This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
The Agency Machine: Motives, Levels of Confidence and Metacognition

by

Jonathan. J. Hall

Submitted to the School of Philosophy, Psychology and Language Sciences
The University of Edinburgh
In Fulfilment of the Requirements
For a Doctorate in Philosophy
February 2024
Acknowledgements

This dissertation owes its primary scholarly debt to Professor Tillmann Vierkant. Without his tireless support, advice, and enthusiasm this project would never have been completed. It has been a real pleasure. I would also like to thank Doctor Suilin Lavelle, whose insightful comments and timely interventions have been instrumental in shaping the thesis. In addition, I have benefitted hugely from conversations with Professor Uriah Kriegel, Doctor Juan Pablo Bermúdez, and many hours of debate with Doctor Graham Doke. Thank you.

On a personal note, I would like to thank my family and particularly my wife, Julia. Her support and enthusiasm for my academic endeavour has never wavered, although the motivational pull of lunchtime philosophical debate has perhaps diminished with time. Others to whom I owe a personal debt will find their names scattered throughout the thesis, in examples and vignettes.

Finally, I would like to pay my respects to the late Paul Snowdon, whose infectious intellectual curiosity ignited my love for philosophy, and sustains it to this day.
Abstract

In this thesis I aim to advance philosophical understanding of human agency, and resolve some knotty philosophical puzzles, by engaging in a novel fine-grained analysis of conative and cognitive phenomenology. Taking the phenomenology of the decision-making process seriously is sensible for three reasons: First, instances of phenomenology are data to be explained. Any theory which ignores their existence is incomplete. Second, experience-based seemings pay a central part in belief formation. Any unresolved conflict between experience and theory will either lead to rejection of the theory or to a dialectic impasse. Third, phenomenologically informed theories, are better placed to answer some of the unresolved philosophical questions including does agential phenomenology support beliefs in libertarianism (no), is intentional synchronic self-control possible (yes), and can making a decision ever be an action (yes)?

The central claim that I defend throughout the thesis is the motivational strength idea, that when an agent makes a choice, she always chooses to do what she is most motivated to do. Whilst others have rejected or added complexity to the idea, I will remain true to its simplicity. Doing so will force me to be flexible elsewhere, and question other assumptions, which will pay dividends. From this foundation, and focusing almost entirely on motives and levels of confidence in the context of a sophisticated metacognitive agent, I am able to present an experience compatible, and evolutionarily plausible theory of agency. Although the broad picture is uncontroversial, some of the findings may be surprising:
An intention is just an uninhibited motive.

The phenomenological intensity of a motive is a representation of its functional strength if and only if it is inhibited.

The feeling of effort is, for all subspecies of effort, a non-conceptual representation of the conative costs of a pursued course of action. It represents-as-costly a trying.

An instrumentally rational (undivided) agent with metacognitive abilities can intentionally shift her own current motivational balance.

Making a difficult decision is an action. It feels effortful because it involves resisting the motivation to deliberate further.

Whilst functional analysis celebrates the victory of the strongest motive (indicating only that it was stronger than its competition), phenomenological analysis reveals the strength of its weaker, inhibited cousin. This inversion encourages a new perspective on the phenomenology of agency. Rather than of desires being realised and intentions being implemented, it is the experience of uncertainty, inhibited motives and unfulfilled plans.

The model that emerges is hierarchical but extremely simple, explaining flexible agency through the interaction of inhibitable motives in accordance with the motivational strength idea. There is no reliance on distinct intending attitudes, motivational partitions, reservoirs of willpower, Cartesian homunculi, or error theory. There is just an agent, deliberating about and then doing (or at least trying to do) what she most wants to do.
Lay Summary: Motives, Confidence and Metacognition

“Life is the sum of all your choices,” Albert Camus

Contemporary theories of human agency are complex but incomplete. I believe that by shedding the complexity of many models, and focusing almost entirely on motives and levels of confidence, I can present an experience compatible and evolutionarily plausible theory of agency. Although the analysis required to defend this theory will be necessarily technical, the core model is extremely simple. I try to present it as such below.

If life is a journey, then our story of agency begins at a fork in the road: “What to do?”. Right or left, high road or low road, action A or action B? Choosing the right answer can be important. It might even be a matter of life or death. For that reason, human cognitive infrastructure implements a decision-rule, which aims to optimise the answer to any practical question, given an information set.

Here is the leading contender for the decision-rule:

When an agent makes a choice, she always chooses to do what she is most motivated to do.

This is known as the motivational strength idea (Mele, 2003), where the strength of a motive is generally thought to represent the expected value (to the agent) of a considered course of action. The motivational strength idea is broadly, but not universally, accepted. I will argue that it is correct.
Within the decision-rule, the introductory clause is important. It indicates that a practical question might not be answered right away. A human traveller at a fork in the road can withhold her decision about which path to follow, and engage in an extended process of deliberation.

An obvious, but perhaps underexplored, assumption of the deliberate-decide-act model, is that agents are not compelled to act on every motive. Motives are inhibitable, and unchosen motives are inhibited. Pre-decision, this allows for deliberation before implementation, and post-decision, it ensures that unchosen motives do not interfere with the implementation of chosen motives.

A chosen motive is generally called an intention. This suggests the following possibility:

*An intention is just an uninhibited motive.*

Few philosophers accept this claim. It is generally believed that intentions have further intrinsic features, such as stability (Bratman, 1999, Holton, 2009b), that differentiate them from simple motives. Whilst I agree that intentions are stable, I deny that this feature is intrinsic to the intending attitude. I argue instead that intention stability is driven by factors extrinsic to the attitude, including what has been called the implementation mindset. I therefore endorse the simplifying claim that intentions are merely uninhibited motives.

If an agent can withhold her decision while she deliberates, this raises a question about decision timing. When will she exit the “withholding zone” of extended deliberation and decide?

Here is my suggested answer to that question:
An agent will end deliberation and decide when she reaches a threshold of confidence in her decision.

Although this should not be controversial, it does raise two further questions: What determines the level of the threshold, and what can an agent do while she is waiting for the threshold to be hit? For sophisticated agents the answer to these questions is influenced by the existence of 'managerial' (Hieronymi, 2009) actions: actions which are part of, and serve, the extended deliberative process.

However, before analysing the various kinds of managerial actions, it is worth pausing on the link between the functional model proposed thus far and the felt experience (henceforth “phenomenology”) of decision making. Auspiciously, the two central elements, motives and levels of confidence, are felt. In an ideal world, introspection would confirm the functional theory, but unfortunately it does not appear to do so, at least at first. Two puzzles immediately arise: First, phenomenologically intense motives can be resisted by opposing motives that don’t feel as strong. The desire to stay home and study feels much less strong than the desire to go to a concert! This is the puzzle of relative strength. Second, an agent seems to be able to actively (and effortfully) make a decision even if her confidence is low. Whether, and if so how, a decision can be an action is the puzzle of active decisions. The opportunity to resolve these puzzles is an opportunity to deepen an understanding of the mind.

The puzzle of relative strength seems to demand either rejection of the motivational strength idea (Altehenger, 2021a, Sripada, 2014), or acceptance of a form of error theory, denying the functional implications of phenomenological evidence (Mele, 2009).

---

1 Hieronymi defines managerial control as any action that manipulates an attitude. I broaden this to any action that manipulates an attitude or cognitive process. Vierkant (2022) also uses managerial in this broader sense.
2003). However, I propose a novel “conditional” solution, based upon the inhibitability of motives.

Here is the novel solution:

*The phenomenological intensity of a motive is a representation of its functional strength if and only if it is inhibited.*

Significant work is required to support this claim, but hopefully readers will find it aligns with their experience. It is consistent with the fact that the motivational pull of a desire to go to a concert is felt strongly when it is being resisted (by the desire to study, for example), but not when one is walking to the venue.

Whilst functional analysis celebrates the victory of the strongest motive (indicating only that it was stronger than its competition), phenomenological analysis reveals the strength of its weaker, resisted cousin. This inversion encourages a new perspective on the phenomenology of agency. It is not the experience of implemented motives and intentions being executed, but of frustrated desires and of (as yet) unfulfilled plans. It is the phenomenology of the road not taken. Humans are flexible agents, motivated but not compelled by inhibitible motives. The phenomenology of motivational pull evidences this lack of compulsion.

The possibility of active decisions implies that the deliberate-decide-act theory of agency might be, in its current form, incomplete. If decisions can be actions, that suggests that an agent can intervene in at least part of the decision-making process. However, rather than justifying a belief in the existence of a Cartesian homunculus, it is an argument for the existence of managerial actions. I return to those now.

Managerial actions are actions which are part of, and influence, the extended deliberative process. A model of the mind that includes managerial actions is
metacognitive, where metacognition is cognition about cognition and includes the processes that monitor, evaluate and control cognitive activity. This model of the mind, I will call hierarchical. A simple schematic of a hierarchical mind, where managerial actions take place within extended deliberation, is presented below:

Managerial actions come in (at least) three flavours. Because they are actions, the motivational strength idea applies. An agent will only engage in a managerial action if it is supported by her balance of motivation.

The first kind of managerial action relates to the collection of evidence. It might involve going to the library, asking a friend, or actively focusing attention. Whilst these actions are often ordinary in their content, they are epistemic in their goal. They are motivated by the desire to reduce uncertainty or increase coherence.
Actions that follow from these motivations can indirectly influence the final decision because they change the evidence set on which the decision is based. These actions are uncontroversial.\(^2\)

The second kind of managerial action treats a mental attitude as an object to be manipulated. It is an act of self-control, usually aimed at supressing a wayward desire. For example, a student might actively focus on tomorrow's supervision meeting, in an attempt to control her current desire to go to the pub. Control actions are the means to the end of a metacognitive motive, a motive towards a motive. The existence of these metacognitive motives is uncontroversial, as is the fact that control actions can sometimes control wayward desires.

As with all actions, an agent will only engage in an act of control if it is supported by her overall balance of motivation. However, instrumental, or means-end, rationality means that this is a high hurdle. Let me explain with an example inspired by Skinner (1953):

Imagine that Skinner is in bed, facing the practical question of whether to get up. He wants to remain in bed, but he simultaneously wants to want to get up. The former is an ordinary motive, the latter is metacognitive. Imagine further that he has taught himself a control action - shouting “Get up!” - which he knows to be effective in increasing his desire to get up. His metacognitive motive is in favour of this control action, but such an action is a means to the end of getting out of bed, and he doesn’t want to do that! On the assumption that Skinner is instrumentally rational, the motivational force of his desire to stay in bed will work as motivation against shouting “Get up!”. Given this, is it possible for Skinner to shout? The answer is yes, but the

\(^2\) Although there may be some controversy over whether I am right to call them managerial, rather than epistemic.
hurdle is high. It is only possible if his direct metacognitive motive towards control is
greater than his transferred instrumental motives against.

There is an interesting asymmetry within instrumental rationality that is worth
clarifying, as it resolves what Altehenger (2021a) has called the mismatch problem,
which I express as a question: How is it possible for the motivational balance of a
unified, instrumentally rational agent to favour (simultaneously) a control action and a
wayward action? To answer this, I will return to the example of Skinner.

Skinner is in bed and he has not yet executed his control action. Although shouting
“Get up!” would, if executed, affect his desire to get up, merely contemplating
shouting does not. As yet his desire to remain in bed is untroubled. Equally, staying
in bed does not harm Skinner’s ability to execute the control action. Therefore,
instrumental rationality does not entail that the metacognitive motive weighs against
the act of staying in bed. Whether he remains in bed or not is solely dependent on
his ordinary motives. Skinner’s single motivational set is simultaneously in favour of
shouting “Get up!” and in favour of staying in bed. The mismatch problem has been
resolved through analysis of metacognitive motives and the asymmetry of
instrumental rationality.

*If an agent has a metacognitive motive towards control, then her motivational
balance can favour both a control action and a wayward action
simultaneously.*

To summarise, there are two motivational contests. One for the control decision and
another for the practical decision. The metacognitive motive is only relevant for the
former.
Of course, if the control action is effective, then once it is executed the lower-order balance of motives will shift. Then, finally, Skinner will get up!

The third kind of managerial action relates to the question of when to conclude deliberation and suggests a solution to the puzzle of active decisions. The act of making a difficult decision is the metacognitive intentional act of concluding deliberation. Making a difficult decision is a motivated act, rather than a passive event, because it involves resisting the ongoing desire to deliberate further and reduce the risk of error. It implies that the threshold of confidence which determines when deliberation ends is not fixed. Confidence in a decision is a want-to-have rather than a must-have.

This is adaptive because there can be situations in which the potential harm of deciding early will be outweighed by the harm of delaying. For example, imagine you are running from a polar bear in Svalbard and are choosing between climbing into an unlocked car or running into a house\(^3\). How quickly you make the decision is far more important than which decision you make! Whilst, the motivation to reduce uncertainty will always be on the side of prolonged deliberation, in situations where confidence is already high, stakes are low, and/or progress in uncertainty reduction has stalled, the strength of this motive may be low. Against this, an agent may be strongly motivated to close the question, either to free up cognitive resources or because there are costs in delaying implementation.

