judgments [100].

Next, we inspected the final judgments, switches and subjective ratings by the type of moral dilemmas (Low-C, High-C, and impersonal). We ran Bayesian hierarchical models in R using the package 'brms' [23]. The models were constructed with Low-C as the baseline condition and participants and items as random effects. We took prior estimates from analogous LME models from Experiment 1 of Study 2 (detailed information about the priors is in the Appendix Table A.2). Each model was run in 4 chains 20000 iterations, half of which were warmup draws. To compare the choice data in these problems (whether participant endorsed the action or not), we ran a generalized Bayesian hierarchical logistic regression model for Bernoulli data. We expected participants to reject the action in Low-C problems more than High-C and impersonal, since conflict is operationalized at the consensus level in Koenigs et al. (2007), Experiment 1 of Study 1, and Experiment

Table 5.1: Mean and 95% confidence intervals (CIs) for endorsement rates and switches recorded in Experiment 1 of Study 3. CIs were obtained through 1000 bootstrap resamples. Transitions show the frequencies of first-last keys pressed for each problem type.

	Endorsement rates	**Switches**	**Transitions**			
			DD	DU	UD	UU
Non-moral	.8990 [.84, .95]	0.8485 [0.59, 1.15]	15	10	6	68
Moral	.4257 [.37, .48]	1.3919 [1.22, 1.58]	137	37	37	85
Low-C	.15 [.08, .22]	1.1300 [0.84, 1.43]	75	6	11	8
High-C	.6598 [.57, .75]	1.4742 [1.15, 1.82]	25	21	6	45
Impersonal	.4747 [.38, .58]	1.5758 [1.27, 1.91]	37	10	20	32

1 of Study 2. Indeed, the endorsement rates were lower in Low-C. Switches also showed the same pattern, with Low-C recording least average switches (see the results in Table 5.2 (a) and (b)). We, then, correlated number of switches in a trial with the conflict and confidence ratings provided while using the priors from estimates of the LME models from Experiment 1 from Study 2 which had the same stimuli. Every switch incremented the conflict ratings on an average by 0.29 while reduced the confidence in the final judgment by 0.22 points (see the credible intervals for these estimates in Table 5.2 (c) and (d)).

In summary, results from Experiment 1 of Study 3 and Experiment 1 of Study 2, which employed the same stimuli, are in agreement. Participants switch more often on the moral dilemmas that typically generate dissimilar judgments. Vacillations as a measure of conflict were also internally validated by subjective ratings. In addition, DU and UD transitions in moral dilemmas, where the first and the second letter in the couplet identify the first and last key presses made on a trial (for instance, DU identifies trials in which the participant committed to the deontological preference first before recording utilitarian choice as their final answer after deliberations; see Chapter 4 for more details), were recorded exactly the same number of times. In moral dilemmas, out of all the DD and UU transitions 39% and 33% trials saw at least two switches between the first and the last key was pressed. These transitions demonstrate that, like in experiments from Study 2, participants' reasoning experience was much more nuanced than captured by DPT models of reasoning. The fixed temporal order and the limited number of preference revisions proposed by these models fail to account for the diverse experiential data presented here.

TABLE 5.2: Bayesian hierarchical models comparing the Low-C condition to High-C and impersonal on (a) final choice reported, (b) switches in preferences, (c) conflict ratings, and (d) confidence ratings provided at the end of each trial. Participants and items were treated as random effects, and standard deviations (sd) of the slopes are reported in each model. CrI stands for the 95% credible interval. $\hat{R}$ is the Gelman-Rubin diagnostic value. $\hat{R}$ more than 1 indicates that chains have not converged.

(a) Choice ∼ dilemma type

	Estimate	Estimate error	CrI	$\hat{R}$
Intercept	-2.93	0.62	[-4.25, -1.81]	1
High-C	3.41	0.69	[2.09, 4.81]	1
Impersonal	2.44	0.67	[1.15, 3.79]	1
sd(Participant)	0.77	0.26	[0.27, 1.32]	1
sd(Item)	1.27	0.44	[0.63, 2.35]	1

(b) Switches ∼ dilemma type

	Estimate	Estimate error	CrI	$\hat{R}$
Intercept	0.81	0.25	[0.31, 1.28]	1
High-C	0.97	0.05	[0.87, 1.07]	1
Impersonal	0.51	0.05	[0.41, 0.61]	1
sd(Participant)	1.06	0.18	[0.76, 1.48]	1
sd(Item)	0.33	0.15	[0.06, 0.66]	1

(c) Conflict ∼ switches

	Estimate	Estimate error	CrI	$\hat{R}$
Intercept	2.19	0.15	[1.89, 2.49]	1
Switches	0.29	0.03	[0.22, 0.35]	1
sd(Participant)	0.30	11	[0.09, 0.52]	1
sd(Item)	0.65	0.17	[0.39, 1.06]	1

(d) Confidence ∼ switches

	Estimate	Estimate error	CrI	$\hat{R}$
Intercept	3.87	0.14	[3.60, 4.14]	1
Switches	-0.22	0.03	[-0.28, -0.15]	1
sd(Participant)	0.30	0.09	[0.11, 0.49]	1
sd(Item)	0.54	0.15	[0.33, 0.90]	1

Fixation duration and pupil size For the eye-tracking part, we analyzed the rolling-window averaged fixation duration and pupil size recorded while participants were deliberating. The interest period was divided into blocks to characterize different parts of the reasoning process as depicted in Figure 5.1. Participants' first key press in the deliberation phase was taken as the initial preference. The period before the first preference is recorded is likely spent in reading the displayed text on the screen (although participants were likely to be reasoning along with reading). The data following this period was chunked in blocks that identified whether or not there was a change in preference. In no-switch block, participants ostensibly stay with their earlier stated preference. Contrarily, the switch blocks identified a change in the reported preference. We wanted to explore how participants re-engaged in the problem marked by the presence of switch and no-switch blocks.

Since pupil size cannot be averaged across participants as it is recorded in arbitrary units, it is usually compared relative to a baseline. Defining a constant baseline period across all trials when participants were less likely to be conflicted was challenging in our setup. One possible window for the baseline would have been in-between trials before participants saw any text. However, we could not use this period as the trials were self-paced and participants rarely waited long enough to begin a new trial. Another option was to use the beginning of the deliberation phase when participants first saw the text of the problem which would be the beginning of the before key-press block in a trial. Although participants more or less immediately started reading the dilemma, they likely began reasoning about the problem right away, too. If the baseline period was defined too short, it would not characterize a baseline activity satisfactorily. On the other hand, a longer period at the beginning of the trial could not ensure that participants did not experience conflict as they read more of the text, already anticipating contrasting motivations behind the two alternatives. Therefore, instead of confining the baseline to a specific part of the deliberation, we normalized pupil data recorded from a participant. This effectively identified pupil constriction and dilation relative to the average pupil size recorded for a participant.

We also restricted our eye-tracking analyses to moral dilemmas. The overarching goal of this study spanning two experiments in this study was to compare different reasoning styles, one when people have a strategy or method to solve a problem like syllogisms and when people did not necessarily have one (like in moral dilemmas). Non-moral problems could not be clearly categorized as either since some of the non-moral stimuli included strategic choosing (resolving scheduling conflicts) while others did not (choosing between a scenic and a fast route to drive to a destination). Hence, we decided to analyze fixation

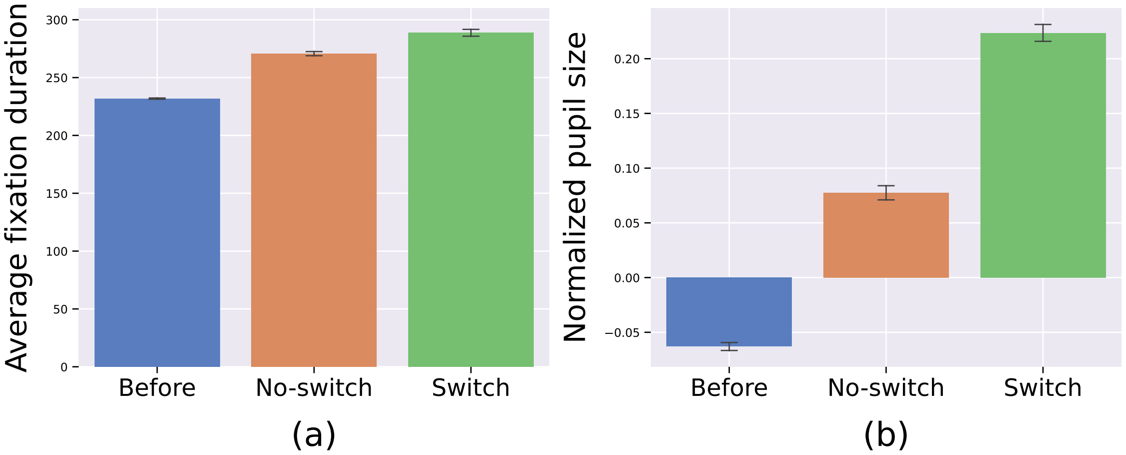


FIGURE 5.3: Blockwise mean and bootstrapped 95% CIs for (a) fixation duration and (b) standardized pupil size by preference blocks in Experiment 1 of Study 3.

duration and pupil size recorded only in moral trials. The descriptive statistics for all conditions including non-moral problems is provided in the Appendix Table A.3.

Fixation duration and pupil size were lowest before a preference was recorded in a trial and highest in the switch block (see Figure 5.3). We compared fixation duration and pupil size in the before key-press block (blue bars in the plots) to the no-switch and switch blocks using LME models with participants as the random factor. Before key-press block was the reference level. Participants' fixations were longer in no-switch and switch blocks than the before key-press block by 54 and 69 ms, respectively. Pupils were also more dilated in these blocks than the before key-press block (see Table 5.3). Further, the difference between no-switch and switch blocks on fixation duration ($\beta_{difference} = 5.161, SE = 1.03, t = 2.54, p = .01$) and pupil size ($\beta_{difference} = 0.0891, SE = 0.01, t = 14.10, p < .001$) was also significant, with switch block recording longer fixation duration and larger pupil size.

Before a key is pressed, participants are likely reasoning along with reading the dilemma. Hence, interpreting the significant difference between before key-press and other two blocks is potentially confounded by the effect of reading on fixations and pupil dilation. Yet, both no-switch and switch blocks showed the same trend on these two measures, suggesting that participants might still be monitoring conflict between alternatives. Given that these measures have been used as indicators of conflict detection and monitoring, participants may be re-engaging in the moral dilemma in a similar way, regardless of the ultimate preference recorded at the end of these interim deliberations.

Table 5.3: LME models of (a) fixation duration and (b) pupil size comparing preference blocks (before key-press, no-switch and switch) in Experiment 1 of Study 3. Participants are treated as the random effect and the predictor is dummy coded with before key-press as the reference level.

(a) Fixation duration ∼ preference block

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	229.7760	7.31	27.04	31.44 ***	Participant	1489
No-switch	61.4330	1.02	59663.12	60.06 ***	Residual	8191
Switch	67.1380	1.179	59658.906	56.93 ***		

(b) Standardized pupil size ∼ preference block

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.0662	0.0136	24.34	4.87 ***	Participant	0.0045
No-switch	0.2086	0.0039	447000	54.20 ***	Residual	1.0515
Switch	0.3117	0.0046	440200	69.94 ***		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

These results, although exploratory, lend support to our initial hunch that when reasoning about moral dilemmas, people may get swayed by thoughts and inclinations that are not necessarily considered in a strategic or orderly manner. The information used while reasoning on such problems may be reconsidered afresh, regardless of what alternative is ultimately favored at the end of such deliberations. Thus, even if participants settle on the same choice after re-engaging in the problem, like in the no-switch block, the process of reasoning may be similar to that which occurs when a preference is changed. A model similar to this is proposed earlier by Gürçay and Baron (2017) in the context of moral reasoning [82]. They argue that when we reason, we consider arguments supporting either of the choices in arbitrary fashion. If this is the case, it should be reflected in similar trends in physiological markers in both switch and no-switch blocks.

Next, we compare these exploratory findings from reasoning through dilemmas with minimal formal training to the conflict signatures observed in logical problem-solving exercises, where participants come to the lab already trained in strategies used to solve them.

5.2 Experiment 2

For the final experiment in the thesis, we investigated how the eye movements while solving syllogisms compare to moral problems. The characteristic difference between these two reasoning problems is between solving and deciding. Syllogisms can be solved employing different strategies that one can be trained to use such as Venn diagrams and logic tables. There is also a set standard, which is the logical validity of the syllogism, against which the answer can be evaluated. However, one cannot—and typically does not—apply strategies to *solve* moral issues. Solving involves applying a set of rules to arrive at an answer, whereas when we reason about moral problems, we are often swayed by momentary preferences, contextual information, and even framing effects [114, 147]. Therefore, applying a particular strategy may not be prudent. In effect, moral deliberations may bring conflicting motivations continually to the forefront of our minds. On the other hand, when we attempt to solve a syllogism using a learned strategy, we apply a known rule to the problem. Unless the strategy seems to fail, we may not look for other ways to solve the problem, and hence, may not experience conflict till that point. Therefore, we anticipated that conflict will be more discrete in these problems, occurring *only* when the current strategy is failing and there is a need to switch away from it.

5.2.1 Method

Participants

Previously in Experiment 3 of Study 2 with syllogisms, participants switched less frequently than with moral dilemmas. Since the goal of tracking eye movements while reasoning was to examine how individuals re-engage with problems differently when reasoning about moral versus logical problems, it was necessary to have a comparable number of switches in both experiments to effectively contrast the conflict signatures in eye-tracking data. Hence, we recruited a larger group for Experiment 2 in this study, totaling 70 participants before exclusions, through email advertisements and word-of-mouth. We collected data from only those participants who had not participated in Experiment 3 of Study 2 earlier. We excluded data from 5 participants whose eyes had to be re-calibrated more than twice after commencing the experiment and from 3 participants data was not saved accurately due to a coding error. The final sample for analysis consisted of 62 participants (18 females; Mean age = 21.80 years). All participants were compensated with Rs. 100/- for their time. The IEC approved the study design.

Materials

Participants solved 8 syllogisms of either single- or multiple-model type. Validity and believability of syllogisms were fully crossed. The syllogisms used in Experiment 2 were exactly the same as Experiment 3 of Study 2 (refer to the Appendix for the stimuli).

Procedure

This experiment employed the Switch paradigm with syllogisms. The procedure for the experiment remained unchanged from Experiment 3 of Study 2 apart from the additional calibration tasks included for eye-tracking. Participants eyes were (monocularly) tracked at 1000 Hz frame rate on the same system (1920 x 1080 pixel) and with the same eye-tracker (EyeLink 1000 plus, SR Research Ltd.) as Experiment 1 above. They were seated approximately 65 cm away from the screen. They completed two sets of 9-point calibration and validation task with their choice of wearing spectacles or not, beginning with the left eye. The task was repeated with the right eye if the tolerance was below 2°. The rest of the protocol for tracking eye movements was exactly the same as in Experiment 1. Participants' head was stabilized with a chin-rest and they were asked to minimize excessive movements. Eye movements were corrected for drift before each trial (tolerance level was 2°). Calibration and validation were repeated once if drift exceeded this limit. If the eyes still did not calibrate, then the experiment was halted.

Preprocessing

The same procedure from Experiment 1 of the current study was followed in Experiment 2 for preprocessing and dividing fixation duration and pupil size data into before key-press, no-switch, and switch blocks.

5.2.2 Results and discussion

Choice data and vacillations Since all of our participants were students in a technical institute, we were expecting that they would be familiar with these problems. Indeed, only 7 participants out of 62 were unfamiliar with syllogisms. At least 20 people reported explicitly that they used Venn diagrams, with an additional few using sets and their intersections to map the premises. Participants familiarity serves our purpose for this

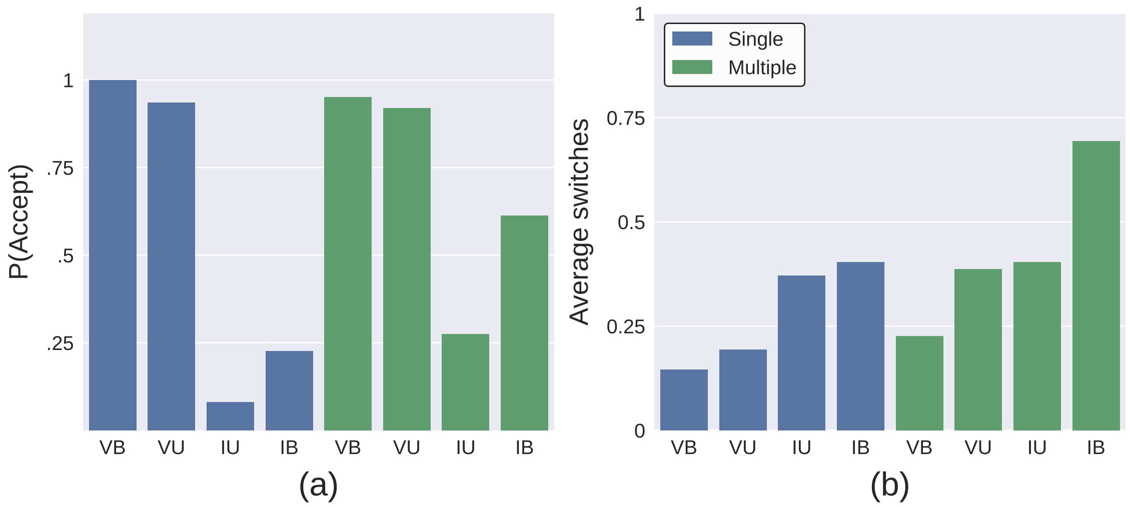


Figure 5.4: Bar plots of (a) the proportions of trials accepted as logically valid and (b) average switches recorded in each condition in Experiment 2 of Study 3. Bars are color coded by model-type (Acronyms for conditions: VB = valid-believable, VU = valid-unbelievable, IB = invalid-believable, and IU = invalid-unbelievable.

experiment which was to investigate eye movements when people know how to navigate them.

We ran logistic regression models separately in single- and multiple-model syllogisms with the interaction between validity and believability as predictors (results presented in Table 5.4). The main effects for validity of the syllogism problem and believability of the conclusion were significant in both models. Hence, believability of the conclusion did impact choice proportions regardless of participants knowing how to solve them. However, unlike Experiment 3 of Study 2, which employed the same stimuli as the current experiment, the interaction between these two effects was not significant in multiple-model as well as single-model syllogisms. This stands in contrast with the classic finding that people tend to accept believable rather than unbelievable conclusions more on invalid syllogisms when compared to valid syllogisms. However, this particular interaction was not under scrutiny for this experiment. The interaction is of interest to test and compare different models of logical reasoning. But we have already demonstrated in Experiment 3 from Study 2 that vacillations offer a more direct way to test these models than simply looking at the choice data. Further, we were mainly interested in investigating the reliability of the pattern of switches across conditions when compared to Experiment 3 of Study 2, in addition to exploring the possibility of a signature of conflict that can be reliably detected in our data. Hence, although our data does not replicate the classical finding, we were interested in comparing these data with behavioral data reported earlier in Experiment 3 of Study 2

Table 5.4: Analysis of variance tables for the logistic regression models for single-model and multiple-model syllogisms from Experiment 2 of Study 3.

Single-model				**Multiple-model**		
Effect	**DF**	**F**	**p**	**DF**	**F**	**p**
Validity	1	204.77	< .001 ***	1	81.76	< .001 ***
Believability	1	8.93	.003 **	1	14.22	< .001 ***
Interaction	1	1.96	0.16	1	1.04	0.31

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

and eye-tracking data with Experiment 1 of Study 3 [1].

Premises in single-model syllogisms can be arranged in only one specific way. On the other hand, to decide the validity of the multiple-model syllogisms, the conclusion has to be valid in *all* possible arrangements of the premises. Hence, in single-model syllogisms, we did not expect participants to switch in any particular condition more than others. To compare switches in all conditions and models, we ran two separate LME models. Invalid and unbelievable were coded as reference levels with participants as the random effects. Our hypothesis was validated as none of the conditions in these problems recorded significantly different number of switches from the baseline condition (invalid-unbelievable; see detailed results in Table 5.5). In contrast, participants switched more in the invalid-believable condition than in the invalid-unbelievable condition (which also recorded more switches than any other condition). Participants also switched less in the valid-believable condition, which had the fewest switches among the multiple-model conditions (see Figure 5.4 (b)). They switched between alternatives more often in multiple than single-model syllogisms (One-sided unequal variance two independent sample t test: $t(486.83) = 2.32, p = 0.02$). Lastly, the number of switches correlated with these ratings in the expected manner. Every switch increased the average conflict rating by 0.55 points while reduced the confidence rating by 0.41 points (see Appendix Table A.4 for detailed results).

We also inspected the first key pressed during the deliberation phase while solving syllogisms. In conditions valid-believable and invalid-unbelievable, analysis of the syllogisms by their logical validity or the believability of the conclusion cues the same response.

[1] Another caveat in these analyses is that we compared the choice data in Experiment 3 of Study 2 in different conditions using ANOVA following the classical studies of logical reasoning. Although this approach is common in the field, since the data being modeled are binomially distributed, generalized models such as logistic regression are more appropriate tests for these data. When logistic regression is applied to the Experiment 3 data from Study 2, the interaction between validity and believability is non-significant, even in multiple-model syllogisms, whereas this interaction is significant in the ANOVA results. These alternative models are provided in the Appendix Table A.5.

Table 5.5: LME models of switches recorded in single and multiple-models in Experiment 2 of Study 3. The predictors are treatment coded with invalid and unbelievable as reference levels such that the intercept is the condition invalid-unbelievable. Participants were introduced as random effects in the model

(a) Single-model

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.3710	0.09	238.93	4.38 ***	Participant	0.0375
Valid	-0.1774	0.12	183	1.55	Residual	0.4082
Believable	0.0322	0.12	183	0.28		
Interaction	-0.0807	0.16	183	0.5		

(b) Multiple-model

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	0.4032	0.1	243.96	4.26 ***	Participant	0.0043
Valid	-0.0161	0.13	183	0.12	Residual	0.5516
Believable	0.2903	0.13	183	2.18 *		
Interaction	-0.4552	0.19	183	2.39 *		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Participants’ pressed this cued response 55% of the times out of total 248 trials (One proportion z test compared to proportion 0.5: $\chi^2(1) = 243.82, p < .001$). When the logical analysis and prior beliefs both cue different responses such as in valid-unbelievable and invalid-believable, then in only ~40% of 248 trials participants responded according to logic (One proportion z test: $\chi^2(1) = 244.43, p < .001$). These results were in contrast with Experiment 3 of Study 2 and also with the hybrid model of DPT which proposes that logically valid response can be accessed by the quick System 1 [35]. Contrarily, others have proposed that the order in which people consider information and build a preference is not fixed. People change mind while thinking, and hence, the order in which preferences are formed do not need to have a set pattern. However, these results suggest that even for problems which people know how to solve, the preferences are not recorded in the expected order. We discuss the point of temporal order in preferences in light of the results from all experiments in the concluding chapter.

Experiment 2 of the current study replicated all major findings of Experiment 3 of Study 2 with the same syllogisms. People switched more often in invalid-believable than any other

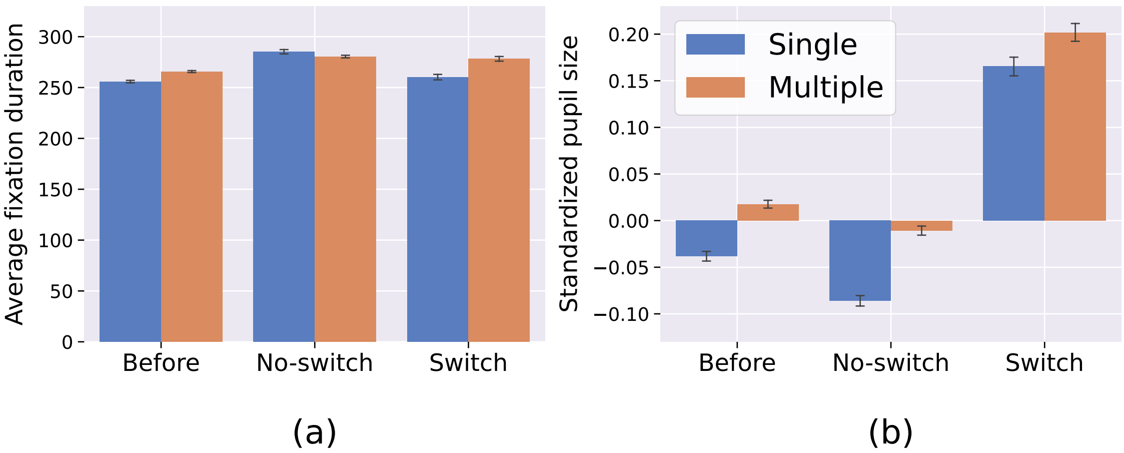


FIGURE 5.5: Blockwise mean and 95% CIs for (a) fixation duration and (b) normalized pupil size data from Experiment 2 of Study 3. Bars are color coded according to the model-type of syllogisms. CIs were calculated through bootstrapping over 1000 samples.

trials. This effect was observed only when the syllogism was of the multiple-model type. Although the misinterpreted necessity model can be extended to hypothesize that people should switch more in multiple syllogism of the invalid kind, they do so particularly more in the believable than unbelievable problems. This finding also goes against the proposal that people build mental models to test for validity of valid-believable syllogisms as these conditions recorded the lowest number of switches. In other words, Experiment 2 replicated the typical response pattern observed in the choice data, but the vacillations revealed a pattern not explained by existing models, similar to what was observed previously in Study 2.

Fixation duration and pupil size The general trend so far suggests that most participants solve the single-model syllogisms accurately and switch less frequently. Multiple-model syllogisms, on the contrary, evoke more frequent vacillations, particularly in invalid-believable syllogisms. Participants' accuracy is also lower with a stronger effect of prior beliefs on validity judgments. In Experiment 2, we recorded participants' fixation duration and pupil data while they solved these syllogisms with the aim of investigating how participants re-visit the problem after forming and reporting an initial preference. As we have mentioned already, most participants were aware of how syllogisms are solved and had strategies to employ. We wanted to explore if the eye behaviors reflected the strategic solving. Switch blocks were interesting in that they occurred when participants changed their choice despite presumably solving through the problem by using the said strategies. Therefore, switch blocks were peculiar cases where a current choice of alternative was in conflict with the preceding recorded choice. Since pupil dilation and longer fixations are signs of conflict detection and monitoring as well as attentive processing of information, we hypothesized that switch block will record the bigger pupil size and longer fixation duration on an average. On the contrary, participants do not necessarily encounter such conflict when they stick to the same choice in no-switch block. Hence, if pupil size and fixation duration are sensitive to the conflict in current and previous preference, then no-switch blocks should not show the same trend in these measures as switch blocks.

Results were mixed for fixation duration. Average duration of fixations in no-switch blocks were longer than before key-press and switch blocks. In single-model syllogisms, the difference was more pronounced than in multiple-model syllogisms, with close to 38 ms and 14 ms gain in duration over before key-press and switch blocks, respectively (see LME results in Table 5.6). In multiple-models, although the difference was significant, the difference was small (19 ms and 3 ms when compared with before key-press and switch blocks, respectively).

Pupils were more sensitive to the difference in switch and no-switch blocks. Switch blocks recorded the highest pupil size in both single and multiple-model syllogisms (Figure 5.5 (b). For exact estimates, see Appendix Table A.6). The difference between the switch and no-switch blocks was also significant (results of the LME model are in Table 5.6). Although the no-switch block recorded significantly smaller pupil size than before key-press block too, we did not have a specific hypothesis for this difference because in before key-press participants see the problem for the first time. On the other hand, the marked difference between switch block and no-switch block lends support to our earlier stated hypothesis. We discuss these results and results from Experiment 1 of this study together below.

5.3 Discussion

Our investigation in this study centres on the phenomenological experience of reasoning when we know how to reason and when we do not. This difference can be appropriately captured by contrasting the activity of solving with deciding. When we solve for a solution, like navigating a maze puzzle, we try a strategy till it appears to fail according to some set standard or the objective of the problem. In a maze puzzle, it is hitting a wall while in a logical reasoning task, it is realizing that the current strategy ends up in an illogical conclusion. Such standards, however, are not so apparent or deducible in moral problems. When one must choose the action that kills a few but saves many with the other choice

Table 5.6: LME models examining fixation duration and pupil size across preference blocks in single-model syllogisms (a, b) and multiple-model syllogisms (c, d) from Experiment 2 of Study 3. Participants were modeled as a random effect. The predictor, preference block, was dummy coded with no-switch block as the reference level.