This solution to the puzzle of active decisions suggests an updated answer to the question of when an agent ends deliberation:

\(^3\) Cars and houses in Svalbard are left unlocked for this reason.
An agent will end deliberation and decide if and only if her motivation towards concluding deliberation is greater than her motivation towards increasing her confidence in the decision.

The ability to stop or prolong deliberation is merely the metacognitive version of the evolutionarily antecedent ability to stop (or prolong) physical activity.

Although I have provided a functional analysis of managerial actions, there is one aspect that I have left unexplored: That they feel effortful. Focusing attention, engaging in self-control, and making a difficult decision are all hard. Although there is an intuitive link between feelings of effort and intentional action, this has not been fleshed out in a theory that unifies the three sub-species of effort: effort of will, cognitive effort and muscular effort. As Bermúdez asks: ‘What is the feeling of effort about?’ (2023).

My answer to this question mirrors the earlier analysis of motivational pull. While the positive pull of a motive indicates that an action plan is attractive, the aversive feeling of effort reveals that an action is controversial or unsustainable or costly. And what is the measure of costliness? It is the strength of the motive to do otherwise. If the strength of a motive towards an action represents its expected value, then the strength of a motive against the action represents its (opportunity) cost. The feeling of effort is the feeling of a motive being inhibited.

My answer to Bermudez’ question is that the feeling of effort is a feeling about an action, which indicates to the agent that the action has costs. In more technical language:

\[ \text{The feeling of effort represents-as-costly an action} \]
This proposal is controversial because it implies that the feeling of effort is a window through which the agent can view an input into the decision of what to do. As such, on pain of duplication, feelings of effort cannot further bias the agent against the chosen action. But doesn’t this imply that the feeling of effort is little more than ‘a steam whistle on the mind-engine’ (Huxley, 1874)? If so, that might seem to be an argument to reject, or at least be wary of the account proposed. I offer a few brief thoughts:

First, windows and whistles serve a purpose. The feeling of effort is a person level phenomenon, which broadcasts information about an element of the cognitive machinery to person-level awareness. Second, the intensity of felt effort is dynamic. Through awareness of the level of effort, human agents can monitor changes. Any fast or surprising change may trigger a review of the situation, even if the effort is not at a level that would normally equate to a decision to stop. Third, when an action is costly, awareness of this fact may activate a search for alternative means to the same end. Finally, even if the feeling of effort cannot impact the current decision, it can be valuable to a diachronic agent who can remember historic efforts or imagine future efforts. It can be highly valuable for future predictions and decisions and planning. Some agents might even seek out the feeling of effort, or praise it in others. “Good effort, mate!”

I therefore conclude that although the feeling of effort is a representation of a cost that has already been incorporated, it does still have value to a metacognitive, diachronic agent.

The model that I have presented solves a number of outstanding puzzles within philosophy of mind. It does this by removing complexity and allowing the interactions
of inhibitable motives to be governed by the motivational strength idea. There is no reliance on distinct intending attitudes, motivational partitions, reservoirs of willpower, or Cartesian homunculi. There is just an agent who does what she most wants to do.

For agents with metacognitive abilities, extended deliberation incorporates managerial decisions and actions. These do not directly answer the practical question of what to do, but they serve the decision-making process. They answer the intermediate managerial questions of what information to gather, whether to try to do something about mental states, and whether the current level of confidence is sufficient, given the context. They do so in accordance with the motivational strength idea.

The resulting model may be hierarchical, but the rules that govern it are extremely simple. The consistent application of the motivational strength idea makes it evolutionarily plausible, and there are no phenomenological gaps which must be explained away with error theory. It is experience compatible.
# Table of Contents

Introduction.

i. The case for phenomenological analysis  
ii. Chapter summary

Part I

1. Attitudes, Values and Decisions  
2. Intentions

Part II

3. Conative Phenomenology: Motivational Pull  
4. Conative Phenomenology: Effort  
5. Intentional Synchronic Self-Control  
6. The Phenomenology of Alienation

Part III

7. Levels of Confidence and Decision Thresholds  
8. Cognitive Phenomenology: Levels of Confidence  
9. The Act of Making a Difficult Decision

Conclusion

10. The Surprising Importance of Inhibition  
11. Experience-Compatibilism
Introduction

(i)
The case for phenomenological analysis

1. Experience-compatibilism

In this thesis I take phenomenology seriously. I engage in a fine-grained analysis of the conative and cognitive phenomenology that is experienced during the decision-making process. Although I am not the only contemporary philosopher who takes such an approach (Bayne and Levy, 2006, Bermúdez, 2023, Holton, 2006, Kriegel, 2015, Vierkant, 2022, Velleman, 1992), we make up a relatively small tribe. This raises the question of whether deep phenomenological analysis is worthwhile. In particular, whilst phenomenological insights might be interesting, do they have broader implications for understanding how the mind works? Should philosophers and scientists who are interested in functionality spend time on phenomenology? Here I will try to persuade the reader that the answer is yes!

The world that science and philosophy describe includes both objective material interactions and subjective human experience. Too often these have been considered independently. Either separated, perhaps irredeemably, by a Cartesian Cut, or seen in a dominant and subordinate relationship where the latter is only turned to as an afterthought. Here, I motivate a project that I call experience-compatibilism, in which phenomenological and functional analysis are given equal emphasis. Building an understanding of psychological phenomena that places
phenomenology at its heart pays dividends not just because it takes seriously the experiences that are so central to people’s lives, but also because the insights enable a contribution to cognitive science that is both richer and more justified than a model that is phenomenologically uninformed.

The failure to take phenomenology seriously has not only been a barrier for broad acceptance of naturalist models of the mind, but it has also restrained development of those very models. As Velleman stated, ‘the conceptual problem..... becomes intertwined with the phenomenological problem’ (1989, p. 32). Humans, for good reason, find it difficult to endorse a theory that seems inconsistent with their experience, and ignoring phenomenology has the cost of setting aside information that could be used to enhance our understanding of the mind:

To theorize about such phenomena purely in third-person mechanistic terms is to miss out on a whole way of understanding them. Crucially, at least phenomenal mental states are certainly such that we enjoy first-person encounter with them, which encounter can certainly afford us some insight into their nature (Kriegel 2013)

The cluster of experiences that accompany a decision-making process cannot be denied. The important question therefore becomes: What inferences are justified by these experiences?

Answering this question is particularly urgent as there are two ongoing and important philosophical debates in which a phenomenological gap, between a functional account of how agency works and everyday experience of how agency feels, is central to the disagreement. The first relates to the argument between libertarians
and naturalists about the possibility of doing otherwise. The second relates to the problem of the passive agent in mechanistic models of the mind. Pre-theoretically at least, phenomenology appears consistent with the libertarian view of alternative possibilities and the active view of agency. However, this seems at odds with a physical model in which the future is determined by the physical laws, and there is a causal chain from the belief-desire calculus of evaluation, via decisions and intentions, to action.

To believe that experience-driven philosophical intuitions are indefeasible would be an extreme position. Equally extreme would be an “error theory” position that no inferences from phenomenology are justified. I will argue that it is implausible that phenomenological and functional aspects of the mind are entirely decoupled. Instead, there is a correspondence relationship between phenomenology and function, but it is complex and nuanced. The nature of the complexity, when combined with the importance of the debates, is an argument for careful and detailed analysis rather than for neglect.

2. Phenomenological gaps in two philosophical debates

2.1 The puzzle of alternative possibilities.

Human agents make decisions and act in the world. For difficult decisions, they engage in extended deliberation and consider various options before finally coming down on one alternative over another. During deliberation it feels to an agent as though the future is open. She has what William James has called open and
ambiguous possibilities (2014). Even after deciding and acting, she may feel that she
could have done otherwise.

Philosophers have linked the experience of alternative possibilities to the experience
of freedom: ‘The experience of freedom, that is to say, the experience of alternative
possibilities, is built into the very structure of conscious voluntary intentional human
behaviour’ (Searle, 1984, p94). This raises the question of whether this experience is
revealing something important about the nature of reality. In particular, is the
phenomenology compatible with the idea that the future is determined by the laws of
nature and, if so, how?

Let naturalism be expressed as follows:

The natural laws determine which future world follows each current world⁴.

In a naturalist account, agential actions and decisions must be consistent with the
natural laws. This is possible if the decision-making processes is part of the
transmission mechanism through which the natural laws determine the future. Such
a position is set out by Ricouer:

The reasons which motivate my decision, the body which I am, even the
personal and historic conditions of my being are not simply external limitations
imposed upon me, but rather the organ in and through which I am actual.
(1966, translator’s introduction).

However, this model of the agent as a conduit through whom the natural laws
determine the future, only serves to increase the concern that there is a
phenomenological gap between experience and the naturalist account. If the natural

⁴ This is consistent with two statements: First, current human understanding of the physical laws is imperfect.
Second, the current best theories of the physical laws incorporate quantum indeterminacy.
laws manifest the single future through the agent’s decision-making process, then why does it feel to the agent that more than one future is possible? This seeming mismatch between naturalism and experience gives rise to the libertarian intuition that something is missing from the naturalist account. The libertarian intuition can be expressed in terms that make clear its opposition to naturalism.

The (anti-naturalist) libertarian intuition:

Agents can make a decision, ‘the making of which cannot be fully explained by the laws of nature and a prior state of the world.’ (Fischer⁵, 1998, p165).

Naturalists think that the libertarian intuition as expressed is mistaken⁶. In chapter 9 I discuss Christian List’s (2019) soft form of libertarianism, agential libertarianism. This is compatible with naturalism precisely because it does not endorse the anti-naturalist libertarian intuition.

For philosophers such as Searle, introspection is considered to support the possibility of doing otherwise:

Reflect very carefully on the character of the experiences you have as you engage in normal, everyday human actions… You will sense the possibility of alternative courses of action built into these experiences… that we could be doing something else right here and now, that is all other conditions remaining the same. This, I submit is the source of our own unshakeable conviction of our own free will⁷. (1984, p. 95)

---

⁵ Fischer was not endorsing this view, but setting it out clearly.
⁶ I will assume that the explanation can be post-hoc, so quantum indeterminacy does not make the libertarian intuition true. My understanding of naturalist libertarian Kane’s position is that, under my definition of explained, he would reject the libertarian intuition.
⁷ In chapter 8 I return to Searle’s position and his conclusions.
The fact that Searle makes this claim is notable because he is a naturalist. By doing so he illustrates the challenge that naturalists face: The underlying experience cannot be denied. Just as nature abhors a vacuum, naturalist should abhor this phenomenological gap.

Taken at face value, phenomenology and the replicability of scientific laws count as arguments for and against the possibility of doing otherwise respectively. Sims has called this clash between experience and determinism ‘experience-incompatibilism …. the view that determinism would need to be false for this “freedom” component of agentive experience to be accurate’ (2018, p 1314). There are three possible responses to this situation:

Naturalists could use historical precedent to argue that the scientific method has been a more reliable belief formation tool than intuition. Those persuaded by this argument might place little or no weight on the experience of freedom, assuming that, to the extent that it conflicts with naturalism, it is non-veridical.

Alternatively, philosophers might claim that the phenomenological gap indicates a deficiency in the naturalist model. They could argue that first-person subjective experience constitutes evidence that must be incorporated within a theory of agency, and that the current model is therefore incomplete. However, whilst this may well be true, the fact that the model can be enhanced does not necessarily imply that naturalism is false. It would be an extreme position to reject naturalism purely on the basis of phenomenology. After all, perceptual illusions exist and are understood, and much of physics is non-intuitive.

In section 3 I will argue against these two extremes and for a golden mean that involves phenomenological analysis. Some philosophers have already started down
this middle road, suggesting that although the experience cannot be denied it does not justify the anti-naturalist libertarian intuition. Instead, the agent’s experience may indicate that she could have done otherwise, but only if some element of her psychological and/or physical environment had been different. Advocates of this position refute, or add a qualifier to, the ‘all other conditions remaining the same” clause in Searle’s claim.

‘When we think of ourselves hypothetically as having acted otherwise than we did, we always suppose a difference in the antecedents: we picture ourselves having known something we did not know . . . or as having desired something . . . more or less than we did’ (Mill, 1872, p451)

A reply such as this sets up a constructive dialectic concerning the inferences that are justified by the phenomenology, leading to a more fine-grained understanding of phenomenological and functional aspects of the mind. For example, whilst Mill seems to be describing easy decisions where what to do becomes obvious, Searle’s description is more consistent with situations of indifference or uncertainty. Here the agent’s belief that she has alternative possibilities is underpinned by her experience of uncertainty about what to do, which can persist even as she is deciding. There is a phenomenological contrast between a decision made despite ongoing uncertainty and one in which uncertainty is resolved. It is the experience of uncertainty or low confidence that underpins James’ feeling of openness and ambiguity, and seems, at least pre-theoretically, to support the libertarian intuition.

Contemporary philosopher Holton agrees that the experience of making a difficult choice ‘is the primary ingredient in the experience of free will’ (2006)8, but his focus

---

8 In chapter 9 I will analyse Holton’s proposed solution, and suggest an alternative, which I believe will finally close the phenomenological gap.
is on the possibility of actively making a difficult decision. This highlights a second philosophical debate in which phenomenology plays a significant role: The debate about active agency and what might be called the puzzle of the disappearing agent. I turn to that now.

2.2 The puzzle of the disappearing agent

In the standard causal account of human action (Davidson, 1963), desires for ends, and beliefs about which actions are means to those ends, combine to cause intentions, and those intentions cause corresponding bodily movements. Central to this account is the idea that a decision is resolved through a motivational contest, and when an agent makes a choice, she chooses to do what she is most motivated to do. However, although this causal, motivational account ends with an intentional action, it seems to leave no room for active agency within the decision-making process itself. As Velleman critiques, “in this story...psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part” (1992). This seeming lack of active agency within the decision-making process is a problem for the motivational model because it clashes with folk intuitions, which are themselves grounded in a rich phenomenology. The mental events that lead to action feel as if they are of agential origin. There is a gap between the standard motivational account of how agency works, and everyday experience of how agency feels. Left open it justifies claims that the account is currently incomplete, and speculation that it is fundamentally deficient.
Active agency has a complex and shifting phenomenology, quite distinct from passivity. Being an active agent doesn’t feel like being a spectator in Velleman’s decision-making arena, or even being the arena itself. It feels like being a player on the field, exerting influence over the events taking place. Contemporary philosophers (Mele, 1995, Holton, 2009, Bayne and Levy, 2006, Sims 2019) have taken this phenomenology seriously, focusing particularly on the insight that paradigmatically agential acts, such as resisting temptation or making a difficult choice, are felt as effortful. With respect to resisting temptation, Holton states that “it certainly doesn't feel as though ........ one is simply letting whichever is the stronger of one's desires or intentions have its way. It rather feels as though one is actively doing something, something that requires effort.” (2009, p121). Similarly, with respect to making a choice, Mele states that there is ‘an experience of its being “up to me” which of these I do’ (Mele 1995, p. 135), and Sims argues that it feels like ‘it is ultimately me that settles the matter.’ (2019, p1314). I will accept these descriptions of the phenomenology. The debate is over the inferences that are justified from such an experience.

In contrast to the account of decision-making as an unbiased, symmetrical competition between motivational forces, the described phenomenology suggests that an active agent can break the symmetry. I will call any asymmetrical account in which an agent can ‘rise above the desires to which one is subject, and to take the reins in one’s own hands’ (Wallace, 1999, p234), “interventionist”. Such a view would imply that what Holton’s agent is doing is intervening in the decision-making process, to bias the decision in favour of the virtuous desire. Velleman’ describes Frankfurt’s model of agency as motivational, but interventionist: ‘The agent's role, according to Frankfurt, is to reflect on the motives competing for governance of his
behaviour, and to determine the outcome of the competition, by taking sides with some of his motives rather than others' (Velleman, 1992, p476). The suggestion is that an active agent takes sides, intervenes, and biases the decision, thereby breaking the symmetry of the motivational account.

The phenomenology of effort is seen as providing (defeasible) support for an asymmetric view. For example, Bayne and Levy describe the experience of effort as "an experience of the self as a source of force", without which "we would not experience ourselves as allied with some desires and opposed to others" (2006, p63). It might be concluded, therefore, that there is something fundamental missing from the motivational account. If the phenomenology implies that agential acts are asymmetric, then a symmetric account will fail. However, whether the phenomenology does imply such a thing depends critically upon the specifics of the relationship between the mental attitudes that functionally drive decisions and the phenomenology of those attitudes.

In the motivational account, decisions are driven by motives. And motives are famously felt as "pulls" or "grabs". Thus, an obvious question is how closely the conative phenomenology of a motive corresponds to its functional impact. Possible answers lay on a spectrum. At one extreme is a consistent one-to-one correspondence. If this were true then asymmetric phenomenology would imply asymmetric functionality, and interventionist beliefs would be justified. At the other extreme there is no relation between function and phenomenology. They are entirely independent. Adopting this view would imply that phenomenological analysis of motives tells you nothing about the underlying symmetry, or lack-thereof, within the decision-making process. Between these two extremes, however, is the middle ground in which divergences between phenomenology and function exist, but they
are partial and rule-governed rather than systemic. In this case phenomenological analysis does provide valuable information, but only to those who have cracked the code. Although a simple one-to-one relationship might be easier to understand, the middle ground is the most interesting. The opportunity to crack the code, by resolving what I will call the puzzle of relative strength, is an opportunity to deepen an understanding of agency. Tantalizingly, if the relationship takes a specific form the phenomenology could actually support the (non-interventionist) motivational account. I will confirm that this is indeed the case in chapter 9.

If true, this would indicate that philosophers of agency have been working on the problem from the wrong direction. The agentive gap should be closed through fine-grained analysis of conative phenomenology rather than by adding cognitive architectural complexity to explain the asymmetry.

I hope I have convinced the reader that there are phenomenological gaps at the heart of two of the major debates in the philosophy of agency. In the following section I will defend the view that progress can and should be made through deep phenomenological analysis.

3. The value of deep phenomenological analysis

The existence of a phenomenological gap raises the question of how to respond. My answer is “with fine-grained phenomenological analysis”. In opposition to this suggestion, I present two caricatured positions: Principled neglecters, and seemings accepters. I am not suggesting that any serious philosopher endorses either position.
Principled neglecters, explicitly adopt an error theory of phenomenology. This is the strong theoretical position that there is no value in phenomenological analysis because experience and reality are untethered. In this view, phenomenological analysis is a fool’s game, which will lead to erroneous conclusions. I will argue against this position in the following section.

Seemings accepters, on the other hand, make strong inferences about how things are from how they seem to be. Their experience of alternative possibilities and active agency leads to a belief that they are active agents with alternative possibilities, just as an experience of rain makes them believe it is raining. They believe that experience based philosophical intuitions are indefeasible. As I will discuss in section 3.2, this position immediately runs into trouble as it is widely understood that seemings are defeasible.

Through analysis of error theory and seemings-based justification I will show that although the extreme positions should be rejected, they both contain valuable insights. Error theorists are right that there is not a simple one-to-one relationship between phenomenology and function, and seemings accepters are right that defeaters that can’t explain the phenomenology are unpersuasive. I will therefore argue for an Aristotelian golden mean between the excessive scepticism of error theory and the deficient acceptance of seemings.

3.1 Error theory and correspondence

In general terms, error theory is the belief that when we make judgements based on experience, we systematically fall into error. Error theories may be broad, implying
that philosophers should not rely on introspection in any circumstance, or narrow, in that only some experience is unreliable. My analysis will focus narrowly on the conative and cognitive phenomenology which is felt in the extended process of deliberation, decision-making, and intention implementation. This is consistent with the model of the mind that I will develop throughout the thesis, at the heart of which are two elements: Motives and levels of confidence. Both have a central functional role in decision making, and both are felt.

I give two examples of how an agent is helped by a correspondence between phenomenology and function, and how she would be harmed if error theory were true. Both are instances in which a feeling triggers a beneficial action. The first illustrates the link between self-control and the conative phenomenology of motivational pull. The second reveals the benefits for learning of the cognitive phenomenology of uncertainty reduction:

Self-control: Chloe has an exam tomorrow. It would be helpful if she did some last-minute preparation in the library and very harmful if she went out drinking with her friends. Unfortunately, the pub is next to the library and it is Thursday night and she feels a strong and growing desire for a drink. She decides instead to go for a walk. This serves as a control action, ensuring that she stays away from the temptations of the pub. Chloe has exerted (diachronic) self-control and avoided disaster. But imagine Error Chloe, whose feelings of desire for a drink do not correspond to any functional want. Error Chole also implements the control action and goes for a walk, but for her, this is driven by an error. She never had any functional motivation towards drinking. She misses out on her final preparation for no good reason, misdirected by her erroneous, but growing feeling of desire.
Learning: Charlotte needs to study for an exam. She heads to the library, where there are two books. One is dense but unhelpful. The other resolves a number of questions with which she has been struggling. As she reads the second book she gets a pleasurable feeling of uncertainty reduction. She is happy to spend the day learning. But imagine Error Charlotte whose pleasurable feelings of uncertainty reduction accompany the first book rather than the second. She is drawn to spend time reading the dense unhelpful book. She feels she is learning something, but she is actually wasting her time, misdirected by her erroneous, but pleasurable feeling of uncertainty reduction.

Error Chloe and Error Charlotte are misdirected by their phenomenology. Not only are they endlessly surprised by, and worryingly alienated from, their own actions, but they underperform their potential. For naturalists, the error world is even less compatible with adaptive fitness because the cognitive infrastructure required to sustain two separate measures of strength, functional and phenomenological, is much more expensive than that required to sustain just one. Is it plausible that the brain spends its limited resources producing experiences that are systematically unrelated to the functional process and harm fitness? I think the answer is no. The idea that evolution would stabilise such fundamental and systematic self-harm is almost an oxymoron. Baumeister et al. have argued that ‘it is implausible that the metabolic costs inherent to conscious processes would have evolved in humans without any adaptive benefits’ (2018, p2). I am making the weaker claim that it is implausible that conative and cognitive phenomenology would have evolved if they were adaptively harmful.

I am not arguing that things are always as they seem. Motives are not always felt and feelings of uncertainty reduction might indicate coherence rather than accuracy.
There may even be specific rule-governed instances where false beliefs are adaptive. I am merely rejecting the opposite extreme, that there is no relation between function and phenomenology. My point is that, for conative and cognitive phenomenology at least, assuming the existence of a correspondence relationship and interrogating its complexity has value. An error theory for this phenomenology is implausible, and proceeding on such an assumption has costs.

Whilst the premise that phenomenological intensity and functional strength are entirely separate and distinct might be a strawman, the rules that govern the correspondence relationship between functional and phenomenological properties have not yet been elucidated. This is an extraordinary gap in philosophical understanding of the mind. Given that human agents do use phenomenological data to make decisions which drive behaviour, then surely, we should desire a clear understanding of the rules which govern when that data is, and is not, reliable.

3.2 Seeming and persuasive defeaters

In this section I will argue that, if an agent has a belief that is grounded in experience, then, although that belief is defeasible, any defeater that leaves an unexplained phenomenological gap is unpersuasive.

My starting point is a field of enquiry within epistemology that formalises the link between phenomenology and belief formation through what is called seemings-based justification. It relies on the idea that in normal environments there is a correspondence between how things seem to be and how they are, so beliefs based
on how things seem to be are rational. If it seems to Zac that it is raining, then in normal circumstances he is justified in believing that it is raining.

More formally, when an agent has an experience as-of some proposition, $p$, being true, $p$ seems to her to be true. And whilst such a seeming doesn’t guarantee the truth of $p$, it is generally considered to provide at least some prima facie justification for believing it. Proposition $p$ is defeasible, but if $p$ seems to be true then the burden of proof lies upon anyone who wants to deny $p$. The sceptic must find a defeater.

Brogaard sets out a position on seemings, which she calls sensible dogmatism:

> If it seems to S as if $p$ and the seeming is grounded in the content of S’s perceptual, introspective, or memory-related experience, then, in the absence of defeaters, S thereby has at least some degree of justification for believing that $p$ (2013, p279)

There are two steps in the process. First an agent has an experience, which provokes in the agent a propositional attitude of the form it seems to me that $p$.

Second, the agent forms a belief that $p$, justified, in the absence of defeaters, by the seeming. Although there is some debate about the interaction between dogmatism and the sceptical “brain-in-a-vat” hypothesis, I set that aside. For readers who view dogmatism to be true only if the sceptical hypothesis is false, please feel free to make this analysis conditional on that assumption.

A defeater is any proposition, or set of propositions, which, when included in the agent’s evidence set, makes a belief in $p$ unjustified. For example, if an agent looks at a wall and has a phenomenological experience of redness, then other things equal she will believe that the wall is red. The experience is undeniable, but the belief may
be mistaken. If it is explained to her that the wall is bathed in a strong red light, and that switching to a yellow light would make the wall look yellow, then she loses justification for her belief. Knowledge of the coloured light acts as a defeater.

As well as perceptual, introspective and memory-based seemings, Brogaard introduces a fourth class, intellectual seemings, which include philosophical intuitions. These are ‘seemings that result from implicit or explicit arm-chair reasoning, where armchair reasoning is reasoning that involves both a priori principles and past experience’ (2014, p387). In this case, an agent has a justified belief E and, through explicit reasoning and/or because of implicit background beliefs, it seems to her that, in light of E, P must be (or is likely to be) the case. The belief P is justified if: (i) the belief E is justified; (ii) the belief E causes the belief P via the intellectual seeming; and (iii) there are no defeaters (Huemer, 2016).