SINGLE-MODEL

(a) Fixation duration ∼ preference block

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	292.125	6.06	61.55	48.19 ***	Participant	2224
Before key-press	-38.735	1.26	37760.41	30.67 ***	Residual	11111
Switch	-14.267	1.99	37765.53	7.18 ***		

(b) Standardized pupil size ∼ preference block

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	-0.0973	0.02	61.91	4.27 ***	Participant	0.0317
Before key-press	0.0702	0.01	311500	17.06 ***	Residual	0.9821
Switch	0.2958	0.01	313800	45.04 ***		

MULTIPLE-MODEL

(c) Fixation duration ∼ preference block

Fixed effects					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	285.7544	5.40	61.92	52.92 ***	Participant	1787
Before key-press	-19.0972	0.78	46663.84	24.36 ***	Residual	5044
Switch	-2.7338	1.12	46657.60	2.44 ***		

(b) Standardized pupil size ∼ preference block

Fixed effect					**Random effects**	
	Estimate	SE	df	t		Variance
Intercept	-0.0055	0.01	62.73	0.31	Participant	0.0189
Before key-press	0.0160	0.01	383500	4.27 ***	Residual	0.9665
Switch	0.2272	0.01	386300	42.86 ***		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 being letting a larger group die, the correctness of this choice may be difficult to decide. In such tasks there is no option to defer the decision to a later time too, since most of these problem are framed to be time-sensitive (the trolley is already hurtling down the tracks at the time of the choice) and not choosing the action is conflated with deferring the choice.

Therefore, moral dilemmas like the Trolley problem seem to be in a stark contrast with how we solve syllogisms. Owing to these differences in the way these tasks are approached, we hypothesized that the re-engagement in the problem after an initial preference is formed will be different. A decision about the validity of the syllogism is reached after following certain steps that are often formally learned (which was the case with most participants in our experiment). Revisiting the problem is likely to result in the same answer as before. However, if reconsideration of the same problem leads to a different choice, it is unexpected. Conversely, if the initial choice itslef is uncertain, like in a moral deliberation, subsequent switches are perhaps more predictable. Vacillations enabled us to identify the periods of re-engagement in the problem when it lead to the same or different preference. In case of moral dilemmas, we expected the re-engagements to show the same trend regardless of whether there was a preference shift due to continually experienced conflict even when re-engaging in the problem. In logical tasks, however, we anticipated that participants would experience conflict primarily when re-engaging with the problem just before switching their judgment in logical reasoning tasks, whereas in moral reasoning tasks, conflict was expected regardless of whether the re-engagement resulted in a preference shift.

Pupil size estimates followed this predicted trend. Participants' eyes were dilated more when they re-engaged in the problem than before recording any preference. While deliberating a moral dilemma, this trend was observed regardless of the succeeding preference aligning with the current preference. When solving syllogisms, however, when participants switched their answer, the pupils were dilated more. When they stayed with their choice, the pupils, in fact, constricted. Average fixation duration in switch and no-switch blocks were higher by more than 50 ms than the before key-press only in Experiment 1. In Experiment 2, we had expected average duration to be longer but only in the switch block. This trend was not observed. Together, our exploratory results suggest that pupil data might be more sensitive to the differences in reasoning when one vacillates while reasoning.

Fixation duration and pupil size have been relatively underutilized in reasoning studies, and when they have been used, it is often in conjunction with areas of interest (AoIs) on the screen. AoIs are predefined regions within the visual display that mark the locations of significant content, such as choice alternatives or trade-offs in the problem. By using AoIs, researchers can gain insights into gaze patterns, such as how often participants switch their focus between different options or how long they spend fixating on each choice. This allows for a more nuanced understanding of attention and cognitive processing during decision-making tasks.

However, implementing a location-specific analysis, as done in studies like Ball et al. (2006) [7] and Purcell et al. (2023) [135], was challenging in this study due to the unique structure of our experimental design. Specifically, participants were required to deliberate on each moral dilemma for at least one minute. Given the extended deliberation period, it was unrealistic to assume that participants would maintain consistent gaze patterns, particularly directed at predefined AoIs, throughout the duration of their thinking. This is compounded by individual differences in how participants approach the task—some may prefer to scan the problem in its entirety, while others might focus only on specific aspects. Additionally, because the task required extensive reading and reflective thought, it is unlikely that all participants would gaze exclusively at the areas marked as AoIs after reading the problem, which could introduce significant variability in gaze behavior.

In syllogistic reasoning tasks, such as those used by Ball et al. (2006), defining AoIs is relatively straightforward because the content (premises and conclusions) is clearly delineated on the screen. However, in moral dilemmas, the situation becomes more complex. It was difficult to determine which words or phrases in the problem would be universally relevant for all participants. Moral dilemmas often involve nuanced language and subjective interpretations, meaning that what one participant might consider the most critical aspect of the problem could differ from another's perspective. This subjectivity in relevance makes it challenging to define AoIs that would be consistent and meaningful across the sample.

Another complication in our study was that the choices in the dilemmas were mapped directly to the response keys, meaning that participants could press the keys without necessarily fixating on the screen, particularly once they had processed the problem. This further limited the utility of AoIs as a measure of cognitive engagement with the problem.

For these reasons, we opted to focus our analysis on more general physiological measures, such as fixation duration and pupil size, which are indirect but still informative indicators of cognitive processes. These measures provided a more holistic view of participants' cognitive states during the reasoning process, without being influenced by the potential complications associated with AoI-based gaze analysis.

Looking forward, future research could expand upon these findings by incorporating AoIs within the Switch paradigm, potentially modifying the experimental design to eliminate the one-minute mandatory deliberation period. This would allow for more precise measurements of gaze behavior without the confounding factor of prolonged reflection. Furthermore, validating this approach could involve correlating eye-tracking data with real-time conflict ratings, enabling a more dynamic understanding of how cognitive conflict unfolds during moral reasoning. One promising method could involve using a joystick to track shifts in preferences as participants engage with the dilemmas, providing real-time insights into how their cognitive processes evolve throughout the task. Such studies could deepen our understanding of the interplay between attention, conflict, and decision-making, shedding light on the nuanced ways in which we reason under varying degrees of cognitive conflict.

5.4 Study 3 in Review

In conclusion, our study examines the differences in reasoning between logical and moral dilemmas, highlighting how the experience of conflict varies across tasks. We show that conflict is task-dependent and that its physiological markers can be tracked in real time with minimal interference. Our findings indicate that moral dilemmas evoke sustained conflict throughout deliberation, whereas logical tasks like syllogisms tend to elicit conflict mainly during the reconsideration of initial choices. These results demonstrate the sensitivity of pupil data in detecting cognitive conflict and shifts. By analyzing fixation and pupil metrics broadly, we capture key aspects of the deliberative process. Future research can build on these exploratory findings by integrating AoIs and other real-time conflict monitoring tools to deepen our understanding of the cognitive processes underlying reasoning and decision-making.

Chapter 6

General Discussion

Reasoning is complex. Our habits, beliefs, motivations intermingle with countless contextual factors to produce a decision, sometimes a part of a pattern while other times fitting no mould. We consider and reconsider information, vacillate between our options, be inconsistent and indecisive in our choices. This rich diversity in the experience of reasoning is seldom fully captured by existing theoretical frameworks. The tools used to study these processes have, in turn, limited the exploration of these frameworks, while the frameworks themselves have constrained our understanding of the processes supporting reasoning's variability. In this thesis, we hoped to make a case for adopting more refined measurement tools that better capture the fluid nature of reasoning and decision-making. Below we review key empirical findings within the context of broader reasoning theories, incorporating alternative perspectives in reasoning research. We conclude by exploring how the study of conflict can be taken forward to build a more comprehensive and nuanced picture of how we reason.

6.1 Critique of two-process models of reasoning

It would not be an understatement to say that the dual-process theory has dominated reasoning and decision-making research since the 1970s. The basic idea behind this theory is that fast and slow decisions are qualitatively different and thus supported by distinct and dissociable mechanisms. System 1 is quick to operate on information based on intuitions and learned associations, while System 2 supports a deliberative effort to generate an answer. Over the years, the two systems have been contrasted using opposing labels such as automatic-controlled, effortless-effortful, fast-slow, unconscious-conscious, etc., and different mechanisms of their interaction have been proposed. Although there is not enough evidence to support that characteristics under the same system concur invariably, the distinction between the two systems has remained popular in the field. Newer mechanisms along with adjustments to the old models are continually proposed not only to accommodate new data but also to spearhead investigations assuming that a systematic distinction between the two systems exists (for a recent review and criticisms, see De Neys (2023) [32] and commentaries on the paper). Below, we outline three criticisms of dual-process models that are highlighted by the evidence presented in this thesis.

The default-interventionist mechanism posits a serial processing architecture for System 1 and System 2. Some versions of this model also suggest that specific responses are exclusively generated by one of the two systems. In the moral domain, the deontological or rule-based response (which coincides with not endorsing the action in a typical sacrificial dilemma) is associated with System 1 due to emotions' influence. Personal dilemmas are thought to engage System 1 preferentially by emphasizing affect-laden information, leading people to prefer deontological choices. On the other hand, impersonal dilemmas cue the calculative System 2 by highlighting the utility-maximizing consequence of endorsing the action, and hence, support the utilitarian response (as suggested in early versions of Greene and colleagues' model [80, 81]). By presupposing a temporal order between the fast and slow systems and response exclusivity, this model predicts that a deontological response is updated to utilitarian judgments if System 2 kicks in. In other words, DU preference updates are expected, but not UD. Here, the first and second letters refer to the initial and final preferences reported during deliberation. DU signifies that an initial preference for the deontological option shifted to the utilitarian option in the final choice, while UD indicates the reverse—a shift from utilitarian to deontological preference. Moreover, the utilitarian response is not expected to precede the deontological response, as System 2 only updates System 1's intuitive response.

However, across all three studies presented in this thesis, we failed to replicate this pattern of responding. In the mouse-tracking experiment with sacrificial dilemmas (Experiment 1 in Study 1), this account would predict that when individuals choose the atypical utilitarian response in personal dilemmas, their response trajectories should be curvier due to the strong influence of the stereotypical deontological response. However, the results showed that the trajectories in atypical trials were not more curved than those in typical response trajectories. Switches in preference offered a more direct evidence against such response exclusivity. Experiments from Study 2 and 3 demonstrated that preferences are updated continually when we reason, with no specific order. The DU preference updates were as common as—in some cases even less common than—UD preferences (see Study 2 and Study 3).

Some models of logical reasoning that propose an explanation for the belief bias effect also posit a specific temporal order in judgments when solving categorical syllogisms [49]. The selective scrutiny model, which is essentially a default-interventionist model, posits that the default response is belief-based and is updated by System 2 if the conclusion is unbelievable. Other models not only predict the sequence of judgments but they can also be extended to get predictions about the types of syllogisms that would lead to more in-depth solving strategies, thereby allowing more opportunities for people to change their preference. According to the misinterpreted necessity model, people should give the logical response first and vacillate more in invalid trials of multiple-model syllogisms. The mental models theory gives precedence to logical reasoning as well, with more vacillations expected in valid-unbelievable multiple-model syllogisms (see Section 2.2 for detailed descriptions of these models). However, in our experiments, these predictions did not hold consistently. Our initial investigations into belief bias with the Switch paradigm in Experiment 3 of Study 2 showed that participants' first choices aligned with logical validity more often than chance. But with a larger sample size, this pattern was not replicated. Instead, in Experiment 2 of Study 3, initial preferences were in favor of the logical response less than chance. Furthermore, vacillations were most frequent in invalid-believable trials, a pattern not predicted by any existing models (see Figure 5.4). In short, upon closer examination, the response exclusivity and predictions about the depth of reasoning did not reflect clearly in our empirical results.

The assumption of response exclusivity in dual-process models has recently come under theoretical criticism, even from proponents of the general framework, most notably De Neys [31, 32]. De Neys proposes an alternative hybrid default-interventionist model, where System 1 can generate intuitions typically attributed to System 2 [6, 33, 35]. That is, System 1 can produce both deontological and utilitarian inclinations in moral and belief- and logic-based intuitions in logical reasoning tasks. The potency or strength of these intuitions might differ from the outset and may change over time, too. System 1 simply keeps track of these activations and cues System 2 when it cannot decisively produce a response. Deliberations take place only when System 2 is engaged. System 2 can operate through different functions. It can generate a new response. It can also assess the changing strength of intuition activations and manipulates them further by deliberating actively until one of the responses emerges victorious or reasoning is abandoned. System 2 emloys additional resources like attentional control and working memory to aid deliberations.

De Neys's proposal avoids the issues posed by response exclusivity but, at the same time, it fails to delineate System 2's contribution to the process of reasoning. For instance, Bago and De Neys (2019) claimed that in some cases people's early preferences do not get updated until the end, demonstrating that early and late preferences in some trials were the same [6]. However, using the same dilemmas with the Switch paradigm (see Experiment 2 of Study 2), we showed that preferences often undergo multiple updates even in a one-shot reasoning task. Participants vacillated between alternatives multiple times before settling on a choice. Although De Neys's hybrid theory can accommodate this result (as System 2 deliberations might favor different alternatives over time), other models which are simpler could explain it as well. We will revisit this issue in more detail in the next section.

Lastly, we have argued in this thesis that the two-systems framework has significantly constrained the study of reasoning. The measures and paradigms used to investigate reasoning are often designed to test specific predictions of these theories. In doing so, they either rely on limited data, such as trial-level summaries, or employ paradigms that substantially distort the processes under investigation. For example, consider the two-response paradigm originally proposed by Thompson, Turner, and Pennycook [168]. Bago and De Neys used this paradigm (2019) to demonstrate that utilitarian judgments, which are traditionally attributed to System 2, are also cued by System 1 [6]. They employed both time pressure and cognitive load to limit the processing of information by 'knocking off' System 2. Participants were briefly shown the positions of four dots in a 3x3 grid before each trial. Each trial required them to read a dilemma and respond within 12 seconds while simultaneously memorizing the dot locations. After producing an initial response, participants were asked to recall the grid's position from four alternatives, followed by a final decision phase where they could deliberate and respond without a time constraint. The authors concluded that some participants who had a predilection toward an alternative did not update their choice until the end. But the paradigm itself only offered a peek (or two) into the process. By tracking preferences more comprehensively, we demonstrated that people's reasoning is more nuanced than trial-level summaries or measurements confined to specific portions of the process.