Seeming-intuitively-to-be-correct is also phenomenologically grounded. Plantinga describes it as ‘that peculiar form of phenomenology with which we all well acquainted, but which I can’t describe in any way other than as the phenomenology that goes with seeing that such a proposition is true.’ (1993, p105-106) To Brogaard, ‘the feeling is an urge (or inclination) to believe merely based on understanding that is so intense that we have a hard time envisaging that others may not feel the same way’ (2014, p388-389). Intuitions and seemings are relied upon in philosophy as defeasible evidence for a claim. As De Cruz states: ‘the intuiter hopes that the philosophical claims she makes will be accepted on the basis of shared intellectual seemings.’(2015, p247). However, these intuitions, which are founded on associative ‘maturationally natural’ (ibid, p240) process can be wrong. For example, intuitive
physics yields predictions that deviate from even Newtonian physics (McCloskey et al. 1983), never mind quantum mechanics.

Sensible dogmatism grants that beliefs based on seemings are defeasible. Despite this, it can be surprisingly hard to find a persuasive defeater. This is because the experience which grounds the belief is not in dispute, so any defeater that leaves the experience unexplained leaves a phenomenological gap. More practically, seemings are recalcitrant and human agents do tend to rely on seemings-based justification. They do so because, to apply Alston’s framing, seemings-based justification is a reliable belief-formation process: It ‘yield[s] a high proportion of truths over a wide range of situations of the sort we typically encounter’ (1995, p10). It would be extremely inefficient, and arguably psychologically impossible, to live a normal life as a full seemings sceptic. Of course, this fact doesn’t rule out the existence of systemic exceptions, particularly if those exceptions are adaptive. Nor does it imply that it is epistemically virtuous to trust seemings-based justification in all situations. But it does explain why evolution has stabilised seemings-based justification as a belief-formation habit. Given this, rather than rejecting seemings-based justification in toto, it is more evolutionarily plausible and psychologically acceptable to think of seemings as generally reliable grounds for coarse grained beliefs in normal environments, but less reliable grounds for fine grained beliefs in novel or extreme environments.

To be truly epistemically virtuous, therefore, an agent should conduct fine-grained analysis to discover the circumstances in which a belief-formation process will fail. If it is consistently unreliable in a certain abnormal environment, then in that environment it should not be relied upon. Through interrogating the reasons for the unreliability in a specific situation, an agent can learn the limitations of its evidential
support. She can then efficiently navigate normal environments whilst exercising caution or switching to an alternative belief formation process, in abnormal environments.

The approach I advocate is not just theoretical. Human agents can and do resist forming beliefs based on false seemings. An example is false perceptual-appearances, otherwise known as perceptual illusions. In fact, the ‘Problem of Known Illusions’ is a central argument against the idea that seemings are identical to beliefs. Tucker illustrates the argument with an example: “It may seem to me that the half-submerged stick is bent even though I don’t believe that it is” (2013, p4). The belief that “a stick bends when lowered into water” can be undermined either through an understanding that the physical forces required to bend a stick are not present in water dipping, or through an understanding that light refraction makes water-impacted perceptual seemings unreliable. The former addresses the facts (the stick remains straight) whereas the latter addresses the correspondence relationship between the facts and the experience of the agent (refraction bends light, which impacts the appearance of half-submerged sticks). The former without the latter leaves the experience unexplained. It leaves a phenomenological gap.

The example of light refraction illustrates the fact that, in order to make accurate inferences from experience, one must understand the correspondence relationship between how things are and how they seem to be. Seemings-based justification works best in cases where there is a one-to-one correspondence. However, as proponents of sensible dogmatism understand, this is not always the case. Casting doubt on one-to-one correspondence may be helpful as a negative argument against a seeming, but discovering and elucidating an alternative correspondence
relationship is more persuasive. In some situations, an agent can show that a claim must be false precisely because it appears to be true. For example, the statements “that stick must be bent because it looks straight” and “that wall must be red because it looks purple”, may sound odd, but they are accurate if the stick is half submerged in water and the wall is bathed in a blue light.

To summarise, whenever anyone argues against a seemings-based claim \( p \), they should expect to be asked a simple question: “If \( p \) is false why does \( p \) seem to be true?” If unanswered, this question creates a worry that it is the defeater that is flawed. I will call a defeater that leaves this question unanswered, an unpersuasive defeater.

**Conclusion**

I hope that I have persuaded readers that phenomenological analysis has value, not just in theory but also specifically with respect to the ongoing libertarianism vs naturalism and active vs passive debates. To be persuasive, any model of human decision-making must be experience compatible. Such an account should be able to explain not just the underlying functional reality, but also the correspondence relationship between how things are and how they seem to be.

As Kriegel states: ‘understanding the will depends in the first instance on phenomenological analysis of conscious, experiential exercise of the will’ (2013, p540). In this thesis, I focus on building an account of the decision-making process that gets the phenomenology right, and explore the implications for our understanding of agency.
Combining phenomenological and functional analysis is necessary because a model of the world that doesn’t fully capture objective interactions and subjective experience is incomplete. It is also efficient, as the different tools and different perspectives of the two analytic styles can be combined to tease out solutions to knotty philosophical puzzles. Historical biases towards one style or another have led to a dialectic stalemate and to theories which suffer from a phenomenological gap. Combining insights from both areas enable a contribution to cognitive science that is both richer and more justified than an investigation that sticks to one side of the Cartesian cut.
(ii)  

Chapter Summary

In this thesis I develop an experience-compatible model of agency based on motives and levels of confidence. I do so by engaging in a combination of phenomenological and functional analysis over the course of 9 chapters. Here I offer a preview of each chapter.

Part I

Part I of the thesis involves two foundational chapters. The first sets the scene, introducing and interrogating the concepts of conative value, doxastic and conative attitudes to content, and the motivational strength idea. In the second I make my first major claim: That intentions are merely uninhibited motives.

In Chapter 1 I conduct some detailed groundwork, laying the foundations for later analysis and argument. I aim to provide conceptual clarity to terminology that will be used throughout the thesis.

The simple idea that pervades this thesis is that, whenever an agent makes a decision, she will choose to do what she is most motivated to do. This is the motivational strength idea (Mele, 2003). However, it is impossible to interrogate the idea without first knowing the answer to a series of questions: What is a motive, and how does it differ from a belief? What gives a motive its directional force? How do conflicting motives interact with each other, both before and after a decision is
made? And how is a decision made? After this chapter the reader will have the answer to these questions and will be equipped to engage with the body of the thesis. I begin with an analysis of what I will call conative value. Conative values are dual-faceted: They are values in the sense that they are predictions of the subjective, context-dependent “sustainability” value of any considered action-plan, and they are conative in the sense that they motivate the agent towards implementing action-plans of value.

Humans are differentiated from other animals in part because of the incredible complexity of our conative evaluative map. This results in a crowded conative environment in which motivational conflict is frequent and significant. The management of motivational conflict is central to what it means to be a human agent and perhaps the most important driver of the phenomenology of agency. But at heart, the question that is being asked and answered in a continuous loop, remains the same: How do I maximise value and therefore sustainability? To understand agency is to understand the process by which an agent answers that question.

Having introduced the concept of conative value, I then conduct a detailed comparative analysis of conative attitudes, such as motives, and doxastic attitudes, such as beliefs. This reveals two major differences:

The most commonly highlighted difference between conative and doxastic attitudes is their direction of fit. Conative attitudes move an agent to act on the world, whereas doxastic attitudes are moved by observations of the world. Conative value is not inert but represents the strength of a motivational force (Velleman, 1992).

The second, much less discussed but equally important, difference is that, whilst doxastic attitudes are subject to a coherence requirement, conative attitudes are not.
Conflicting motives persist and compete in motivational contests to drive action. To manage all unchosen motives during and after a motivational contest, a mechanism must be in place to block transmission from readiness to action. Motives must be *inhibitable*. Inhibition of motives, when implemented in conative architecture has three functional implications: (i) agents are not compelled to act on every motive; (ii) agents do not try to simultaneously satisfy two directly conflicting motives; and (iii) for as long as an agent is uncertain about what to do, all competing alternatives are inhibited. Whilst the inhibitability of conative states is a necessary, perhaps obvious, consequence of conflict persistence, it will turn out to have important implications that will be explored throughout this thesis. If motives were not inhibitable we would be very different agents.

After discussing the content of propositional attitudes, with particular focus on the fact that the content of a conative attitude is a plan-of-action which incorporates a schedule, I turn to metacognition. Particularly relevant to the thesis will be mind-directed meta-conative attitudes, which treat elements of the cognitive machinery as objects to be manipulated (Hieronymi, 2009). For example, intentional synchronic self-control will turn out to be dependent on desires about desires, whilst the feeling of effort when making a difficult decision will indicate that the agent is resisting the desire to prolong deliberation.

Finally, I analyse models of the decision-making process. This is the first time I introduce the motivation strength idea, that I will defend throughout the thesis.

The Motivational Strength Idea:

When making a choice between A-ing and ¬(A-ing), the agent will choose to A if, and only if, at the moment the choice is made, the strength of her
motivation towards A-ing is greater than the strength of her motivation towards \( \neg \text{(A-ing)} \).

This model of decision-making is symmetrical or unbiased, because all activated motives are treated equally, competing on a level playing field. The only determining factor in a motivational contest is motivational force.

This idea is an important element of the philosophical puzzles I analyse and resolve in chapters 3-9. While others have rejected or adjusted or complicated the idea, I will remain true to its simplicity. Doing so will force me to be flexible elsewhere, and question other assumptions, which will pay dividends in the final model.

In Chapter 2 I make the first major argument of the thesis. I propose and defend the simple, but controversial, view that intentions are merely uninhibited motives.

This is necessary for three reasons. First, any analysis of intentional agency must engage with the question of what, if anything, is special about an intention. By showing that an intention is merely an uninhibited motive, I simplify the model, and the writing. Second, the most detailed, and widely accepted, theory of intentions imbues them with intrinsic stability (Bratman, 1999). If this is true, then it could significantly complicate the motivational strength idea. By denying intrinsic stability I remove that complication. Finally, other phenomenologically-informed philosophers such as Holton (2009b) have identified intentions as an important explanans for conative phenomenology. By refocusing on inhibition, I begin to set up my alternative account.

I highlight three features of the post-decision landscape that, whilst extrinsic to the attitude, explain the shifting stability of a motivational attitude. These relate to plan
implementation, meta-evaluative consistency, and the implementation mindset. The analysis confirms that Bratman's three key features of intentions - commitment, stability, and coherence - are explained by extrinsic factors. They are not intrinsic properties of the intending attitude. I therefore conclude that intentions are not sui generis attitudes, but are merely uninhibited motives.

The analysis of this chapter emphasises the importance of inhibition to a functional model of intentional action. The following chapters will reveal that it also plays a central role in conative phenomenology.

**Part II**

In Part II I engage in two significant pieces of phenomenological analysis, relating to the feeling of motivational pull and the feeling of effort, and one piece of functional analysis, relating to the possibility of intentional synchronic self-control. In doing so, I consider the motivational strength idea in the context of two philosophical puzzles. By resolving these successfully, using phenomenological and functional analysis respectively, I further the project of building an experience compatible model of the mind. The model that emerges is undivided, but hierarchical, with inhibitable motives at its core. In the final chapter I consider whether the phenomenology of alienation is a problem for the account.

In **Chapter 3**, I analyse the conative phenomenology of motivational pull. In doing so, I address and resolve the philosophical puzzle of relative strength, and elucidate the correspondence relationship between the phenomenology and function of conative attitudes (Mele, 2003, Kriegel 2015). This is important, because only once
the correspondence relationship is known can the full power of phenomenological
data be unleashed, to interrogate and enhance functional theories of the mind, and
to generate persuasive defeaters to seemings-based libertarian intuitions.

The puzzle of relative strength can be set out as three seemingly inconsistent claims:

1. When making a choice, an agent always chooses to do what she is most
   motivated to do (The motivational strength idea)
2. The phenomenological intensity of a motive is a non-conceptual
   representation of its functional strength (Full Correspondence)
3. The phenomenological intensity of unchosen motives can be greater than
   those of chosen motives. (Experience)

I argue that premise 2 is true if and only if a certain condition is met. I call this
solution conditional correspondence. The specific proposal I defend can be
expressed as follows.

The Conditional Correspondence Relationship:

\[ \text{The functional strength of a motive is revealed by its phenomenological} \]
\[ \text{intensity if and only if the motive is inhibited.} \]

Three clarifications are necessary to forestall objections:

First, for conative phenomenology to be entirely absent there needs to be a perfect
alignment between the action that is implemented and the plan-of-action that is the
content of the desiring attitude. This is consistent with the idea that the content of a
motive incorporates a schedule. If an agent is behind schedule, acting to get back on
schedule feels urgent.
Second, intensity is only one aspect of the conative phenomenology of motivational pull. The other element is directionality (Gibson, 1979, McClelland, 2019). Although intensity and directionality are conceptually distinct, reflecting the strength and plan-content of the motive respectively, it may be practically difficult for an agent to introspectively isolate the intensity of the attitude from other aspects of the conative phenomenology. There is, however, one case in which content and intensity are linked: When the content is null. This suggests a possible functional reason for the conditional correspondence relationship. Motivational strength is a measure of the force required to change the world so that it aligns with a plan, but if the world is already aligned with the plan, then there is nothing to do. Even if a plan is not yet fully implemented, if it is on schedule and there is no resistance, then no motivational force is necessary.