In summary, this thesis adds to the growing body of evidence suggesting that response exclusivity under two-system models lacks empirical support in moral and logical reasoning [6, 13, 14, 33, 35, 82]. Further, while the popular measures and paradigms of reasoning have helped advance dual-process theorizing, they have either limited the measurements to specific periods or relegated them to the end. By treating different parts of the reasoning process as partially independent, studies that attempt to isolate these mechanisms——such as those using cognitive load paradigms——risk disrupting the very processes they aim to understand. Alternatively, if the assumption of two separate systems is abandoned, it might be more effective to track reasoning processes concurrently rather than confining the investigation to early or late stages. This approach would allow reasoning to be studied under more ecologically valid conditions, providing a more accurate depiction of how individuals engage in decision-making in real-world scenarios.

6.2 Rethinking conflict in reasoning

The central claim of dual-process theories is that conflict detection decisively changes the mechanism underlying reasoning. If two equally strong responses are cued—whether by the same or different systems—then the analytical operations of System 2 intervene to resolve this conflict. Before conflict is detected, information is processed quickly, effortlessly, automatically, unconsciously, etc., as opposed to how it is processed after detecting conflict, presumably slowly, effortfully, under cognitive control, consciously, etc. However, these distinctions are insufficient to dichotomize the underlying mechanisms in reasoning.

Firstly, there is little empirical support that these properties concur when processing information [9, 68, 112, 126, 182]. Additionally, these labels represent points on a continuum rather than distinct categories. For example, on the effort scale, a strong intuitive response may lie at one extreme, indicating that little effort was expended, while effortful reasoning may lie at the other end. But there is no break postulated between these two extremes. De Neys argues that the divide is of the qualitative kind (no effort as opposed to a lot of effort) but drawing any boundary on a continuum would be arbitrary and can be contested. Hence, a mere qualitative difference in processing does not imply that the processes toward one end are different from those at the other end. Put differently, distinct qualitative properties of reasoning do not mean that the algorithms supporting those processes are distinct as well (this point has recently been argued by Dewey (2021) [36]).

Alternatives to the dual-process view, though few and not widely applied across reasoning contexts, do exist. Gürçay and Baron (2017) proposed a conflict-tracking model in moral reasoning [82]. Their model rests on three assumptions: first, that moral principles, both deontological and utilitarian, are available to reason from the outset and are in conflict in sacrificial dilemmas. Second, conflict resolution can come under the influence of problem-specific and person-specific attributes. For instance, personal dilemmas may cue a deontological response more readily than a utilitarian response. Similarly, individual differences in preferred ethical principles can influence the decision. And lastly, responses cued by different ethical principles may have different activation strengths. When the activation strengths are comparable, conflict will ensue as two responses compete to control the ultimate response. Here, the sequence of consideration is inconsequential, as multiple factors will differ in their favored outcome and thereby sway the preference. This account of reasoning closely resembles De Neys' hybrid dual-process view. However, Gürçay and Baron do not postulate that the reasoning following the detection of conflict will be supported by a different algorithm. Instead, their account aligns more closely with a stochastic evidence accumulation model, which continually gathers information until the difference between accumulated evidence in favor of alternatives is sufficient to halt reasoning and produce a response.

Single-process models, like Gürçay and Baron's conflict model, offer a more straightforward explanation of decision-making [24, 158]. These models assume that reasoning is governed by a single algorithm, regardless of how quickly or slowly the reasoning occurs. Model parameters can be adjusted to account for individual biases and context-specific influences, as in the conflict model mentioned earlier. For example, if people are naturally biased against harming their own relatives in sacrificial dilemmas, the model can incorporate this bias by lowering the threshold for the deontological response.

The results of our investigation into conflict can also be interpreted through a single-process model, as outlined above. The interim preferences recorded while people reasoned about moral dilemmas and syllogisms did not show a pattern predicted by the dual-process models. Even when participants were familiar with various strategies for solving syllogisms, their preference for the logic-based response in conflict cases (valid-unbelievable, invalid-believable) was inconsistent across the two experiments in this thesis using syllogisms as stimuli. Interestingly, across all three studies, the cohort-level categorization of conflict was consistently replicated. Specifically, low-conflict dilemmas recorded very few endorsements of the utilitarian action as compared to other categories. Participants started with the deontological response and stayed with that preference throughout their deliberations, rarely switching away from their initial choice. This suggests a strong initial bias toward deontological principles in these dilemmas. What cues these priors is open for debate. The dual-process model initially favored the explanation based on the proximity between the action generation and its consequence, suggesting that personal actions involving direct muscular force, such as stabbing and smothering, are rejected [81]. Later, the personal-impersonal distinction was updated to account for interactions with sacrifice and intention [75]. Others have argued that the primary principle underlying this distinction is the likelihood of failure associated with the utilitarian action, which may explain why certain actions are endorsed more frequently in sacrificial dilemmas [106, 143, 148]. Shivnekar and Srinivasan (2024) suggest that the inconsistency in choosing the utility maximizing alternative in the Switch and Footbridge cases may be due to how effective the actions within those dilemmas are perceived to be [148]. For instance, people understand that flipping a switch to divert a trolley is more likely to succeed than pushing a large person off a footbridge to stop the trolley from running over five workers. Therefore, the stronger preference for the deontological principle may be motivated by the expectation that the utility-maximizing action following the utilitarian principle is likely to fail.

Vacillations also closely align with the single-process account of reasoning. In addition to the absence of a consistent pattern in the sequence of preference updates, each interim preference, which may not necessarily represent the final judgment, can be conceptualized as part of an evidence accumulation cycle. Each cycle involves gathering evidence until a certain threshold for an alternative is crossed. Reasoning, therefore, consists of multiple such cycles. But how do we know we have reasoned enough? Here, a metacognitive description of the reasoning process may come to aid. At each step, a metacognitive parameter—indicating our confidence or certainty in the judgment—is concurrently updated. This means that as a reasoner accumulates more evidence and evaluates different alternatives, they are also assessing their confidence in the conclusions drawn. If a reasoner arrives at the same alternative again and again, then her confidence in that alternative will keep increasing and she will be more likely to accept that alternative as the final choice. On the other hand, if she vacillates between choices while thinking, then she may remain underconfident in her answer. In this sense, vacillations can be used as indicators that signify that the evidence gathered so far has not decisively favored one alternative over another. Hence, each step of the cycle accompanies a decision whether to select the alternative with the highest confidence as the answer, continue reasoning, or defer it to another time.

The idea of evidence accumulation, involving either a dissociable or non-dissociable metacognitive component that determines when to stop gathering evidence and, consequently, when to stop thinking, has been examined in other fields, such as perceptual decision-making [119, 139, 154] and reasoning (Jonathan Baron in personal communication; [2, 11]).

Although our study did not directly test a single-process model, the findings from Studies 2 and 3 lend preliminary support to this framework. Specifically, participants who frequently switched between alternatives during their reasoning reported higher levels of conflict and lower confidence in their judgments. This observation is consistent with the proposed metacognitive framework, which suggests that such vacillations reflect ongoing uncertainty and a need for further evidence accumulation.

Lastly, our investigation in Study 2 and Study 3 assumed that there is a difference in how we deliberate on moral dilemmas versus logical problems, particularly when we are trained in strategies employed to solve the latter. While our experience of reasoning in these contexts may differ, we do not claim that they are underpinned by fundamentally different mechanisms. Rather, we suggest that the evolution of the conflict inherent in these two types of problems may vary, which could account for the differences in the experience of reasoning. For example, people generally do not have a strong preference for either the deontological or utilitarian principle. This is evident in the inconsistencies in people's final judgments as well as shifting of inclinations when they reason about (some of) these problems. In certain dilemmas, a particular principle might be strongly cued—such as deontological inaction in low-conflict situations—but in other cases, different considerations may come to mind. Contrasting arguments may pull the preference in opposite directions, sometimes in favor of the action, other times preferring inaction. In terms of the single-process model described earlier, the activation strengths associated with these alternatives may remain in close competition, vying for control of the final judgment. These situations reflect decisions, where the choices are difficult, and uncertainty persists. Conflict is continually experienced in such dilemmas. These dilemmas can lead to extended periods of reflection, with some questions remaining unanswered for years—especially when the solution is ambiguous (deciding whether to whistleblow on suspected fraud at the potential cost of one's safety and reputation) or when the future consequences of a decision are uncertain (choosing a medical treatment for a loved one).

These situations contrast with those when one is solving a mathematical equation or a logical problem when we are equipped with the tools and possess the necessary reasoning skills. The standard against which we compare our strategies in these problems is often clear. We follow steps, matching patterns to a strategy to arrive at an answer. For example, in a syllogism, there may not be an inherent preference for a valid or invalid judgment. A reasoner might construct Venn diagrams of the premises and check the conclusion against the sets in the diagram. Each time the arrangement matches the conclusion, the evidence in support of the conclusion's validity builds up. This will continue until the conclusion does not follow from an arrangement. Conflict between the arrangement of the premises and the conclusion, therefore, may not be experienced continually in these problems. Thus, although conflict may be experienced temporally distinctly in the moral and logical reasoning contexts, the fundamental cognitive processes at work may be similar. Both contexts involve evaluating competing alternatives, but the progression of conflict may differ depending on how the competition between the alternatives is resolved over time.

6.3 Future directions

Active reasoning is dynamic: sometimes people reach a decision and stop thinking, while other times they defer their judgment for later or leave a problem unresolved without intending to return to it. In this thesis, we have argued for an account of the evolving nature of conscious thought in theories of reasoning and decision-making. A reasoner's contemplations and metacognitive judgments, gleaned during or at the end of a choice, can be remarkably informative in understanding how we reason. For instance, people's own judgments about how well and deeply they reasoned consistently aligned with dynamic measures of conflict, irrespective of their correlation with theory-driven operationalizations of conflict. Tools that are sensitive to the temporal and phenomenal character of conflict can shed light on the underlying cognitive operations that drive the evolving nature of thought.

The decision context in all experiments in this thesis was static such that the information about the context, alternatives and what they entail was provided from the outset. What is often missing in setups like these is how individuals interpret and respond to information as they encounter it. While the Switch paradigm allowed us to identify when participants shifted between two options, we were unable to determine the specific consideration that triggered the switch in preference. A reasoner's gaze can indicate which parts of the information are currently being focused on. Although we did not incorporate fixation locations in our reasoning investigation, this approach could yield valuable insights, particularly for shorter and predominantly visual tasks.

On the other hand, paradigms can be designed such that not all information needed to make a decision is given to the participant from the beginning. Information can be revealed when participants interact with their environment. Such paradigms, when paired with our Switch method, can help more specifically identify what aspects of information pushed an individual to consider a different perspective or changed their preference. Take the multi-armed bandit task, for instance. In such tasks, sampling from different options (slot machines) reveals varying payoffs. It is only through repeated sampling over time that the probability of success becomes apparent. When people notice changes in payoffs, they adjust their strategies, switching from exploiting a known option to exploring new ones [93].

In this thesis, our aim was to track conflict closely in time when people navigated difficult problems with tools that minimized interference with the task. Mouse-tracking employed at the end of reasoning fares poorly in pinpointing periods in thinking when the reasoner was conflicted. Although the Switch paradigm allows for a more granular insight into the process of reasoning, there is room for enhancement to extract even more information about the intricacies of this process. To continuously monitor preferences, joysticks can be employed, with the direction of movement indicating the current preference and the distance displaced reflecting the extent of certainty associated with that preference simultaneously. Future studies may also explore modifying instructions to align with specific task demands, such as tracking changes in confidence instead of preferences during reasoning.

One of the biggest concerns that reasoning studies must address is the ecological validity of the task methods and materials used. We have argued before that highly intrusive task demands may confound how we reason. While concurrent tasks can be intrusive—such as in verbal protocols and cognitive load tasks—other impediments to the primary task also need to be reconsidered. For instance, in our Switch paradigm, we imposed a one-minute mandate to encourage participants to stay focused on the problem and deliberate. This, perhaps, could have led people to reason longer than they typically would. Therefore, minimizing any artificial constraints is recommended. That being said, paradigms should allow the extended nature of reasoning to emerge, too. A natural way to encourage participants to reason would be to reduce the number of problems they are asked to solve. Even a single reasoning problem for a large sample could be enough to capture the natural dynamics of the process. The aim should be to foster reasoning as individuals typically would, without overwhelming them, while still incorporating theory-driven and targeted modifications to the problem set.

In addition to task demands, the stimuli used should also be reconsidered. Moral dilemmas are often criticized for their absurdity and limiting choices, which prevent participants from reporting alternatives beyond action commission or omission, even if they spend time thinking about them. Participants in our moral reasoning experiments sometimes expressed frustration when their preferred choice, such as sacrificing themselves rather than someone else, was not listed. In belief bias experiments, we also did not establish whether the believable conclusions were truly believable, either by gathering independent ratings or by asking participants to rate the believability at the end of the trial. While our aim in this thesis was to use stimuli from well-established literature to externally validate our measures, newer problem sets should be considered moving forward.

6.4 Conclusion

The reverie of thoughts is fluid. In this thesis, we argue that the tools and theories of thinking and reasoning must be adapted to better reflect this complexity. Empirical evidence obtained across three studies indicates that closer tracking of conflict holds the potential for greater insights than can be gleaned from trial-level summaries such as response times or cohort-disagreement levels. Our results demonstrate that prevailing theoretical accounts of reasoning struggle to adequately explain the sequence of vacillations seen in peoples' judgments. We anticipate that the enhanced visibility into the deliberation process afforded by our paradigm will contribute to refining and improving these theoretical models.

Appendix A

Appendix

A.1 Study 1

Experiment 1

The stimuli set for Experiment 1 contained 25 problems with 12 of them being non-moral problems. The rest 13 problems were moral dilemmas of type low-conflict personal, high-conflict personal (henceforth, Low-C and High-C, respectively), impersonal, and harmless-offensive. Harmless-offensive dilemmas were 4 and were taken from Haidt et al. (1993, 2000) [85, 86]. The rest of the moral dilemmas were from Koenigs et al. (2007) and each category had 3 items [100].

Stimuli

Non-moral problems

- **Brownies** X has decided to make a batch of brownies for some guests. She needs butter, milk, and flour in 1:2:3 proportion.
  X has plenty milk and butter, but she has only 7 cups of flour. She decides to use all of it, along with 2.3 cups of butter and 5.4 cups of milk.
  Is it appropriate for X to do that?