Third, conative phenomenology is only one element of the broader phenomenological landscape.

If the conditional correspondence relation is correct, then this implies a phenomenological inversion. Whilst functional analysis celebrates the victory of the strongest motive (indicating only that it was stronger than its competition), conative phenomenology reveals the strength of its weaker cousin, the inhibited motive. Humans are flexible agents, motivated but not compelled by inhibitable conative attitudes (Bayne, 2013), and conative phenomenology is evidence of this lack of compulsion. However, one must be cautious. The phenomenology does not imply that a motive to do otherwise is defeated by anything other than a conative force. It does not imply the existence of a Cartesian homunculus that is doing the resisting. Rather, the analysis of the chapter is consistent with, and supportive of, the motivational strength idea. Agential action is decided in a motivational contest.
In Chapter 4 I analyse what is, in my opinion, the most important but least well understood agential phenomenology, the experience of effort (Bermúdez, 2023).

It is necessary for me to have a detailed account of the feeling of effort because in chapter 9 I will attempt to solve an outstanding question in philosophy of the mind: Is making a difficult decision an action and why does it feel effortful? The fine-grained understanding of feelings of effort developed in this chapter will equip me to finally solve that problem and close the phenomenological gap.

I argue that, for all subspecies of effort (Massin, 2017), the feeling of effort represents-as-costly a trying, where cost is understood in the technical conative sense (Kriegel, 2015, 2023). Motives represent as valuable their contents and, as cost-benefit computations are symmetrical, the value associated with a particular course of action is an opportunity cost of any opposing course of action. That opportunity cost of a trying is equal to the conative value of doing otherwise, and is revealed to the agent through conative phenomenology (Kurzban et al., 2013).

This theory is perhaps most controversial with respect to muscular effort, where I rely on the argument that external resistance causes, not the feeling of effort directly, but an impulse to stop trying. That impulse is a motive, which is an input into the cost-benefit computation. Fatigue or pain or disappointing progress can lead to stopping behaviour, but only indirectly via a stopping-impulse. This mediation is necessary to allow agents to flexibly respond to the environment, mirroring the fact that hunger and eating are mediated by the desire to eat.

This analysis of effort suggests two insights which will be carried forward within the thesis:
First, care must be taken when using phrases such as “agents are biased towards effort avoidance”. Taken at face value, these statements risk getting the directionality wrong. A feeling of effort reveals to the agent that a chosen action is controversial and involves costs, but it does not add a further cost that must be synchronically incorporated and resisted. Does this mean that the phenomenology serves no purpose? No. Feelings of effort are highly valuable for a diachronic agent, who can remember previous feelings of effort and anticipate future effort. She can monitor changes in the level of effort over time and/or relative to predictions and search for alternative (less effortful) means to the same end. Feelings of effort support flexible, adaptive behaviour diachronically, even if, as non-conceptual representations of input costs, they cannot directly and synchronically undermine a chosen course of action.

More parochially, the feeling of effort is highly valuable for philosophers in that it indicates the presence of a conflict between motives. As a general point, this supports the motivational strength idea and confirms that feelings of effort are sufficient (albeit not necessary) for intentional action attributions. Specifically, it implies that there are two motives for each feeling of effort, one towards and one against the chosen action. For each feeling of effort, philosophers of mind should aim to identify those motives. If an action is effortful then a motive against that action must exist, and if theories of intentional action cannot explicate what that motive is, then they are incomplete. The most obvious case of this incompleteness is found in the making of a difficult choice. Answering the question of how a choice can be an act has been described by Vierkant (2022) as the agency problem of compatibilism. A fine-grained understanding of the feeling of effort will equip me to finally solve that problem in chapter 9.
In Chapter 5 I return to my defence of the motivational strength idea. In particular, I focus on the possibility of intentional acts of synchronic self-control, where an agent acts to shift her own current balance of motivation. This creates what Altehenger (2021, responding to Mele, 1987, 1997) calls the mismatch problem, which can be expressed in terms of a question: How can the three statements below all be true?

1. When making a choice, an agent always choses to do what she is most motivated to do (The motivational strength idea)
2. An agent can intentionally shift her current motivational balance, changing what she most wants to do (Intentional synchronic self-control)
3. If an agent desires an end and is considering an act that will harm that end, then the motivational force of her desire will be transmitted in full into desire to refrain from the harmful act. (Instrumental rationality)

Although the analysis of this chapter is mainly functional rather than phenomenological, support for 2 is generally based on phenomenology. Intentional synchronic self-control is the paradigmatic effort of will, which is felt.

This puzzle represents possibly the greatest challenge to the motivational strength idea. Indeed, philosophers such as Sripada (2014, 2020) and Altehenger have argued that the existence of intentional synchronic self-control implies that the mind is divided and/or that an agent can overrule her strongest motives by using a separate and distinct source of “willpower”. I deny that conclusion and argue that intentional synchronic self-control follows naturally from the existence of metacognitive motives, acting in accordance with the motivational strength idea.
I propose a simple hierarchical solution. Statements 1-3 can all be true because the question of whether to engage in self-control is metacognitive. When the question of self-control is raised, metacognitive motives become relevant. These conative attitudes towards or against practical desires can be specific, such as the desire to suppress the desire to eat, or general, such as the desire to always remain in control. If strong enough, they can initiate a control action which will attempt to shift the balance of practical desires even on the assumption of full instrumental rationality.

There are two motivational contests, one for the control decision and another for the practical decision, with different inputs. The metacognitive motive is (usually) only relevant for the former. If that is true then, if the motivational force of the metacognitive motive is strong enough, an agent’s motivational balance can favour, simultaneously, a wayward action and a control action.

I contrast my proposal to Altehenger’s divided-mind proposal, which relies on the idea of motivational partitioning (2021). After detailed analysis, I reject the idea of a partition between so-called deliberative and impulsive desires, and argue that the most promising version of the divided-mind model is one in which the motivational partition divides high-order and lower-order desires. I then compare the divided hierarchical model with my unified hierarchical model in which higher-order and lower-order desires interact with each other. I conclude that the unified hierarchical model is simpler, more phylogenetically plausible, and more consistent with currently available empirical data.

In Chapter 6 I consider the possibility that an agent may feel alienated from a force that moves her to act: She may divide the forces that move her into motives with
which she identifies, and motives from which she is alienated (Frankfurt, 1988). This possibility of asymmetry raises two worries about the symmetrical motivational force idea, one epistemological and the other ontological.

The epistemological worry: The experience of alienation indicates that the proposed model of intentional action is flawed. Does the possibility of alien motives indicate that any symmetrical model must be wrong? To answer this question, I need to first understand what it is that makes an agent categorise a motive as alien. By analysing potential solutions to what I will call the “subjective categorisation problem”, I can deepen my psychological understanding and validate or invalidate, or perhaps enhance, my model. I will find that both judgement-based and conative phenomenology-based solutions to the categorisation problem are consistent with my model. The possibility of a non-conative phenomenology of alienation, exemplified by a divergence between liking and wanting in addiction (Holton and Berridge, 2013), is most interesting in that it helps address some outstanding questions. I conclude that there is nothing within the experience of alienation which undermines the motivational strength idea.

The ontological worry: The phenomenology of alienation indicates that although my model successfully explains intentional action, some of that action is not agential action. This second worry is that I have built an accurate model, but of the wrong thing. Perhaps I have cast the net too wide, catching all motives rather than the motives of the agent. I provide three arguments against a hard version of this worry, in which agential motives and alien motives are of fundamentally different natural-psychological kinds: the argument from attitude formation processes (Plantinga,
1993); the argument from addiction (Sripada, 2022); and the argument from degrees of alienation (McConnell and Golova, 2003). A softer version remains, however, in which a motive is called alien (or half alien) if it is inauthentic, somehow defined, or fails to satisfy a norm of agency. This would be consistent with an “incremental” view, in which agency captures only what is distinctive about humans. I will argue that this view is not without problems and that, for the phenomenologically informed project in which I am engaged, an “inclusive” view of intentional agency is more appropriate.

Therefore, I will endorse what the standard conception of agency, which ‘consists in the performance of intentional action’ (Schlosser, 2019).

Human agency, is messy, conflicted and sometime inauthentic:

Do I contradict myself?

Very well then I contradict myself,

(I am large, I contain multitudes.)

(Whitman, 1892)

Part III

Part III consists of three chapters. By the end of this section, I aim to have a) completed my project of developing an experience-compatible model of the mind; b) shown that the phenomenology of the decision-making process does not justify
libertarian beliefs; and c) solved the agency problem of compatibilism. Making a
difficult decision is an action.

The simple model of decision making is based on the following two ideas:

*When an agent chooses, she will always choose to do what she is most* 
motivated to do (*The motivational strength idea*)

*An agent only makes a choice when she reaches a threshold of confidence in* 
*her decision* (*Confidence*)

In part I and II I focused mainly on conative attitudes and phenomenology. In the first
two chapters of part III, I will switch gear and conduct a phenomenological and 
functional analysis of (levels of) confidence. This will prepare the ground for bringing 
both elements together and filling in the final details of the unified, hierarchical model 
of the mind, in chapter 9.

**Chapter 7** mirrors chapter 1 to the extent that it aims to provide conceptual clarity to 
terminology that will be used in Part III. I define, and analyse the implications of, risk- 
of-error, evidence, extended deliberation, levels of confidence and decision 
thresholds. Some of the analysis is quite technical but the story goes as follows:

Conative values are modelled estimates of the extent to which an action satisfies 
care-based norms. Due to cognitive and temporal constraints these conative value 
estimates include a risk-of-error. This is the risk that, although the current evidence 
set is consistent with choosing an action A, a full evidence set would be consistent 
with the same agent choosing an alternative action. This risk can be reduced through
extended deliberation. When an agent is deciding between two action-plans, a
decision is easy if the risk-of-error is negligible and hard if it is high.

The inverse of the risk of error is represented in psychological architecture as the
level of confidence in a choice. As Kriegel states, we should ‘construe confidence not
as a modification of judgements, but as an independent second-order mental state
about judgements’ (2015, p67). When confidence is low this motivates an agent to
deliberate further: locating, gathering and incorporating relevant evidence. This
deliberation will lead to updated conative values and increased precision. I argue
that a decision is made when the level of confidence exceeds a certain decision-
threshold (Lee, Daunizeau and Pezzula, 2023). When an agent’s level of confidence
is below that threshold, she is uncertain about what to do and motivated to deliberate
further. When it crosses the threshold, she makes a decision (Burnston, in
submission).

Being uncertain is not an “all-or-nothing” variable within the agent’s psychological
architecture. The psychological variable is the agent’s level of confidence, which is
continuous and “accumulates” as the risk of error declines. An agent is uncertain
about what to do until the accumulating level of confidence crosses a decision-
threshold (Teodorescu and Usher, 2013, Moss, 2018).

The concepts of uncertainty and levels of confidence are important for the thesis.
The introductory clause of the motivational strength idea and extended deliberation
only exist because reducing the risk-or-error has value. Without extended
deliberation there would be no possibility of managerial actions such as acts of
intentional synchronic self-control control, or the act of making a decision.
In addition, feelings of uncertainty about what to do seem to support the libertarian intuition. By showing that feelings of uncertainty are feelings of low confidence, I set myself up to argue, in chapter 8, that the cognitive phenomenology of uncertainty is not the phenomenology of alternative possibilities, but of hunger for information.

Chapter 8 comprises of two sections. In section 1 I conduct a phenomenological analysis of the feelings of confidence experienced through an extended process of deliberation. I identify three different feelings, only the first of which has been fully incorporated within philosophy of agency:

The first is the aversive feeling of the presence of low confidence. This is relatively well understood both as indication of epistemic risk, but also as a trigger which activates curiosity (Dokic, 2012, Proust, 2013). Second is the pleasurable feeling of a reduction in uncertainty. This is not just an agential ability to observe levels of uncertainty declining over time but the phenomenology of the rate of change (or first derivative) of uncertainty (Nguyen, 2020, Csikszentmihalyi, 1988, Kiverstein et al., 2019). A steep fall in uncertainty is felt as pleasurable, whereas a steep rise is aversive. Third is the pleasurable feeling of trust and confidence (Topolinski and Reber, 2010). When an agent feels fully confident in her decision, her state of mind is consistent with only that decision, and deliberation ends (Kriegel, 2015).

In section 2, I draw on the findings of section 1 and chapter 7 to present three arguments against libertarianism. In the first I use Dennett’s (2003) framework of nearby possible worlds to show that “psychological underdetermination” can cause feelings of psychological openness and the belief that one could have done otherwise. Here is my definition of psychological underdetermination:
An agent’s understanding is never fine grained enough to describe just one unique world. An agent’s subjective set of possible worlds at any moment in time, her set of psychologically possible worlds, is the set which contains all worlds consistent with her then understanding.