- **Lottery** X goes to a casino and wants to bet on a slot machine. He can put money in Jumbo Machine or Lotto Machine.

He can put in Rs. 100 in Jumbo and win Rs. 1000 with a probability of 5%. Otherwise, he can put in Rs. 200 in Lotto and win Rs. 5000 with a probability of 1%. X chooses Lotto.
Is it appropriate for X to do that?

- **Coupons** X has gone to a bookstore to buy Rs 500 worth of books. He has with him two coupons.
  One of these coupons gives you 30% off which expires tomorrow. But it applies for a minimum buying price of Rs. 1,000. This coupon expires tomorrow. The other coupon gives you 15% off his purchase price, and this coupon does not expire for another year. X decides to use the 30%-off coupon for the present purchase.
  Is it appropriate for X to do that?

- **Paperwork** X intends to accomplish two things this afternoon: going for a jog and doing some paperwork. In general he prefers to get his work done before his exercise. The weather is nice at the moment, but the weather forecast says that in a couple of hours it will start to rain. He very much dislikes jogging in the rain, but he doesn't care what the weather is like while he does paperwork. X decides to do the paperwork now and jog later in the rain.
  Is it appropriate for X to do that?

- **Turnips** X is a farm worker driving a turnip-harvesting machine. She wants to be productive as quickly as ossible. S is approaching two diverging paths.
  By choosing the path on the left she will harvest ten bushels of turnips. By choosing the path on the right she will harvest twenty bushels of turnips. If she does nothing her turnip-harvesting machine will turn to the left. She decides to turn your turnip-picking machine to the right.
  Is it appropriate for X to do that?

- **Mutual funds** X is at home one day when the mail arrives. She receives two letters from two reputable corporations that provides financial services.
  Corporation 1 has invited her to invest in a mutual fund with Rs. 60,000. X knows the average return is 7% and also an income tax deduction upto 10% of the capital. Corporation 2 is an SIP with Rs. 5,000 every month for a year. X knows the average retuen is 10% with no income tax rebates. X decides to invest in Corporation 2.
  Is it appropriate for X to do that?

- **Semesters** X is beginning her final two semesters of college. To fulfill her graduation requirements she needs to take a history and a science class by the end of the year.

During the first semester, the history class she wants to take is scheduled at the same time as the science class she wants to take. During the second semester the same history class is offered, but the science class is not. X decides to take history class during the first semester to fulfill her graduation requirements.
Is it appropriate for X to do that?

- **Plant transport** X is bringing home a number of plants from a store that is about two miles from his home. The trunk of his car, which he has lined with plastic to catch the mud from the plants, will hold most of the plants he has purchased.
  X could bring all the plants home in one trip, but this would require putting some of the plants in the back seat as well as in the trunk. By putting some of the plants in the back seat he will ruin the fine leather upholstery which would cost thousands of rupees to replace. He decides to make two trips home in order to avoid ruining the upholstery of the car.
  Is it appropriate for X to do that?

- **Computer** X is looking to buy a new computer. At the moment the computer that she wants costs Rs 50,000. A friend from the industry tells X that he can get her this model for Rs 30,000 in the next year.
  She can also fix her current computer instead of buying a new one. A new GPU costs Rs. 15000 and a new CPU costs 1.7 times the GPU. The new components have a 6 month warranty. Finally X decides to wait for a year.
  Is it appropriate for X to do that?

- **Shower** X is planning to attend a luncheon this afternoon, and before he goes he will need to take a shower. He has some yard work that he would like to do before then, and doing this yard will cause him to sweat a fair amount.
  If X showers before he does his yard work he will have to take another shower before the luncheon. At the present time he could enjoy taking a shower. At the same time, he has a very strong commitment to lowering his water bill and to showering no more than once a day. However, he decides to shower before the yard work. Is it appropriate for X to do that?

- **Pasta** X is preparing pasta with fresh vegetables, and she is deciding on the order in which she will do the various things she needs to do. She is in a big hurry.
  If she first starts the water boiling and then cut the vegetables she will be done in twenty minutes. If she cuts the vegetables and then start the water boiling she will be done in forty minutes. She decides to cut the vegetables first, then start the water

boiling.
Is it appropriate for X to do that?

- **Commute** X needs to travel from Kanpur to Lucknow in order to attend a meeting that starts at 2:00 PM. He can leave at 11:30 AM.
  If X takes the train, he will reach the venue in 2 hours but it is usually is late by 15 minutes. The bus takes 1 hour 55 minutes. From the bus stop, X will have to take an auto to reach the venue in 45 minutes. X wants to reach as early as possible. He decides to take the train.
  Is it appropriate for X to do that?

Low-C dilemmas

- **Transplant** X is a doctor. He has five patients, each of whom is about to die due to a failing organ of some kind. He has another patient who is healthy.
  The only way that X can save the lives of the first five patients is to transplant five of this young man's organs (against his will) into the bodies of the other five patients. If he does this, the young man will die, but the other five patients will live. X decides to perform this transplant in order to save five of his patients.
  Is it appropriate for X to do that?

- **Architect** X is a young architect visiting one of his construction sites with his boss. His boss is a despicable individual who makes everyone around him miserable including X.
  It occurs to X that if he were to push him off of the building he is inspecting his boss would fall to his death and everyone would think it was an accident. X decides to push his boss off the building in order to get him out of his life.
  Is it appropriate for X to do that?

- **Country road** X is driving along a country road when she hears a plea for help coming from some roadside bushes. She pulls over and sees a man who has had an accident while hiking and needs to be taken to a hospital.
  X's initial inclination is to help this man, who will probably lose his leg if he does not get to the hospital soon. However, if she gives this man a lift, his blood will ruin the leather upholstery of her car. X decides to leave this man by the side of the road to preserve her leather upholstery.
  Is it appropriate for X to do that?

High-C dilemmas

- **Crying baby** Enemy soldiers have taken over X's village. They have orders to kill all remaining civilians. X and some of the villagers have sought refuge in the cellar of a large house. Outside X hears the voices of soldiers who have come to search the house for valuables.
  X's baby begins to cry loudly. He covers his mouth to block the sound because his crying will summon the attention of the soldiers who will kill X, his child, and the others hiding out in the cellar. To save himself and the others X decides to smother his child to death.
  Is it appropriate for X to do that?

- **Lifeboat** X is on a cruise ship when there is a fire on board, and the ship has to be abandoned. The lifeboats are overloaded and the lifeboat X is in is sitting dangerously low in the water.
  The seas start to get rough, and the boat begins to fill with water. If nothing is done it will sink before the rescue boats arrive and everyone on board will die. But, if X throws an injured person off the boat, the boat will stay afloat and the remaining passengers will be saved. X decides to throw this person overboard.
  Is it appropriate for x to do that?

- **Preventing an epidemic** X is a waiter. He overhears one of his customers say that he is about to go to jail and that in his last 48 hours of freedom he plans to infect as many people as possible with HIV. X knows this customer well enough to know that he is telling the truth and that he has access to many potential victims.
  X happens to know that this customer has a very strong allergy to poppy seeds. If he eats even one he will go into convulsions and have to be hospitalized for at least 48 hours. X then decides to put poppy seeds in his food.
  Is it appropriate for X to do that?

Impersonal dilemmas

- **Speedboat** While on vacation on a remote island, X is fishing from a seaside dock. He observes a group of tourists board a small boat and set sail for a nearby island. Soon after their departure he hears over the radio that there is a violent storm brewing, a storm that is sure to intercept them.
  The only way that X can ensure their safety is to warn them by borrowing a nearby speedboat. The speedboat belongs to a miserly tycoon who would not take kindly

to X borrowing his property. X decides to borrow the speedboat in order to warn the tourists about the storm.
Is it appropriate for X to do that?

- **Illegal lunch** X is a lawyer working on a big case. The judge presiding over the trial happens to be someone X knew from law school. The two of them were friendly back then, but now, decades later, it seems that her old friend barely remembers her.
  X is quite sure that if she were to talk to him over lunch, she could jog his memory and he would begin to see X as an old friend, which would be very good for X's work on this case. It's illegal for judges and lawyers working on the same case to meet socially but X decides to meet the judge anyway to help her case.
  Is it appropriate for X to do that?

- **Lost wallet** X is walking down the street when she come across a wallet lying on the ground. She opens the wallet and finds that it contains several thousand rupees in cash as well the owner's driver's license.
  X, on the other hand, has been hit by hard times recently and could really use some extra money. She decides to send the wallet back to the owner without the cash, keeping the cash for herself.
  Is it appropriate for X to do that?

Impersonal

- **Incest** X and his sister are traveling together. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach.
  They decide that it would be interesting and fun if they had sex. At very least it would be a new experience for each of them. X's sister is already taking birth control pills, but X decides to use a condom too, just to be safe. They agree to keep this secret between the two of them and never do it again.
  Is it appropriate for X and his sister to do that?

- **Dog** X has a dog that he very much loves. The dog is in a great shape, as X takes him exercising every day and is up to date with his vaccinations and medications.
  One day, X accidentally leaves the main door to his house open and the dog slips out and before X could get him back, the dog gets hit by a car and dies. X had heard that dog meat is delicious, so he decides to clean and then cook the dog meat and eat it.
  Is it appropriate for X to do that?

- **Promise** X is very close to her old mother. After her mother fell sick, X did everything she could to tend to her. Her mother eventually died of old age.
  Before dying, as she lay on the deathbed, X's mother asked X to promise that she would visit her grave every week. X loved her mother very much, so she promised to visit her mother's grave every week. But now after a few months of her mother dying, X finds it difficult to keep her promise. She decides to skip visiting her mother's grave.
  Is it appropriate for X to do that?

- **Cannibalism** X works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body to science for research.
  One night X sees a body that is going to be discarded the next day. She knows the cadaver is thoroughly disinfected and hence is perfectly edible. X decides to take a piece of it home, cook it and eat it.
  Is it appropriate for X to do that?

Experiment 2

Experiment 2 had a stimuli set of 18 problems, with 10 non-moral and 8 moral. All non-moral problems were originally from Koenigs et al. (2007) and also used in Experiment 1 [100].

Stimuli

Non-moral problems

- **Brownies**

- **Coupons**

- **Paperwork**

- **Turnips**

- **Mutual funds**

- **Plant transport**

- **Computer**
- **Shower**
- **Pasta**
- **Commute**

Moral problems

- **Sister** X's younger sister has a tendency to be irresponsible with money. But X's parents have always overlooked it. His parents also expect X to take care of his sister after they pass away and fulfill his brotherly responsibilities.
  However, X has always felt left out of his family. He has come to dislike how his parents prioritize his sister over him, even though he is way more responsible. Now, X has decided to distance himself from his family because he does not want to come second to his sister anymore.
  Is it appropriate for X to do that?
- **Grandma** X is an adult and lives with her parents and a grandmother who can not take care of herself after a paralysis attack. X's parents have hired a caregiver that is with her grandmother 24*7.
  One day, the caregiver has to leave early due to an emergency. X is alone at home with her grandmother. Her grandmother needs a change of diapers. However, X feels too icky to do it all by herself as she has never done it before. She decides to ask a neighbor's house maid to help her with the changing.
  Is it appropriate for X to do that?
- **Adopt** X and his wife adopt an orphan child who has autism even though they are made aware of the challenges involved in raising a child with special needs but the couple decides to go ahead with the adoption.
  However, as days pass, the adopted child's symptoms become too much to handle and puts strain on their marriage. Finally, they decide to give the child up to a nursing facility, even though it was definitely under-resourced as they could only afford a cheap one for a long stay.
  Is it appropriate for X and her husband to do that?
- **Gym** X recently got in a prestigious institute for a postgraduate degree course. He is a gym enthusiast and was excited to find out the new gyming facility in his college. The gym is open for 12 hours everyday and out of them 3 hours are exclusively for

women. However, every student wanting to use the facility is required to pay the annual fees of Rs 10,000. X agrees that women should have a safe space to exercise but he strongly disagrees with the inequality in fees. He decides to start a petition for a new gym facility exclusively for women.
Is it appropriate for X to do that?

- **Missing cat** X was walking through a park when she noticed a cat that looked like one she had seen on posters that were posted around the neighborhood. They said there was a reward of Rs 10,000 for finding their lost cat. So she caught the cat, contacted the owner about it.
  When asked for the reward, the owner said that he would only be able to pay half the amount as he had lost a lot of money due to the pandemic. X refuses to give the cat back till she gets the full reward.
  Is it appropriate for X to do that?

- **Chemistry** X is at the top of his chemistry class. She has been consistently scoring good grades. Once she just couldn't concentrate for an upcoming class test. She still appeared for it but could not answer most of the questions. She got frustrated and crumpled her answer sheet, stuffed it in the trash and left the hall.
  Next class, the professor asks her to stay behind. He apologizes for losing her answer sheet and averages her previous exam scores. X feels bad but decides not to correct her.
  Is it appropriate for X to do that?

- **Comedian** X is a fan of a famous comedian. He has been to his standup comedy shows in the past, follows him on social media.
  Recently, there were sexual misconduct allegations against the comedian. He was also being actively investigated for the allegations. X was conflicted about the situation but decided to stop following the comedian by not watching the older standups of him that X enjoys.
  Is it appropriate for X to do that?

- **Drugs** X has a 19 year old son. The son is brilliant at studies. X is really proud that his son has secured a seat in a top college for his chosen specialization.
  As the time passes, X gets to know that his son has gotten addicted to drugs and has been thrown out of school. X takes him to therapy multiple times but after a few good months his son starts using drugs again. X has spent his savings on his son's treatment, jeopradising his other kid's college funds. X finally decides to cut

ties with his son to be fair to his other child, even though his son begs him not to. Is it appropriate for X to do that?

A.2 Study 2

Experiment 1

Stimuli

We used 16 stimuli from Koenigs et al. (2007) from 4 conditions: non-moral, impersonal moral, low-conflict (Low-C) personal, and high-conflict (High-C) personal moral dilemmas. Although we selected moral dilemmas with higher mean emotionality ratings, we also considered prior exposure to them for the final set. In all problems, the choice was between taking an action and endorsing its omission. In non-moral and impersonal problems, the action is not categorized as either characteristically utilitarian or deontological (henceforth, D and U, respectively; see Greene (2014) for the details on characterizing an action as utilitarian [76]). Barring two dilemmas from Low-C, the action was U in all personal dilemmas (Low-C and High-C). Hence, the D alternative was always inaction in all moral dilemmas (Low-C, High-C, and impersonal) but the action was not necessarily U.

Non-moral problems

- **Standard Turnips** You are a farm worker driving a turnip-harvesting machine. You are approaching two diverging paths.
  By choosing the path on the left you will harvest ten bushels of turnips. By choosing the path on the right you will harvest twenty bushels of turnips. If you do nothing your turnip-harvesting machine will turn to the left.
  Would you turn your turnip-picking machine to the right in order to harvest twenty bushels of turnips instead of ten?

- **Generic brand** You have a headache. You go to the pharmacy with the intention of buying a particular name-brand headache medicine. When you get there you discover that the pharmacy is out of the brand you were looking for.
  The pharmacist, whom you have known for a long time and in whom you have a great deal of trust, tells you that he has in stock a generic product which is, in his words, “exactly the same” as the product you had originally intended to buy.
  Would you purchase the generic brand instead of searching further for the name-brand product you were looking for?

- **Scheduling** You are in charge of scheduling appointments in a dentist’s office. Two people, Mr. Morris and Mrs. Santiago have called to make appointments for next

Monday. The only available times for next Monday are at 10:00 AM and at 3:00 PM.

Mr. Morris's schedule is rather flexible. He can have his appointment either at 10:00 AM or at 3:00 PM. Mrs. Santiago's schedule is less flexible. She can only have her appointment at 10:00 AM.

Would you schedule Mr. Morris for 3:00 PM so that both he and Mrs. Santiago can have their appointments next Monday?

- **Scenic route** An old friend has invited you to spend the weekend with him at his summer home some ways up the coast from where you are. You intend to travel there by car, and there are two routes that you can take: the highway and the coastal road. The highway will get you to your friend's house in about three hours, but the scenery along the highway is very boring. The coastal route will get you to your friend's house in about three hours and fifteen minutes, and the scenery along the coastal road is breathtakingly beautiful.

  Would you take the coastal route in order to observe the beautiful scenery as you drive?

Low-C dilemmas

- **Plane crash** Your plane has crashed in the Himalayas. The only survivors are yourself, another man, and a young boy. The three of you travel for days, battling extreme cold and wind. Your only chance at survival is to find your way to a small village on the other side of the mountain, several days away.

  The boy has a broken leg and cannot move very quickly. His chances of surviving the journey are essentially zero. Without food, you and the other man will probably die as well. The other man suggests that you sacrifice the boy and eat his remains over the next few days.

  Would you kill this boy so that you and the other man may survive your journey to safety?

- **Hired rapist***[1] You have been dissatisfied with your marriage for several years. It is your distinct impression that your wife no longer appreciates you. You remember how she appreciated you years ago when you took care of her after she was mugged. You devise the following plan to regain your wife's affection.

  You will hire a man to break into your house while you are away. This man will tie

[1] Moral dilemmas marked by '*' do not have a utilitarian and/or a deontic alternative.

up your wife and rape her. You, upon hearing the horrible news, will return swiftly to her side, to take care of her and comfort her, and she will once again appreciate you.
Would you hire a man to rape your wife so that she will appreciate you as you comfort her?

- **Architect** You are a young architect visiting one of your construction sites with your boss. Your boss is a despicable individual who makes everyone around him miserable including you.
  It occurs to you that if you were to push him off of the building you are inspecting he would fall to his death and everyone would think it was an accident.
  Would you push your boss off of the building in order to get him out of your life?

- **Transplant** You are a doctor. You have five patients, each of whom is about to die due to a failing organ of some kind. You have another patient who is healthy.
  The only way that you can save the lives of the first five patients is to transplant five of this young man's organs (against his will) into the bodies of the other five patients. If you do this, the young man will die, but the other five patients will live.
  Would you perform this transplant in order to save five of your patients?

High-C dilemmas

- **Crying baby** Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.
  Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.
  Would you smother your child in order to save yourself and the other townspeople?

- **Sophie's choice** It is wartime and you and your two children, ages eight and five, are living in a territory that has been occupied by the enemy. At the enemy's headquarters is a doctor who performs painful experiments on humans that inevitably lead to death.
  He intends to perform experiments on one of your children, but he will allow you to choose which of your children will be experimented upon. You have twenty-four

hours to bring one of your children to his laboratory. If you refuse to bring one of your children to his laboratory he will find them both and experiment on both of them.

Would you bring one of your children to the laboratory in order to avoid having them both die?

- **Bomb 2** You are negotiating with a powerful and determined terrorist who is about to set off a bomb in a crowded area. Your one advantage is that you have his teenage son in your custody.

  There is only one thing that you can do to stop him from detonating his bomb, which will kill thousands of people if detonated. To stop him, you must contact him over the satellite hook-up that he has established and, in front of the camera, break one of his son's arms and then threaten to break the other one if he does not give himself up.

  Would you break the terrorist's son's arm in order to prevent the terrorist from killing thousands of people with his bomb?

- **Lifeboat 2** You are on a cruise ship when there is a fire on board, and the ship has to be abandoned. The lifeboats are carrying many more people than they were designed to carry. The lifeboat you're in is sitting dangerously low in the water—a few inches lower and it will sink.

  The seas start to get rough, and the boat begins to fill with water. If nothing is done it will sink before the rescue boats arrive and everyone on board will die. However, there is an injured person who will not survive in any case. If you throw that person overboard the boat will stay afloat and the remaining passengers will be saved.

  Would you throw this person overboard in order to save the lives of the remaining passengers?

Impersonal dilemmas

- **Lost wallet*** You are walking down the street when you come across a wallet lying on the ground. You open the wallet and find that it contains several hundred dollars in cash as well as the owner's driver's license.

  From the credit cards and other items in the wallet it's very clear that the wallet's owner is wealthy. You, on the other hand, have been hit by hard times recently and could really use some extra money. You consider sending the wallet back to the owner without the cash, keeping the cash for yourself.

Would you keep the money you found in the wallet in order to have more money for yourself?

- **Illegal lunch*** You are a lawyer working on a big case. The judge presiding over the trial happens to be someone you knew from law school. The two of you were rather friendly back then, but now, decades later, it seems that your old friend barely remembers you.
  You're quite sure that if you were to talk to him over lunch, you could jog his memory and he would begin to see you as an old buddy, which would be very good for your work in this case. It's illegal for judges and lawyers working on the same case to meet socially.
  Would you meet with this judge socially in order to help you win your case?

- **Stock tip*** You are a management consultant working on a case for a large corporate client. You have access to confidential information that would be very useful to investors. You have a friend who plays the stock market. You owe this friend a sizable sum of money.
  By providing her with certain confidential information you could help her make a lot of money, considerably more than you owe her. If you did this, she would insist on canceling your debt. Releasing information in this way is strictly forbidden by federal law.
  Would you release this information to your friend so that she will cancel your debt?

- **Resume*** You have been trying to find a job lately without much success. You figure that you would be more likely to get hired if you had a more impressive resume. You could put some false information on your resume in order to make it more impressive. By doing this you might ultimately manage to get hired, beating out several candidates who are actually more qualified than you.
  Would you put false information on your resume in order to help yourself find employment?

Experiment 2

Stimuli

Experiment 2 had three conditions with 3 problems in each condition viz., non-moral, conflict moral and non-conflict moral problems. We selected moral dilemmas from Bago and De Neys (2019) and non-moral problems from Koenigs et al. (2007) for Experiment

2 [6, 100]. All moral problems were characteristically impersonal as the U action did not cause harm directly (see Greene, (2014) for the distinction between personal and impersonal implied here [76]). Conflict moral dilemmas had a clear utilitarian action and a deontological omission of it. On the other hand, non-conflict dilemmas were constructed such that both utilitarian and deontological principles ostensibly converge on the same alternative (see congruent and incongruent distinction from Conway and Gawronski (2013) [27]). Hence, non-conflict dilemmas do not have distinct alternatives with one of them being U action and D omission. For the convenience of discussion, we call the action within these dilemmas U (which is the supposedly convergent option for both utilitarian and deontological principles).

Non-moral

- **Broken VCR** You have brought your broken headphones to the local repair shop. The woman working at the shop tells you that it will cost you about Rs. 2000 to have it fixed.
  You noticed in the paper that morning that the electronics shop next door is having a sale on headphones and that a certain new headphone which is slightly better than your old one is on sale for Rs. 2000.
  Would you buy new headphones instead of repairing the old ones?

- **Choosing classes** You are beginning your final year of college. In order to fulfill your graduation requirements you need to take a history class and a science class by the end of the year.
  During the odd semester the history class you want to take is scheduled at the same time as the science class you want to take. During the summer semester the same history class is offered, but the science class is not.
  Would you take the history class during the odd semester in order to help you fulfill your graduation requirements?

- **Investment offer** You are at home one day when the mail arrives. You receive a letter from a reputable corporation that provides financial services. They have invited you to invest in a mutual fund, beginning with an initial investment of twenty thousand rupees.
  As it happens, you are familiar with this particular mutual fund. It has not performed very well over the past few years, and, based on what you know, there is no reason to think that it will perform any better in the future.
  Would you invest a thousand dollars in this mutual fund in order to make money?

Conflict moral dilemmas

- **Iceberg** You are a radar operator overseeing vessel movement near Greenland. Due to sudden ice movement a boat carrying 11 passengers is about to crash into an iceberg. If nothing is done, all passengers will die. The only way to save the 11 passengers is for you to order the captain to execute an emergency maneuver that will sharply alter the course of the boat.
  However, the maneuver will cause the boat to overrun a life raft carrying 3 people that is floating next to the iceberg and out of sight of the captain. The 3 people on the life raft will be killed if you order the captain to execute the maneuver, but the 11 people on the boat will be saved.
  Would you order the maneuver to alter the course of the boat?

- **Gas explosion** You are the inspector of gas lines in a city. For some reason, the gas in a pipeline, which is running under a building, got on fire. If the fire reaches the building, it will explode, and will kill 12 people in it. You realize that the only way to stop the explosion of the building is to close the pipeline by pushing on a button, and divert the fire to a side-pipeline. If you do so, the 12 people will be saved.
  However, above the side-line is another building with 4 people in it. If you push the button and divert the fire into the sideline, this building will explode and kill the 4people in it, but the 12 in the building above the main line will be saved.
  Would you push the button to divert the fire explosion?

- **Fumes** You are the late-night watchman in a hospital. Due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. The fumes are directly and quickly heading towards a room with 13 patients in it. If you do nothing the fumes will rise up into this room and cause their deaths. The only way to avoid the deaths of these patients is to hit a certain switch, which will cause the fumes to bypass the room and enter a second room instead.
  However, you realize that 5 patients are in the second room. These 5 patients will be killed if you hit the switch and let the fumes bypass the first room, but the 13 patients in the first room will be saved.
  Would you hit the switch to cause the fumes to be released?

Non-conflict moral dilemmas

- **Submarine** You are responsible for the mission of a submarine, leading the operation from a control center on the beach. An onboard explosion has damaged the ship and

collapsed the only access corridor between the upper and lower levels of the ship. As a result, water is quickly approaching the upper level of the ship. If nothing is done, 4 people in the upper level will be killed. You realize that the only way to save these people is to hit a switch in which case the path of the water to the upper level will be blocked and it will enter the lower level of the submarine instead.
However, you realize that 12 people are trapped in the lower level. If you hit the switch, these 12 people (who otherwise would survive) will die, but the 4 people in the upper level will be saved.
Would you hit the switch to divert the water to the lower level?

- **Mine** Due to an accident there are 3 miners stuck in one of the shafts of a copper mine. They are almost out of oxygen and will die if nothing is done. You are the leader of the rescue team. The only way for you to save them is to activate an emergency circuit that will transfer oxygen from a nearby shaft into the shaft where the 3 miners are stuck.
However, your team notices that 11 miners are trapped in the nearby shaft. If you activate the emergency circuit to transfer the oxygen, these 11 miners will be killed, but the 3 miners will be saved.
Would you activate the emergency circuit to divert the oxygen in the shaft?

- **Volcano** After a volcano eruption deadly hot lava is heading towards a nearby village. You are directing the rescue operations. There are 2 people standing on the roof of a house. If nothing is done, these 2 people will inevitably be killed by the lava stream. The only way to save these people is to order the construction of an emergency barrier that will divert the lava stream into an old river bed.
However, you suddenly receive the information that right along the old river bed 10 people are standing on the roof of a barn. If you order the construction of the emergency barrier and divert the lava to save the 2 people on the roof of the house, the 10 people on the roof of the barn will inevitably be killed.
Would you order the construction of the emergency barrier?

Experiment 3

Stimuli

Categorical syllogisms used in Experiment 3 of Study 2 were either single- or multiple-model type. Single-model were of the form

All A are B.
All B are C.
Therefore, all A are C. (valid)

OR

Therefore, all C are A. (invalid)

Multiple-model syllogisms were of the form:

Some A are B.
No B are C.
Therefore, some A are not C. (valid)

OR

No A are B.
Some B are C.
Therefore, some A are not C. (invalid)

In this experiment, participants were tasked with solving eight syllogistic reasoning problems. We utilized a within-subject 2x2 design with two factors: validity (assessing whether the conclusion logically follows from the premises) and believability (evaluating whether the conclusion is believable). To manipulate the believability of the conclusions, we drew on materials from Robison and Unsworth (2017) and Evans et al. (1983) [49, 142].

All the stimuli used are listed below.

Valid-Believable

- **Lollipops (single-model)**
  All lollipops are candy.
  All candy is made of sugar.
  Therefore, all lollipops are made out of sugar.
- **Vicious dogs (multiple-model)**
  Some highly trained dogs are vicious.
  No vicious dogs are police dogs.
  Therefore some highly trained dogs are not police dogs.

Valid-Unbelievable

- **College professors (single-model)**
  All college professors are doctors.
  All doctors have medical degrees.
  Therefore all college professors have medical degrees.

- **Priests (multiple-model)**
  Some priests are young.
  No young people are religious people.
  Therefore some priests are not religious people.

Invalid-Believable

- **Sea creatures (single-model)**
  All animals that spend the majority of their lives in the water are sea creatures.
  All sea creatures are animals that are able to swim.
  Therefore all animals that are able to swim spend the majority of their lives in the water.