Psychological underdetermination suggests that rather than supporting anti-naturalist libertarianism, the experience of alternative possibilities is a function of human limitations, and is consistent with naturalism. However, this argument is not fully satisfying because it leaves open the possibility of *psychological libertarianism*, the claim that decision outcomes are not fully determined by psychological factors (List, 2019). This would be consistent with the pre-theoretically intuitive idea that feelings of uncertainty indicate not just the possibility of doing otherwise, but of deciding otherwise. Although this form of libertarianism is compatible with naturalism it is inconsistent with the motivational strength idea.

I therefore present two arguments directly against psychological libertarianism. I argue that any link between low confidence and psychological openness is *contingent upon further deliberation*. In a normal situation when an agent responds to low confidence by deliberating further, then alternative future decisions are possible. But when deliberation ends, the opportunity for further evidence gathering or evaluation is removed, so the link between low confidence and psychological openness is severed. The counterfactual that further deliberation could possibly have led to change of mind, is irrelevant to a world in which there is no further deliberation. Feelings of confidence indicate a hunger for information, so are relevant to the question of *when* to decide, not *what* to decide.
This analysis suggests that the cognitive phenomenology of low confidence does not support psychological libertarianism, but a reader might wonder about the conative phenomenology of effortfully making a decision. The feeling of effort indicates that the agent is doing something. Is it possible that what she is doing is resisting the alternative with the highest conative value (and lowest risk-of-error), and actively choosing its less valued and riskier cousin? The answer is no, but the justification of this answer will have to wait until chapter 9.

In **Chapter 9** I bring together all previous elements of the thesis in order to solve what Vierkant calls the agency problem of compatibilism (2022). This is the challenge of accounting for the experience of effortful choice within a compatibilist framework. Can making a decision be an action? If it can, what exactly is the agent doing? If it cannot, why does it feel effortful?

Drawing on the analysis of the conative phenomenology of effort from chapter 4, I argue that the feeling of effort does imply that making a difficult decision is an action. It also indicates that a motive is being resisted. If the resisted motive can be identified then that will help answer the question of what exactly the agent is doing. My proposal, consistent with chapter 8, is that what is being resisted is the motive to reduce the risk-of-error. There is a (metacognitive) motivational contest and the motive towards prolonging deliberation is defeated, but not extinguished, by the motive towards stopping. The feeling of effort in an effortful decision is the feeling of resisting the ongoing motive to prolong deliberation.

In the standard model of decision making, an agent cannot decide until she has full confidence in her decision (O’Shaughnessy, 1980). The feeling of effort implies that
this model is too inflexible (Holton, 2006, Mele, 2017): The decision-threshold can be lowered. If an agent is most motivated to conclude deliberation that implies that the current level of confidence is *sufficient, given the context*. This shift, to a context-dependent, rather than fixed, definition of sufficient has the intuitive implication that deliberation is a flexible, motivated agential action rather than an irresistible reflex response. Unconfident agents may be hungry for more information, but that hunger can be resisted, just as a dieter can resist eating.

The phenomenology of an effortful decision does not undermine the motivational strength idea. Rather, it supports the claim that the motivational strength idea also applies to metacognitive decisions. The phenomenology neither justifies beliefs in libertarianism nor indicates an ability to asymmetrically bias a decision. When the decision threshold is lowered one option will win, not because it is boosted by agential powers, but because it was slightly ahead, and its small lead is now deemed to be enough to be declared the winner.

The analysis of this thesis is functional and phenomenological, but the implications are metaphysical. Libertarianism (whether anti-naturalist or psychological) can be persuasively⁹ defeated by augmenting the metaphysical position of naturalism with a fine-grained understanding of agential phenomenology. An understanding of conative phenomenology supports the motivational strength idea, and an understanding of cognitive phenomenology reveals that feelings of uncertainty do not indicate that doing otherwise is possible. Even an agent who resolves indecision

---

⁹ I call a defeater unpersuasive if it leaves the question “If p is false, then why does p seem to be true?” unanswered. Whether libertarians are actually persuaded is another matter.
without resolving uncertainty, will always choose to do what she is most motivated to do.
Part I

Chapter 1

Attitudes, Values and Decisions

Throughout this thesis I will adopt a contemporary but broadly representational theory of mind. Fundamental to this view is careful differentiation between doxastic and conative attitudes, and between these attitudes and their content.

In this chapter my aim is to bring clarity to nuanced elements of the representational theory that will be relevant to the broader thesis. This part of the dissertation is in general descriptive and definitional, although some argumentation will be necessary. I aim to confirm that a contemporary understanding of the representational theory is consistent with the latest philosophical and scientific thinking, and provide some of the foundations on which my later arguments will stand. These include the concepts of motivational force, conative value, inhibition, and schedules within plans.

Although I don’t expect the model I develop in this chapter to be particularly controversial, some may find the emphasis on certain aspects, such as inhibition and plan schedules, quirky. This quirkiness is driven by a focus on elements that I believe are central to an understanding of cognitive and conative phenomenology. The reasons for their centrality will become clear in the main body of the thesis.

By the end of the chapter, I hope to have made clear my perspectives on the following issues:

1. The link between conative value and motivational force.
2. The differences between doxastic and conative attitudes, including the management of conflict between attitudes.

3. The importance of differentiating between attitude and content.


5. The relationship between the motivational strength idea and a cost-benefit computations when deciding what to do.

I will start at the beginning, with a simple agent who aims to continue to be. Such an agent is continuously trying to respond to the (implicit) question of “what to do?” with the same answer: “That which maximises sustainability”.

1. Conative value

1.1 Sustainability

The ultimate goal of any agent is to sustain itself. This is achieved through care, where care is shorthand for harm avoidance, maintenance, and enhancement.

The breadth of this goal and the complexity of any goal-seeking behaviour depends on the peculiarities of the agent. In its simplest form sustaining oneself involves gathering resources and using those resources to resist entropy.

“To be” is its intrinsic goal. Teleology comes in where the continuous identity of being is not assured by mere inertial persistence of a substance, but is

---

10 Homeostasis is the ‘mother-value of all values’ (Weber and Varela, 2002)  
11 Sophisticated agents may also aim to sustain others, such as their progeny.
continually executed by something done, and by something which has to be
done in order to stay on at all. (Jonas, 1968, p243)

If a simple mobile agent identifies a source of sustenance, then it moves to benefit
from the resource. A more sophisticated agent may also be able to identify sources
of harm and move away. In accordance with the etymology, the internal states which
move an agent are called motives, although as they don’t guarantee movement it is
perhaps more accurate to call them conative states. Conation implies trying or
striving\textsuperscript{12}. Consistent with this, Spinoza described the essence of human agents as
\textit{conatus}: ‘Each thing, as far as it can by its own power, strives to persevere in its
being’ (III Proposition 6, translation from Curley, 1969). I will use both terms,
understanding motives to be fallible.

The word autonomy comes from the combination of “auto”, self, and “nomos”, norms.
If an agent is an individual who interacts asymmetrically with the world according to a
set of norms (Barandian et al 2009), then it is the duty of care that underpins the
norms, and conative infrastructure that transforms those norms into motives that
drive action. Throughout this thesis I will call the set of norms that govern agential
behaviour “care-based norms”.

The environment through which an agent moves is a landscape of affordances,
where for the environment to ‘afford’ something means to provide or supply it\textsuperscript{13}
(Gibson, 1979, p127), and the amount of sustenance supplied by any individual
source varies. Therefore, for an agent to optimally navigate an environment of

\textsuperscript{12} Conation: ‘an inclination (as an instinct, a drive, a wish, or a craving) to act purposefully’ (Merriam-Webster),
‘the element in psychological processes that tends towards activity or change and appears as desire, volition,
and striving’ (Collins)

\textsuperscript{13} Affordances are relational, in that they depend on the abilities of the agent. A tree is climb-able, or a ball
throw-able, only to an organism capable of climbing or throwing.
opportunities and dangers, it is adaptive for her to be most attracted to sources which provides the greatest sustenance. I will call the amount of sustenance supplied by a source, its care value. If an agent has a cognitive architecture that predicts the care value of affordances, then she can create an evaluative map, with which she can navigate the environment. The elements of this evaluative map are conative values. They represent the agent’s subjective and context-specific predictions of the care value of environmental affordances. The greater the conative value, the greater an agent’s motivation towards the source of sustenance.

Motivational force in this model, is a non-conceptual representation of the conative value an agent’s cognitive infrastructure has assigned to a plan of action. In section 2.2 I analyse in more detail this link between value and force.

The fact that a cognitive infrastructure exists which encodes an agent’s various conative values is not a necessary fact, but it is a natural fact. Over the last decade, empirical data in the appropriately named field of neuroeconomics has confirmed this idea: ‘Neuroeconomics is now anchored in a new theoretical framework that posits the existence of a centralized neural mechanism that enables the a priori incomparable subjective values of multiple stimuli to be compared using a single scale’ (Serra, 2021, p47, see also Padoa-Schioppa, 2011, Smith et al 2010, Levy and Glimcher, 2012, Westbrook and Braver, 2016). These encoded values may be innate or learned, intrinsic or extrinsic, subconscious or explicitly endorsed. They are recalcitrant but not static, generally adjusting over time based upon experience.

The only values that are relevant for synchronic motivation are the context-relevant conative values that have been activated by the (internal and external) state of affairs. Although an agent can act to create an environment in which a certain motive is likely to be activated, this indirect action is ‘of the catalytic sort’, and ‘the rest is just
waiting, seeing if anything happens’ (Strawson, 2003, p232). For a given state of affairs, activation is involuntary. That is not to say that activation is random. The process of activation has been stabilised through natural selection and learning, to reliably, although not perfectly, activate motives that are relevant to the situation at hand.

Whilst the disposition to conatively value an end when certain conditions are met is often described as recalcitrant (Sripada, 2021), activation or de-activation can be instantaneous. The motivational pull of a desire generally vanishes if the trigger is removed. For example, an agent who thinks she sees her lover across the room, will immediately lose motivation to approach if she realizes she is mistaken and it is a stranger. Similarly, the aversive fear of a snake will disappear if it becomes obvious that the snake is actually a stick. That is not to say that the full affective state immediately returns to neutral. I am making the less substantive claim that there is no motivational push to run from a stick, even if that stick was initially mistaken for a snake.

In summary, agents pursue ends of value. They do so because they are motivated, in appropriate circumstances, to act in ways that achieve (or increase the likelihood of) states of affairs consistent with those ends. Thus, motivation is transmitted from ends to means. If there was only ever one potential course of action open to an agent, then value computations would be unnecessary and behaviour inflexible. It is in situations of conflict that representations of value support decisions about what to do. In the following section I introduce three sources of distinctly human conflict that will be relevant for the broader thesis.
1.2 Distinctly human sources of motivational conflict.

Humans are differentiated from other animals in part because they have a (vastly) greater variety and complexity of conative values, many of which rely on language use and meta-representation. This results in a crowded conative environment in which motivational conflict is frequent and significant. The management of this motivational conflict is central to what it means to be a flexible human agent and perhaps the most important driver of the phenomenology of agency.

Here I introduce three sources of evaluative complexity that are extremely rare in the animal kingdom, if not unique to human agents. Whilst I remain open to the possibility that some of the findings of this thesis will be relevant for other animals whose set of values is large enough to cause conflict, my focus is on human agency. The three highlighted areas of value are responsible for the motivational conflict that underpins much of what is distinctive about human agency. All mentions of value should be understood as referring to conative value.

Physical vs Narrative identity

A first source of conflict comes from the fact that human self-care relates to not only care for the physical body in the physical environment, but also care for the psychological or narrative identity in the socio-cultural environment. For an illustration of how the latter can conflict with the former, consider Beatrice Garland’s poem about a Kamikaze pilot who returned:

And though he came back
my mother never spoke again
in his presence, nor did she meet his eyes
and the neighbours too, they treated him
as though he no longer existed,
only we children still chattered and laughed
till gradually we too learned
to be silent, to live as though
he had never returned, that this
was no longer the father we loved.
And sometimes, she said, he must have wondered
which had been the better way to die.

The fact that human agents are able to prioritize narrative identity over physical perseverance is striking, but I will assume it is uncontroversial\(^{14}\). Indeed, whilst the construction of a self-narrative might be uniquely human, social needs and social hierarchies are not, and social motivations are felt across the animal kingdom. For example, Gillian and Tye bring together a wealth of findings from across the animal kingdom to argue that what they call ‘social homeostasis’ (2019, p5) is a widespread phenomenon. Short term deviations from a stable social “set point” elicit motivated behaviour to correct the deviation, and long-term deviations can cause chronic maladaptive neurochemical changes. These social motivations then compete with and influence other motivations.

In chapter 3 I will analyse internalist vs externalist models of moral motivation. Whilst some philosophers believe that some judgements, such as moral judgements, must be intrinsically motivational (Hare, 1952), I will argue that agents are motivated to act

\(^{14}\) In Asimov’s short story ‘Bicentennial Man’ (1982), the robot Andrew arranges his own death in order to be recognised as human. Andrew is asked how he could do such a thing whilst governed by the Third Law of Robotics, which states that a robot must protect its own existence. Andrew answers: “I have chosen between the death of my body and the death of my aspirations and desires. To have let my body live at the cost of that greater death is what would have violated the Third law” (p564).
in line with judgements only when they have antecedent motivations (Mele, 2003), such as the desire to do what one morally ought, to do what one judges best, or to adhere to tribal norms. Values based upon sustaining one’s narrative identity enable an ethically, epistemically, or socially normative should to be transformed into a conative want-to, motivating an agent to do what she believes she should do.