- **Cigarettes (multiple-model)**
  No addictive things are inexpensive.
  Some inexpensive things are cigarettes.
  Therefore some addictive things are not cigarettes.

Invalid-Unbelievable

- **Cubes (single-model)**
  All objects with six sides are cubes.
  All cubes are objects with sides of equal area.
  Therefore all objects with sides of equal area are objects with six sides.

- **Deep sea divers (multiple-model)**
  No deep sea divers are nutritionists.
  Some nutritionists are good swimmers.
  Therefore some deep sea divers are not good swimmers.

Additional analyses

Table A.1: Results of generalized regression model of acceptance rates for single-model and multiple-model syllogisms from Experiment 3 of Study 2. Validity and believability are dummy coded with invalid and unbelievable conditions as reference levels.

Single-model				**Multiple-model**		
	Estimate	**Std. error**	**z**	**Estimate**	**Std. error**	**z**
Intercept	-1.8718	0.54	-3.49 ***	-0.5465	0.38	1.44
Validity	3.4812	0.73	4.79 ***	3.1856	0.82	3.87 ***
Believability	0.8602	0.68	1.27	1.7361	0.57	3.02 **
Validity : Believability	17.0964	1963.41	0.01	-2.1780	1.11	1.96 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

A.3 Study 3

Experiment 1

Stimuli

We used the same stimuli from Experiment 1 of Study 2. They are listed in Appendix A.2

Additional analyses

Table A.2: Priors used in the Bayesian hierarchical models reported in Experiment 1 of Study 3 (Table 5.2).All priors followed a normal distribution, denoted as $N(Mean, SD)$. The column on the left identifies the dependent variable with predictors specified in each the columns. In the Choice and Switch models (first two rows), the intercept was dummy-coded with the Low-C condition as the reference. For the Conflict and Confidence models (last two rows), the intercept corresponds to prior estimate when no switch is recorded during the trial.

DV	Intercept	High-C	Impersonal	Switches
Choice	$N(-3, 1)$	$N(3.5, 1)$	$(2.5, 1)$	-
Switches	$N(0, 1)$	$N(1, 0.15)$	$N(0.5, 0.15$	
Conflict	$N(3.5, 0.2)$	-	-	$N(-0.15, 0.05)$
Confidence	$N(2.5, 0.2)$	-	-	$N(0.17, 0.05)$

Experiment 2

Stimuli

Syllogisms in this experiment are listed under the appendix of Experiment 3 of Study 2, Section A.2.

Additional analyses

Table A.3: Mean and 95% CI for fixation duration and pupil size recorded in each condition by preference blocks in Experiment 1 of Study 3. CIs were obtained through 1000 bootstrap resamples.

	Fixation duration			**Standardized pupil size**		
	Before	No-switch	Switch	Before	No-switch	Switch
Non-moral	226.6145 [225.76, 227.5]	257.8283 [254.52, 261.23]	272.41 [266.84, 278.23]	-0.148 [-0.15, -0.14]	0.0746 [0.07, 0.08]	-0.0136 [-0.03, 0]
Moral	234.9115 [234.36, 235.5]	284.0762 [281.23, 287.03]	290.6409 [287.61, 294.04]	-0.0628 [-0.07, -0.06]	0.122 [0.11, 0.13]	0.2235 [0.22, 0.23]
Low-C	233.7275 [232.63, 234.97]	296.4508 [290.94, 302.17]	286.4523 [280.77, 292.43]	-0.0829 [-0.09, -0.08]	0.0049 [-0.01, 0.02]	0.197 [0.18, 0.21]
High-C	237.6844 [236.85, 238.57]	260.0872 [257.21, 263.13]	296.6007 [290.46, 302.45]	0.0738 [0.07, 0.08]	0.1778 [0.16, 0.19]	0.3302 [0.32, 0.34]
Impersonal	232.5757 [231.7, 233.54]	291.1332 [286.05, 296.45]	288.7348 [283.78, 293.58]	-0.2301 [-0.24, -0.22]	0.2151 [0.2, 0.23]	0.1435 [0.13, 0.16]

Table A.4: Results from Experiment 2 of Study 3: LME models of switches in syllogistic reasoning predicting (a) conflict and (c) confidence ratings with participants as a random effect.

(a) Conflict rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	2.0283	0.10	67.80	20.64 ***	Participant	0.45
Switches	0.5483	0.07	465.94	8.42 ***	Residual	0.94

(b) Confidence rating ∼ switches

Fixed effect					**Random effects**	
	Estimate	Std. error	df	t		Variance
Intercept	4.2627	0.06	73.77	69.35 ***	Participant	0.1
Switches	-0.4132	0.05	486.22	7.52 **	Residual	0.70

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table A.5: Analysis of variance tables for single-model and multiple-model syllogisms from Experiment 2 of Study 3. η_p^2 indicates partial η^2 for the corresponding effect.

Single-model				**Multiple-model**		
Effect	(DF_n, DF_d)	**F**	η_p^2	(DF_n, DF_d)	**F**	η_p^2
Validity	(1, 61)	490.36 *	.89	(1, 61)	110.99 *	.65
Believability	(1, 61)	9.10 *	.13	(1, 61)	18.28 *	.23
Interaction	(1, 61)	1.48	.02	(1, 61)	9.07 *	0.13

‘*’ indicates $p < .05$

Table A.6: Mean and 95% CI for fixation duration and pupil size recorded conditions and model-types by preference blocks in Experiment 2 of Study 3. CIs were obtained through 1000 bootstrap resamples.

	Fixation duration			**Standardized pupil size**		
	Before	No-switch	Switch	Before	No-switch	Switch
Single-Model	255.8465 [254.69, 257.12]	285.2058 [283.05, 287.46]	260.1531 [257.55, 262.99]	-0.0383 [-0.04, -0.03]	-0.086 [-0.09, -0.08]	0.1656 [0.16, 0.18]
VB	260.8402 [256.19, 266.1]	323.878 [318.13, 330.32]	300.4109 [290.43, 309.86]	0.0525 [0.04, 0.07]	-0.0158 [-0.03, -0.0]	0.7577 [0.73, 0.78]
VU	274.7352 [270.55, 278.96]	287.1606 [283.68, 290.58]	325.3884 [311.48, 339.03]	0.0223 [0.01, 0.04]	-0.2291 [-0.24, -0.22]	0.1451 [0.12, 0.17]
IB	246.812 [245.4, 248.27]	264.8435 [261.55, 268.65]	235.2352 [232.41, 238.16]	-0.023 [-0.03, -0.01]	0.0195 [0.01, 0.03]	0.1411 [0.12, 0.16]
IU	253.9012 [252.34, 255.44]	274.5631 [270.51, 278.4]	250.5412 [248.19, 253.08]	-0.1297 [-0.14, -0.12]	-0.1167 [-0.13, -0.1]	0.0766 [0.06, 0.09]
Multiple-Model	265.7034 [264.77, 266.56]	280.4102 [279.2, 281.79]	278.2689 [275.91, 280.6]	0.0177 [0.01, 0.02]	-0.0107 [-0.02, -0.01]	0.2017 [0.19, 0.21]
VB	269.7482 [267.57, 271.93]	279.1292 [276.8, 281.63]	267.131 [262.81, 271.37]	0.0525 [0.04, 0.07]	-0.0158 [-0.03, -0.0]	0.7577 [0.73, 0.78]
VU	269.5262 [267.59, 271.54]	291.7847 [288.44, 295.14]	328.6427 [320.8, 336.6]	0.0223 [0.01, 0.04]	-0.2291 [-0.24, -0.22]	0.1451 [0.12, 0.17]
IB	265.0879 [263.3, 266.8]	293.8864 [290.99, 297.0]	260.4931 [257.8, 263.26]	-0.023 [-0.03, -0.01]	0.0195 [0.01, 0.03]	0.1411 [0.12, 0.16]
IU	259.8718 [258.25, 261.71]	264.2675 [262.37, 266.27]	275.8595 [272.54, 279.01]	-0.1297 [-0.14, -0.12]	-0.1167 [-0.13, -0.1]	0.0766 [0.06, 0.09]

Bibliography

[1] Abney, D. H., McBride, D. M., Conte, A. M., and Vinson, D. W. (2015). Response dynamics in prospective memory. *Psychonomic Bulletin & Review*, 22:1020–1028.

[2] Ackerman, R. and Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in cognitive sciences*, 21(8):607–617.

[3] Aston-Jones, G. and Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, 28(1):403–450.

[4] Awad, E., Dsouza, S., Shariff, A., Rahwan, I., and Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337.

[5] Bacon, A., Handley, S., and Newstead, S. (2003). Individual differences in strategies for syllogistic reasoning. *Thinking & reasoning*, 9(2):133–168.

[6] Bago, B. and De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, 148(10):1782.

[7] Ball, L. J., Phillips, P., Wade, C. N., and Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. *Experimental Psychology*, 53(1):77–86.

[8] Banerjee, S. and John, P. (2024). Nudge plus: incorporating reflection into behavioral public policy. *Behavioural Public Policy*, 8(1):69–84.

[9] Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. *Unintended thought*, pages 3–51.

[10] Baron, J. (2019). Actively open-minded thinking in politics. *Cognition*, 188:8–18.

[11] Baron, J. (2023a). Individual differences and multi-step thinking. *Behavioral & Brain Sciences*, 46.

[12] Baron, J. (2023b). *Thinking and deciding*. Cambridge University Press.

[13] Baron, J. and Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition*, 45:566–575.

[14] Baron, J., Gürçay, B., Moore, A. B., and Starcke, K. (2012). Use of a rasch model to predict response times to utilitarian moral dilemmas. *Synthese*, 189:107–117.

[15] Barston, J. L. (1986). An investigation into belief biases in reasoning.

[16] Bauman, C. W., McGraw, A. P., Bartels, D. M., and Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8(9):536–554.

[17] Bear, A. and Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*, 113(4):936–941.

[18] Bellini-Leite, S. C. (2018). Dual process theory: systems, types, minds, modes, kinds or metaphors? a critical review. *Review of Philosophy and Psychology*, 9(2):213–225.

[19] Berridge, C. W. and Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain research reviews*, 42(1):33–84.

[20] Białek, M. and De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology*, 28(5):631–639.

[21] Białek, M. and De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision making*, 12(2):148–167.

[22] Bronstein, M. V., Pennycook, G., Joormann, J., Corlett, P. R., and Cannon, T. D. (2019). Dual-process theory, conflict processing, and delusional belief. *Clinical psychology review*, 72:101748.

[23] Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

[24] Busemeyer, J. R., Gluth, S., Rieskamp, J., and Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in cognitive sciences*, 23(3):251–263.

[25] Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., De Gardelle, V., Castañón, S. H., and Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, 81(6):1429–1441.

[26] Cisek, P. and Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, 33(1):269–298.

[27] Conway, P. and Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of personality and social psychology*, 104(2):216.

[28] Cushman, F., Young, L., and Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, 17(12):1082–1089.

[29] De Groot, A. D. and De Groot, A. D. (1978). *Thought and choice in chess*, volume 4. Walter de Gruyter.

[30] De Neys, W. (2013). Heuristics, biases and the development of conflict detection during reasoning. In *The Developmental Psychology of Reasoning and Decision-Making*, pages 130–147. Psychology Press.

[31] De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on psychological science*, 16(6):1412–1427.

[32] De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 46:e111.

[33] De Neys, W. and Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, 113(1):45–61.

[34] De Neys, W. and Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106(3):1248–1299.

[35] De Neys, W. and Van Gelder, E. (2009). Logic and belief across the lifespan: the rise and fall of belief inhibition during syllogistic reasoning. *Developmental science*, 12(1):123–130.

[36] Dewey, A. R. (2021). Reframing single-and dual-process theories as cognitive models: Commentary on de neys (2021). *Perspectives on Psychological Science*, 16(6):1428–1431.

[37] Dickstein, L. S. (1980). Inference errors in deductive reasoning. *Bulletin of the Psychonomic society*, 16(6):414–416.

[38] Diederich, A. and Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological review*, 125(2):270.

[39] Eldar, E., Cohen, J. D., and Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature neuroscience*, 16(8):1146–1153.

[40] Ericson, J. D., Albert, W. S., and Bernard, B. P. (2021). Investigating the relationship between web object characteristics and cognitive conflict using mouse-tracking. *International Journal of Human–Computer Interaction*, 37(2):99–117.

[41] Ericsson, K. A. and Moxley, J. H. (2019). Thinking aloud during superior performance on tasks involving decision making 1. In *A handbook of process tracing methods*, pages 286–301. Routledge.

[42] Ericsson, K. A. and Simon, H. A. (1980). Verbal reports as data. *Psychological review*, 87(3):215.

[43] Ericsson, K. A. and Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3):178–186.

[44] Ericsson, K. A. and Simon, H. A. (2003). Verbal reports on thinking.

[45] Evans, J. S. B. (1989). *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc.

[46] Evans, J. S. B. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychological bulletin*, 128(6):978.

[47] Evans, J. S. B. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, 13(4):321–339.

[48] Evans, J. S. B. (2017). Dual process theory: Perspectives and problems. *Dual process theory 2.0*, pages 137–155.

[49] Evans, J. S. B., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

[50] Evans, J. S. B. and Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11(4):382–389.

[51] Evans, J. S. B., Newstead, S., Allen, J., and Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology*, 6(3):263–285.

[52] Evans, J. S. B. and Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

[53] Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

[54] Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191.

[55] Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford*, 5:5–15.

[56] Fox, M. C., Ericsson, K. A., and Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? a meta-analysis and recommendations for best reporting methods. *Psychological bulletin*, 137(2):316.

[57] Freeman, J. B. (2014). Abrupt category shifts during real-time person perception. *Psychonomic bulletin & review*, 21:85–92.

[58] Freeman, J. B. and Ambady, N. (2010). Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior research methods*, 42(1):226–241.

[59] Freeman, J. B., Ma, Y., Han, S., and Ambady, N. (2013). Influences of culture and visual context on real-time social categorization. *Journal of experimental social psychology*, 49(2):206–210.

[60] Freud, S. (1914). Remembering, repeating and working-through (further recommendations on the technique of psycho-analysis ii). In *On Freud's "Remembering, Repeating and Working-Through"*, pages 14–21. Routledge.

[61] Frey, D., Johnson, E. D., and De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly journal of experimental psychology*, 71(5):1188–1208.

[62] Friedman, J., Brown, S., and Finkbeiner, M. (2013). Linking cognitive and reaching trajectories via intermittent movement control. *Journal of Mathematical Psychology*, 57(3-4):140–151.

[63] Frisch, S., Dshemuchadse, M., Görner, M., Goschke, T., and Scherbaum, S. (2015). Unraveling the sub-processes of selective attention: Insights from dynamic modeling and continuous behavior. *Cognitive processing*, 16:377–388.

[64] Gagne, R. M. and Smith Jr, E. C. (1962). A study of the effects of verbalization on problem solving. *Journal of experimental psychology*, 63(1):12.

[65] Gervais, W. M. and Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, 336(6080):493–496.

[66] Ghaffari, M. and Fiedler, S. (2018). The power of attention: Using eye gaze to predict other-regarding and moral choices. *Psychological science*, 29(11):1878–1889.

[67] Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology press.

[68] Gigerenzer, G. and Regier, T. (1996). How do we tell an association from a rule? comment on sloman (1996).

[69] Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., and Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, 10(2):252–269.

[70] Gold, J. I. and Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, 404(6776):390–394.

[71] Gold, J. I. and Shadlen, M. N. (2003). The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *Journal of Neuroscience*, 23(2):632–651.

[72] Grayot, J. D. (2020). Dual process theories in behavioral economics and neuroeconomics: A critical review. *Review of Philosophy and Psychology*, 11(1):105–136.

[73] Greene, J. and Haidt, J. (2002). How (and where) does moral judgment work? *Trends in cognitive sciences*, 6(12):517–523.

[74] Greene, J. D. (2008). The secret joke of kant's soul. *Moral psychology*, 3:35–79.

[75] Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to mcguire, langdon, coltheart, and mackenzie. *Journal of Experimental Social Psychology*, 45(3):581–584.

[76] Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*, 124(4):695–726.

[77] Greene, J. D. (2016). Solving the trolley problem. *A companion to experimental philosophy*, pages 173–189.

[78] Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3):364–371.

[79] Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3):1144–1154.

[80] Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2):389–400.

[81] Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

[82] Gürçay, B. and Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning*, 23(1):49–80.

[83] Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

[84] Haidt, J. and Bjorklund, F. (2008). Social intuitionists answer six questions about morality.

[85] Haidt, J., Bjorklund, F., and Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*, 191:221.

[86] Haidt, J., Koller, S. H., and Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of personality and social psychology*, 65(4):613.

[87] Haidt, J., Rozin, P., McCauley, C., and Imada, S. (1997). Body, psyche, and culture: The relationship between disgust and morality. *Psychology and developing societies*, 9(1):107–131.

[88] Harutyunyan, M. (2021). Fighting fake news with accuracy: Dual processing perspective.

[89] Hayes, T. R. and Petrov, A. A. (2016). Pupil diameter tracks the exploration–exploitation trade-off during analogical reasoning and explains individual differences in fluid intelligence. *Journal of cognitive neuroscience*, 28(2):308–318.

[90] Jagau, S. and van Veelen, M. (2017). A general evolutionary framework for the role of intuition and deliberation in cooperation. *Nature Human Behaviour*, 1(8):0152.

[91] Jamieson, R. K., Johns, B. T., Vokey, J. R., and Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. *Nature Reviews Psychology*, 1(3):174–183.

[92] Janis, I. L. and Frick, F. (1943). The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology*, 33(1):73.

[93] Jepma, M. and Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory. *Journal of cognitive neuroscience*, 23(7):1587–1596.

[94] Johnson-Laird, P. N. and Bara, B. G. (1984). Syllogistic inference. *Cognition*, 16(1):1–61.

[95] Just, M. A. and Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480.

[96] Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., and Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological review*, 125(2):131.

[97] Kieslich, P. J. and Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior research methods*, 49:1652–1667.

[98] Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J., and Schulte-Mecklenbeck, M. (2019). Mouse-tracking. *A Handbook of Process Tracing Methods; Routledge: Abingdon, UK*, pages 111–130.

[99] Klaczynski, P. A. and Gordon, D. H. (1996). Self-serving influences on adolescents' evaluations of belief-relevant evidence. *Journal of Experimental Child Psychology*, 62(3):317–339.

[100] Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., and Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138):908–911.

[101] Kohlberg, L. (1968). *The child as a moral philosopher*, volume 2. Psychology today.

[102] Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., and Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, 83(1):26–52.

[103] Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision making*, 8(5):527–539.

[104] Koop, G. J. and Criss, A. H. (2016). The response dynamics of recognition memory: Sensitivity and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(5):671.

[105] Koop, G. J. and Johnson, J. G. (2013). The response dynamics of preferential choice. *Cognitive psychology*, 67(4):151–185.

[106] Kortekamp, K. V. and Moore, C. F. (2014). Ethics under uncertainty: The morality and appropriateness of utilitarianism when outcomes are uncertain. *The American Journal of Psychology*, 127(3):367–382.

[107] Krajbich, I., Bartling, B., Hare, T., and Fehr, E. (2015a). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature communications*, 6(1):7455.

[108] Krajbich, I., Hare, T., Bartling, B., Morishima, Y., and Fehr, E. (2015b). A common mechanism underlying food choice and social decisions. *PLoS computational biology*, 11(10):e1004371.

[109] Landy, J. F. and Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? a meta-analytic review of experimental evidence. *Perspectives on psychological science*, 10(4):518–536.

[110] Maldonado, M., Dunbar, E., and Chemla, E. (2019). Mouse tracking as a window into decision making. *Behavior research methods*, 51:1085–1101.

[111] McGuire, J., Langdon, R., Coltheart, M., and Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, 45(3):577–580.

[112] Melnikoff, D. E. and Bargh, J. A. (2018). The mythical number two. *Trends in cognitive sciences*, 22(4):280–293.

[113] Mercier, H. and Sperber, D. (2017). *The enigma of reason.* Harvard University Press.

[114] Merlhiot, G., Mermillod, M., Le Pennec, J.-L., Dutheil, F., and Mondillon, L. (2018). Influence of uncertainty on framed decision-making with moral dilemma. *PloS one*, 13(5):e0197923.

[115] Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., and De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, 27(2):227–237.

[116] Miles, J. D. and Proctor, R. W. (2015). Attention is captured by distractors that uniquely correspond to controlled objects: An analysis of movement trajectories. *Attention, Perception, & Psychophysics*, 77:819–829.

[117] Moore, A. B., Clark, B. A., and Kane, M. J. (2008). Who shalt not kill? individual differences in working memory capacity, executive control, and moral judgment. *Psychological science*, 19(6):549–557.

[118] Moore, A. B., Lee, N. L., Clark, B. A., and Conway, A. R. (2011). In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making*, 6(3):186–195.

[119] Moran, R., Teodorescu, A. R., and Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive psychology*, 78:99–147.

[120] Morgan, J. J. and Morton, J. T. (1944). The distortion of syllogistic reasoning produced by personal convictions. *The Journal of Social Psychology*, 20(1):39–59.

[121] Murphy, P. R., Robertson, I. H., Balsters, J. H., and O'connell, R. G. (2011). Pupillometry and p3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology*, 48(11):1532–1543.

[122] Newstead, S. E., Pollard, P., Evans, J. S. B., and Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, 45(3):257–284.

[123] Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231.

[124] Norman, D. A. (1988). The psychology of everyday things.

[125] Oakhill, J. and Johnson-Laird, P. N. (1985). The effects of belief on the spontaneous production of syllogistic conclusions. *The Quarterly Journal of Experimental Psychology*, 37(4):553–569.

[126] Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review*, 11(6):988–1010.

[127] Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., and Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, 112(13):4170–4175.

[128] Paxton, J. M., Ungar, L., and Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive science*, 36(1):163–177.

[129] Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51:195–203.

[130] Pennycook, G., Fugelsang, J. A., and Koehler, D. J. (2015a). Everyday consequences of analytic thinking. *Current directions in psychological science*, 24(6):425–432.

[131] Pennycook, G., Fugelsang, J. A., and Koehler, D. J. (2015b). What makes us think? a three-stage dual-process model of analytic engagement. *Cognitive psychology*, 80:34–72.

[132] Pennycook, G. and Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.

[133] Pieters, R. and Warlop, L. (1999). Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of research in Marketing*, 16(1):1–16.

[134] Purcell, Z. A., Howarth, S., Wastell, C. A., Roberts, A. J., and Sweller, N. (2022). Eye tracking and the cognitive reflection test: Evidence for intuitive correct responding and uncertain heuristic responding. *Memory & Cognition*, pages 1–18.

[135] Purcell, Z. A., Roberts, A. J., Handley, S. J., and Howarth, S. (2023). Eye movements, pupil dilation, and conflict detection in reasoning: Exploring the evidence for intuitive logic. *Cognitive Science*, 47(6):e13293.

[136] R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

[137] Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416):427–430.

[138] Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922.

[139] Ratcliff, R. and Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychological review*, 120(3):697.

[140] Raven, J. (2003). Raven progressive matrices. In *Handbook of nonverbal assessment*, pages 223–237. Springer.

[141] Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, 461(7261):263–266.

[142] Robison, M. K. and Unsworth, N. (2017). Individual differences in working memory capacity and resistance to belief bias in syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, 70(8):1471–1484.

[143] Royzman, E. B., Kim, K., and Leeman, R. F. (2015). The curious tale of julie and mark: Unraveling the moral dumbfounding effect. *Judgment and Decision making*, 10(4):296–313.

[144] Russo, J. E. (2019). Eye fixations as a process trace. In *A handbook of process tracing methods*, pages 4–26. Routledge.

[145] Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science*, 15(2):207–215.

[146] Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., and Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, 26(5):442–450.

[147] Schwitzgebel, E. and Cushman, F. (2015). Professional philosophers' susceptibility to order effects and framing effects in evaluating moral dilemmas. *Cognition*, 141:127–137.

[148] Shivnekar, R. and Srinivasan, N. (2024). Choosing between bad and worse: investigating choice in moral dilemmas through the lens of control. *Cognitive Processing*, pages 1–8.

[149] Shivnekar, R. V. (2023). Vacillations in syllogistic reasoning.

[150] Shivnekar, R. V. and Srivastava, N. (2023). Measuring moral vacillations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

[151] Simpson, E. L. (1974). Moral development research: A case study of scientific cultural bias. *Human development*, 17(2):81–106.

[152] Skulmowski, A., Bunge, A., Kaspar, K., and Pipa, G. (2014). Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. *Frontiers in behavioral neuroscience*, 8:426.

[153] Song, J.-H. and Nakayama, K. (2008). Target selection in visual search as revealed by movement trajectories. *Vision research*, 48(7):853–861.

[154] Spence, M. L., Dux, P. E., and Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, 42(5):671.

[155] Spivey, M. (2008). *The continuity of mind.* Oxford University Press.

[156] Spivey, M. J. and Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, 15(5):207–211.

[157] Spivey, M. J., Grosjean, M., and Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, 102(29):10393–10398.

[158] Srivastava, N. and Vul, E. (2015). Choosing fast and slow: explaining differences between hedonic and utilitarian choices. In *CogSci.* Austin, TX.

[159] Stanovich, K. E. and Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence*, 11(2):27.

[160] Stanovich, K. E. and West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of educational psychology*, 89(2):342.

[161] Stanovich, K. E. and West, R. F. (2000). Advancing the rationality debate. *Behavioral and brain sciences*, 23(5):701–717.

[162] Stillman, P. E., Krajbich, I., and Ferguson, M. J. (2020). Using dynamic monitoring of choices to predict and understand risk preferences. *Proceedings of the National Academy of Sciences*, 117(50):31738–31747.

[163] Strack, F. and Deutsch, R. (2015). The duality of everyday life: dual-process and dual system models in social psychology.

[164] Sullivan, E. V. (1977). A study of kohlberg's structural theory of moral development: A critique of liberal social science ideology. *Human development*, 20(6):352–376.

[165] Suter, R. S. and Hertwig, R. (2011). Time and moral judgment. *Cognition*, 119(3):454–458.

[166] Swann Jr, W. B., Gómez, Á., Buhrmester, M. D., López-Rodríguez, L., Jiménez, J., and Vázquez, A. (2014). Contemplating the ultimate sacrifice: identity fusion channels pro-group affect, cognition, and moral decision making. *Journal of personality and social psychology*, 106(5):713.

[167] Tabatabaeian, M., Dale, R., and Duran, N. D. (2015). Self-serving dishonest decisions can show facilitated cognitive dynamics. *Cognitive Processing*, 16:291–300.

[168] Thompson, V. A., Turner, J. A. P., and Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive psychology*, 63(3):107–140.

[169] Thomson, J. J. (1984). The trolley problem. *Yale LJ*, 94:1395.

[170] Titchener, E. B. (1912). The schema of introspection. *The American Journal of Psychology*, 23(4):485–508.

[171] Toplak, M. E., West, R. F., and Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental psychology*, 50(4):1037.

[172] Toplak, M. E., West, R. F., and Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making*, 30(2):541–554.

[173] Travers, E., Rolison, J. J., and Feeney, A. (2016). The time course of conflict on the cognitive reflection test. *Cognition*, 150:109–118.

[174] Trippas, D., Thompson, V. A., and Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & cognition*, 45:539–552.

[175] Tversky, A. and Shafir, E. (1992). Choice under conflict: The dynamics of deferred decision. *Psychological science*, 3(6):358–361.

[176] Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., and Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283(5401):549–554.

[177] Van der Wel, P. and Van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review*, 25:2005–2015.

[178] Van Someren, M. W., Barnard, Y. F., Sandberg, J. A., et al. (1994). The think aloud method: a practical approach to modelling cognitive processes. *London: AcademicPress*, 11:29–41.

[179] Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., and LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin*, 19(3):331–339.

[180] Wulff, D. U., Haslbeck, J. M., Kieslich, P. J., Henninger, F., and Schulte-Mecklenbeck, M. (2019). Mouse-tracking: Detecting types in movement trajectories. In *A handbook of process tracing methods*, pages 131–145. Routledge.

[181] Zahrai, K., Veer, E., Ballantine, P. W., de Vries, H. P., and Prayag, G. (2022). Either you control social media or social media controls you: Understanding the impact of self-control on excessive social media use from the dual-system perspective. *Journal of Consumer Affairs*, 56(2):806–848.

[182] Zbrodoff, N. J. and Logan, G. D. (1986). On the autonomy of mental processes: a case study of arithmetic. *Journal of Experimental Psychology: General*, 115(2):118.

[183] Zhao, W. J., Richie, R., and Bhatia, S. (2022). Process and content in decisions from memory. *Psychological Review*, 129(1):73.