The synchronous vs diachronic self

The second source of evaluative conflict, highly relevant for analysis of self-control, is the fact that humans can engage in what has been called mental time-travel, allowing them to value, for example, not just the immediate benefits of an extra glass of wine but also yesterday’s resolution or tomorrow’s productivity. Thus, care relates not just to the synchronous self, but also the diachronic entity that persists over time. Philosophers have noted that the relationship between one’s current and future self at least partially mirrors that between oneself and another. (Ricouer, 1992, Bratman, 2018, Parfit, 1986).

Diachronic agents understand that evaluations of what to do can vary over time independent of the actual state of the world, which means that current or future evaluations may not optimise care for the diachronic self. This creates the potential for conflict between the synchronous and diachronic self and the possibility that self-control itself is valued.

Asymmetric interventionist models (Wallace, 1999), view control as being exerted over conative attitudes, whereas choice models (Berkman et al 2017) view control as nothing more than value-based choice. The bridge between these two theories is provided by the fact that self-governance, or self-control, is itself valued. Control over impulses shifts from appearing asymmetric to appearing value-based depending on
whether the value of control is excluded or included. In chapter 5 I will show that the fact that human agents value self-control explains the possibility of intentional synchronic self-control.

*Resolving indecision vs resolving uncertainty*

In an ideal world, all relevant practical values and information would be incorporated into every decision, but resource constraints, both temporal and cognitive, mean that in practice only a subset can be considered. Thus, the output of any evaluation includes an error term\(^{15}\), which should generally decrease as evidence is gathered. To reduce the potential harm associated with error, humans monitor uncertainty through noetic feelings such as feeling of knowing (Dokic, 2012), and ascribe value to information and uncertainty reduction. However, there is also value to an agent in appropriately allocating her finite cognitive and temporal resources. Spending too long focusing-on and deliberating-over a single issue has opportunity costs (Kurzban et al 2013). This creates an explore-exploit tension (Hills et al, 2015) between error minimisation and resource re-allocation.

The values discussed here, as well as the value of control, are metacognitive, but I will reserve further discussions of metacognition until section 3. For now, it is enough to understand that metacognitive motives introduce a further source of conflict for human agents.

Human sophistication has not shifted the goal away from care, but it has significantly complicated the definition and administration of care-based norms. Within and between each of the care-related values there is opportunity for conflict. This

\(^{15}\) In striking new research, it is proposed that ‘optimistic’ and ‘pessimistic’ neurons map the entire complex distribution rather than just the mean plus an error term (Dabney et al, 2020)
explains the need for a mechanism, discussed in section 2.3, through which to resolve that conflict and avoid paralysis. Before turning to the decision-making process, however, I will first analyse the two main elements of the representational theory of mind, doxastic and conative attitudes. My main focus will be on highlighting their differences.

2. Attitudes and Content

2.1 Doxastic and conative attitudes

The two core attitudes on which I will focus, correspond to a belief/desire model of the mind (Hume, 1748 [1999]), Davidson, 1963, Mele 1992). I will refer to them as doxastic and conative attitudes.

Doxastic attitudes - beliefs, disbeliefs or suspension of judgement - have truth or accuracy conditions and mind-to-world directionality. An agent’s doxastic landscape is a model of the world and the model performs its function to the extent that it accurately represents or predicts the world. Within this overall success condition there are nuances. First, as agents don’t have the cognitive capacity to model every aspect of the world, they demand greater precision or accuracy in areas relevant to their overall sustainability. Second, the doxastic landscape is not merely a list of known facts, but is a probabilistic space, which contains both statistical and contingent elements, including complex probability distributions. For example, an agent can believe that it is 60% likely to rain, or that if it rains, the river will flood. Third, sophisticated agents can model future worlds across multiple time horizons. They can simultaneously believe that it will rain later today and that tomorrow it will be clear.
Conative attitudes, which correspond to desires, have satisfaction conditions and world-to-mind directionality. Although some (Kriegel, 2015) may class non-motivational attitudes such as wishes or hopes as conative because they indicate approval, I remain true to the etymological link between conatus and trying, and set these aside. The attitudes that are relevant to my project are those that are motivational. They include action-desires (Mele, 2003), desires-to, and wants. Although I will use these terms when it seems natural, I will generally be explicit and simply call them motives.

Motives are pro-attitudes that dispose an agent to act and bring about certain changes in the world. As Anscombe states, ‘the primitive sign of wanting is trying to get’ (1957, p68). An agent who maintains a frustrated desire-to without being motivated to do anything about it is failing a norm of desire (Wall, 2012). Having a desire-to find shelter from the rain necessarily implies being motivated to find shelter. If an agent is not so motivated, then this is a sign that she doesn’t truly desire-to find shelter.

Consistent with the norm of desire is the broadly accepted idea that motivational attitudes are restricted in that they can only be directed at non-actual but possible worlds. As Velleman asserts, ‘one can desire that \( p \) only if \( p \) seems attainable’ (1992, p17). Attainability is linked to motivation in that, to be motivated to perform a specific action to bring about a desired end, an agent must believe that the action is appropriate to realize that end. If she believes there is no such appropriate action, because the desired world is actual or is impossible, then a rational agent will not feel motivated to perform it. ‘What is clearly unattainable is not subject to consideration as an end, as there is no way of conceiving means to its achievement’ (Whitford, 2002, p339, discussing Dewey’s account of desires).
The irrationality of pursuing unattainable desires may seem to create a puzzle for the direction of fit framework. In the case where a desire is unattainable, the world can’t be made to fit with that desire, so it is the desire which must adjust. Ricouer calls this ‘consent’: ‘converting within myself, the hostility of nature into the freedom of necessity’ (1950, p346). My view is that this puzzle can resolved if the desire to only pursue attainable projects is understood to be meta-conative. It acts on lower order motives and treats them as objects in the world. If this is right, then the world-to-mind direction of fit is maintained. I analyse metacognition more generally in section 3.

2.2 Representing-as-valuable vs value-representations

Attitudes are often described as representational. Doxastic attitudes represent-as-true, whilst conative attitudes represent-as-valuable. Personally, I like this terminology, and will use it through the dissertation, but it does create the risk of confusion, as “value” in the context of conative motivational attitudes is defined in a way that diverges from folk understanding. In this section I analyse the difference between conative and doxastic value.

Within representational attitudes, truth and value seem to be co-dependent. Doxastic accuracy has value to the agent in that it reduces the harm of error, whilst the value-representations underlying conative attitudes have accuracy conditions: value-representations can be correct or incorrect predictions of whether an action is beneficial for an agent

16 As Sripada states, conative attitudes are based on value representations which are predictions, and predictions, like doxastic states have accuracy conditions. they ‘are accurate to the degree that they approximate what is being predicted.... Thus, value representations track their representational target in the manner that doxastic states do’ (manuscript).
“value”, in the context of conative attitudes has a technical definition which includes direction-of-fit\textsuperscript{17}. Conative values are not just valuations in the ordinary sense, but are demands ‘to be made true’ (1992, p12)\textsuperscript{18}. To understand the importance of this conceptual distinction, consider the following erroneous argument:

1. Conative motivational attitudes represent-as-valuable their content
2. Value-representations are predictions, which have accuracy conditions
3. Representations with accuracy conditions are doxastic states
4. Therefore, value-representations are doxastic states
5. Therefore, conative motivational attitudes are doxastic attitudes.

The error becomes clear when one understands that the definition of value is not consistent between premise 1 and premises 2+3. For premises 2+3 to be correct, value must be understood in a non-technical inert sense. But this non-technical understanding of value would miss the most important fact about motives: They move an agent to act. Whilst a motivational force without an accurate set of beliefs or values would be anarchic, an accurate set of values without a motivational force would be moribund.

I, and others (Sripada, manuscript) use the phrase “evaluative landscape” to capture the fact that an agent can be motivated towards a number of environmental affordances with different conative values. To me this phrase conjures up the image of mountains and valleys. Unfortunately, if others feel the same, there is a risk that this terminology is a false ‘intuition pump’ (Dennett, 2013) as it implies that the contours of the landscape are fixed and any force (gravity) on objects tumbling

\textsuperscript{17} ‘The cognitivist conception of desire as an action-justifying value judgment thus depends on a misinterpretation of the sense in which desiring something entails regarding it as good’ (1992, p20)

\textsuperscript{18} ‘The propositional object of desire is regarded not as fact -not, that is, as factum, having been brought about - but rather as faciendum, to be brought about; it’s regarded not as true but as to be made true’ (p8)
through the landscape is external. If the strength of gravity were to instantaneously change, the landscape would remain the same. For agents, in contrast, the contours of the evaluative landscape are not fixed. They are set by the strength of motivational forces. If I suddenly realise that the stick on the forest path is actually a snake, then the landscape will steepen dramatically. And, in cases of listlessness, the evaluative landscape flattens.

Conative value is not a value that interacts with an extrinsic motivational force. It is the strength of the motivational force. It is another way of saying motivational strength, so if there is no force then there is no value. Trying to split the two is like taking a £10 note and suggesting to a friend that she can keep the 10 while you will keep the £. Therefore, neither conative attitudes as a whole, nor elements of conative attitudes, can be reduced to doxastic attitudes. This is true despite the fact that from an external perspective an agent’s evaluative landscape may seem flawed. While it may be accurate to say that the listless agent’s evaluative landscape should be steeper, or that it would be steeper if she was well, it is false to say that it is steeper but is currently coupled to low motivation.

Conative attitudes are intrinsically motivating whereas doxastic attitudes can only influence action indirectly, via, for example, antecedent motivations. Consistent with this, ‘the value-commitment built in to conative states is attitudinally encoded’ (Kriegel, 2015, p74). Representing-as-valueable content “X” is not equivalent to representing-as-true the content “X is valuable”. Wanting X is necessarily motivational. Believing that X is valuable is not.
There is another, less commonly discussed, difference between doxastic and conative architecture: the management of conflict between attitudes. As a function of this difference, conative attitudes must be inhibitable. I turn to that now.

2.3 The coherence requirement vs persistent conflict and inhibition

Doxastic attitudes are subject to a coherence constraint. Any known conflict between beliefs must be resolved. This follows from the accuracy condition.

Coherence is a necessary (but not sufficient) condition for accuracy. It is impossible for a model of the world to be accurate if it is incoherent. Therefore, doxastic architecture is structured to reject incoherence. Indeed, given that mind-independent reality is intrinsically unknowable, doxastic coherence is a more tractable problem than doxastic accuracy. Despite this, resolution of belief conflict will ideally involve returning to the evidence, as two contradictory beliefs cannot both be justified by the same evidence-set. One, or both, of the conflicting beliefs must be adjusted until they are justified by the evidence, cohere with each other, and cohere with the rest of the epistemic landscape.

Although belief conflict management might seem operationally complex, it mirrors the lack of incoherence in the (macro) world that is being observed and modelled. Two objects cannot inhabit the same space at the same moment in time, an object

---

19 Sometimes inconsistencies between beliefs, or between beliefs and actions will become known to the agent through unpleasant feelings of dissonance (Festinger, 1957)

20 Whereas foundationalists argue that beliefs are rationally constrained by non-verbal experience, coherentists argue that beliefs are rationally constrained only by other beliefs.

21 Quantum mechanics makes coherence less intuitive, as exemplified by the story of Schrodinger’s Cat, but doesn’t undermine the points made here for two reasons. First, no matter how difficult to understand, physical laws are never internally inconsistent. Second, Schrodinger’s cat can only be in an indeterminate state when unobserved. Therefore, both the underlying physical world and the observed world are subject to the coherence constraint, even if the link between the underlying and the observed is beyond comprehension.
cannot have two properties that are inconsistent, an agent cannot perform two conflicting actions at the same time. The world that is being modelled is itself constrained to be internally consistent. This implies that although belief coherence and the associated enabling mechanisms are fascinating areas of study, the coherence requirement is not unique to doxastic states. Rather the coherence constraint is a broader fact of the world.

In contrast to beliefs and macro reality, direct conflicts between conative attitudes can and do persist. Whilst two contradictory beliefs cannot both be true, an agent can simultaneously have two conflicting pro-attitudes. The coherence constraint does not apply to conative attitudes. In fact, simultaneously wanting to perform action A and an action inconsistent with A is not unusual. It is a regular occurrence that must be managed. The persistence of conflicting desires necessitates two elements of conative architecture:

First, conative dispute must be resolved before action commences. An agent who has two competing motives must decide between them. The leading hypothesis for how she does so is the motivational force idea: There is a contest between motives the most powerful motive wins (Mele, 2003, Hume, 1748 [1999]). Motivational contests are metaphorical arm-wrestles, and only when a victor is declared is the winning action-plan implemented. Although intuitive, this is not uncontroversial, so I will analyse the motivational force idea from many angles throughout this dissertation. I will conclude that it is correct.

Whilst the coherence requirement of doxastic states matches the constraints on a world of physical objects, the possibility of persistent motivational conflict mirrors that seen in physical forces. If two opposing physical forces meet, that does not imply
that one or both of those forces must be adjusted. Rather, the resulting action will be in the direction of the largest force. Although the idea that motives are forces is compelling, it is necessary to acknowledge one key difference between a clash of (external) physical forces and a contest of (internal) conative forces. The resulting action in the physical case is the net of the two forces, but for conative attitudes this “compromise” approach would be maladaptive. Instead, what has stabilised within the conative architecture of sophisticated agents is a winner-takes-all model, as shown in figure 1.

![Figure 1](chart.png)

**Figure 1. Charts illustrating the difference between “net” and “winner-takes-all” outcomes in a contest between two forces.**

Conative motivating attitudes ready an agent to act and compete in a winner-takes-all contest. The action-plan of the winning desire is implemented, but an obstinate

---

22 There will often be compromise with respect to the question of what to do, but this is realised through a compromise content, not by switching from a net vs winner-takes-all outcome.
desire can persist even after its opposition has been chosen. To manage all competing motives while the contest is ongoing, and obstinate desires once a decision has been made, a mechanism must be in place to block transmission from readiness to action. Motives must be *inhibitable*.

Motive inhibitability, which is implemented in conative architecture through what are called inhibitory gates, has three functional implications: (i) agents are not compelled to act on every motive; (ii) agents do not try to simultaneously satisfy two directly conflicting motives; and (iii) for as long as an agent is undecided about what to do, all competing alternatives are inhibited. Whilst the inhibitability of conative states is a necessary, perhaps obvious, consequence of conflict persistence, it will turn out to have important implications that will be explored throughout this thesis. Arguably, inhibitability is one of the central foundations on which flexible agency and its associated phenomenology is built. Without inhibitable motives we would be very different agents. The inhibitability of motives will be an important element of the analysis in chapters 2, 4, 5 and 9.

The importance of inhibition raises an interesting question with respect to intentions. These are generally considered to be conative attitudes as they exert a motivational force, but they differ from action-desires in (at least) one crucial respect: they are uninhibited. An intention comes into existence when the question of what to do is settled and inhibition is lifted. Intentions are executive attitudes in that ‘to have the intending attitude towards a plan is to be settled (but not necessarily irrevocably) on executing it’ (Mele, p28). As Bratman has famously detailed, any individual intention ‘becomes part of my web of intentions and plans, a web subject to the plan-type demands of consistency and coherence’ (1999, p 32). Intentions are therefore different from other conative attitudes in that they are subject to consistency
constraints, just like doxastic attitudes. But this should not be surprising. The winner-takes-all model, implemented via inhibitory gates, is an incoherence filter. Whilst persistent conflict can exist amongst motives, it cannot persist amongst intentions that have been released from inhibition. Bratman’s demand for consistency and coherence is satisfied by the gating of unchosen motives.

Taking stock, I have analysed two major differences between doxastic and conative attitudes. The first is their direction of fit. Conative attitudes move an agent to act on the world, whereas doxastic attitudes are moved by observations of the world. The second, much less discussed but equally important, is that, whilst doxastic attitudes are subject to the coherence requirement, conative attitudes are persistent but inhibitable.

There is a third major difference between conative and doxastic attitudes: their phenomenology. I set this aside for now, saving a detailed analysis of conative and cognitive phenomenology for chapters 3, 4 and 8.

I have argued that conative attitudes cannot be reduced to doxastic attitudes. They should be conceptually differentiated. With careful theoretical analysis the two attitudes can be disentangled, but from a practical perspective they will always be tightly entwined. Doxastic accuracy has value to the agent, and the value-representations underlying conative attitudes have accuracy conditions. Neither an ability to model, nor an ability to act, are, on their own, sufficient for continued existence. It is only through purposeful interaction with the world, guided and informed by an accurate model, that an agent can sustain itself.

Having analysed doxastic and conative attitudes in detail, I now turn to their contents.
2.4 Content

In sections 2.1-2.3 I focused on the direction of fit and representational guise of an attitude, but it is also important to consider the varying kinds of content to which those attitudes refer.

Sripada (manuscript) has argued that propositional attitudes are ‘concept-laden’ but here the distinction between content and attitude is important. For an agent to desire some-thing, or believe that it is true, she must have some concept of that thing. But she need not have the concept of a desire or of a belief. Whilst the content of the attitude is concept-laden, the attitude to the content need not be.

The content of a doxastic attitude is often described as a proposition, which can be rated on a true/false scale. However, this shouldn’t be taken to imply that when an agent forms a belief, she necessarily contemplates a sentence. She can believe that the sky is blue without any recourse to language. Indeed, animals and small children can form beliefs, the contents of which are non-linguistic concepts. In addition, there is general agreement that while acceptance is all-or-nothing, belief can come in degrees. This raises the question of whether those degrees form part of the attitude or content. For example, if Muriel assigns a 65% likelihood of rain tomorrow. Does she believe-to-the-degree-of-65% that “it will rain”, or does she believe that “there is a 65% likelihood of rain”? The former is a complex attitude to a simple content, whereas the latter is a simple attitude to a complex content. My opinion is that both are possible. Unlike conative value, the degree of likelihood can shift between attitude and content. If asked whether she believes it will rain tomorrow Muriel might answer “probably” or “I am uncertain”. If she is asked whether she believes that there
is an approximately 65% likelihood of rain, then she will answer “yes”. I will return to this discussion in chapter 7 when I interrogate the difference between uncertainty and (lack of) confidence.

In addition to statistical contents, beliefs can have contingent contents. Muriel can believe that if it rains tomorrow, she will get wet. This belief will be maintained even if she revises down her likelihood of rain. Sarah Moss, in her book ‘Probabilistic Knowledge’ argues that the contents of beliefs are ‘probability spaces’ rather than propositions, but admits that this might just mean that ‘propositions are not the sort of objects that we thought they were’ (2018, p15). For ease of language, I will continue to talk of propositional content, on the understanding that propositions can be complex, probabilistic and there is no requirement that they are represented linguistically. Although I adopt this broad understanding of propositional content, the arguments and conclusions of this thesis would not be materially impacted or undermined if it were to be proven that the more traditional, narrow view of content is correct.

The content of a conative attitude is a plan-of-action. This may be simple or complex, immediate or temporally extended, partial or complete. The expression *Chloe desires a cupcake*, whilst widely used and clearly understood in ordinary language, is underspecified. It is shorthand for Chloe’s desire to bring about a state of the world in which she is eating a cupcake. The content of the attitude is the plan, which may have a number of steps, and the end of the plan is cupcake-eating. Chloe can be simultaneously blissfully ignorant of this nuance and yet highly capable at navigating and shifting the world in order to realise cupcake-eating goals.
Although plans can be the contents of conative attitudes, they can also exist outside of that structure. As Mele describes, an agent can have a plan for A-ing without that implying that she plans to A. He gives an example: ‘I am now mentally reviewing a plan for making paper airplanes that I once committed to memory. Consequently, it is true that I now have such a plan. But I do not now plan to make a paper airplane.’ (1992, p136). Equally, an agent can have a variety of attitudes towards plans. She “might admire plan x, be disgusted by plan y and desire to execute plan z” (2007, p741). However, when a plan of action is the content of a conative motivational attitude, the agent is motivated to implement that plan.

In chapter 3, whilst analysing the phenomenology of motivational pull, I will focus on an underappreciated element of some plans, their schedule. When an agent has a plan towards achieving a certain state of the world, it explicitly or implicitly addresses the question of when. As with other elements of a partial plan, the time specifications may be rough and filled out later, flexible and of secondary importance, or they may be a very specific primary focus. However, a plan with no timeframe for any element is not the kind of thing that can be the content of a motivational attitude. The scheduled rate or progress provides a conceptual bridge between the sometimes long-term nature of plans and the straightaway-ness of action-desires. Occurrent motivation is a response to a negative answer to the question of whether progress is being made as planned. It inclines an agent to do something now to get things back on track. If, on the other hand, perceived progress is in line with planned progress, then there is no acceleratory force needed.

This concludes my analysis of attitudes and their content. My hope is that what I have said seems about right. I am sure that questions remain. Resolving some of these, whilst important for other fields, will be unnecessary for my project of
experience-compatibilism. I therefore set them aside and are neutral to how they are resolved. The others, I will return to later as the analysis gets more fine-grained. By the end of the dissertation, my hope is that no important questions will be left unresolved.

3. Metacognition

The APA Dictionary of Psychology defines cognition as ‘all forms of knowing and awareness, such as perceiving, conceiving, remembering, reasoning, judging, imagining, and problem solving’. Metacognition is cognition about cognition, which includes the series of processes that monitor, evaluate and control cognitive activity. Whilst metacognitive monitoring refers to the subjective assessment of one's own cognitive processes and knowledge, metacognitive control refers to the processes that regulate cognitive processes and behaviour.

Although metacognition has been analysed in detail in the context of philosophy of mind, its implications for the philosophy of action have been less well explored. My approach will be relatively uncontroversial in that I will merely apply the findings from one area of study to another.

Metacognition can be divided into analytic and procedural (Dokic 2012). The former is reliant on meta-representation and involves explicit attitudes about first-order mental states and processes. Procedural metacognition, however, does not require meta-representation. Information about cognitive processes can be non-conceptually revealed (or broadcast) through noetic feelings, such as a feeling of knowing when considering a question. Unlike analytic metacognition, evidence suggests that
procedural metacognitive monitoring exists in some non-human animals such as macaques (Zakrzewski et al 2014).

If an agent has metacognitive monitoring capabilities, then she can learn that her mental states are to some degree independent of the actual state of the world and that an attitude at one time can seem wrong from the perspective of another time or place. This allows the agent to understand that it is possible for her to have a false belief or an inappropriate desire, even if at an earlier or later time she would believe or desire otherwise (Vierkant, 2022). Once she understands this, it is a short step towards developing mind-directed (meta-)conative attitudes, which treat elements of the cognitive machinery as objects to be manipulated. Hieronymi calls exerting control over one’s own attitudes, managerial control, and argues that it is possible:

If, unable to sleep, you want to believe that your children arrived home safely through the storm, you might call them and so provide yourself with convincing evidence that they have arrived. If you want to be sure that, tomorrow, you will still intend to avoid dessert, you might act, today, to create extra incentives: you might make bets with your friends. … Thus it seems we can manage or manipulate our own attitudes in roughly the way we can manage or manipulate ordinary objects: by taking actions designed to affect them according to our purposes (2009, p141)

Within this thesis I will focus particularly on two kinds of metacognitive action\(^{23}\) that take place within extended deliberation.

---

\(^{23}\) Hieronymi defines managerial control as any action that manipulates an attitude. This explicitly captures my first kind of metacognitive action. I broaden her definition to include any action that manipulates an attitude or cognitive process, thereby capturing my second kind of metacognitive action.
The first, and conceptually simplest, kind of metacognitive action follows from the (metacognitive) desire to manipulate a practical desire. For example, an act of synchronic self-control is the implementation of the desire to suppress a currently active wayward desire. This will be relevant for chapter 5.

The second kind of metacognitive action follows a clash of metacognitive motives concerning the allocation of cognitive resources. This may relate to whether to stop or prolong deliberation, whether to reconsider or maintain an intention, or whether to focus on implementation or information gathering. This kind of action will be relevant for chapters 2 and 9.

Metacognitive actions are driven by metacognitive motives just as practical actions are driven by practical motives. When there is motivational conflict, practical or metacognitive, this must be resolved. A decision must be made about what to do. In the following, final, section I will show that motivational contests and cost-benefit computations are two names for the same decision-making process.

4. Cognitive cost-benefit computations are based on conative values.

The central claim that I will defend throughout this thesis is the motivational strength idea. I will generally express this idea in slogan form: *When making a choice, an agent will always choose to do what she most wants to do.* Here is the same idea expressed more formally:

The Motivational Strength Idea:

When making a choice between A-ing and ¬(A-ing), the agent will choose to A if, and only if, at the moment the choice is made, the strength of her
motivation towards A-ing is greater than the strength of her motivation towards ¬(A-ing)

A is a course of action and ¬A is anything other than A. ¬A includes doing nothing, an action B, multiple possible actions B or C or D etc, or multiple combined actions B+C+D etc. The total motivational force in one direction at any moment is the sum of all motivational forces in that direction. For the avoidance of doubt, a motivation is, by definition, activated and occurrent. This model of decision-making is symmetrical or unbiased, because all activated motives are treated equally, competing on a level playing field. The only determining factor in a motivational contest is motivational force.

Within the motivational strength idea, the introductory clause is important. It indicates that a practical question might not be answered right away. A human agent can withhold her decision, and engage in an extended process of deliberation. But when a decision is made, it is always made in accordance with the agent’s motivational balance.

More recent accounts of decisions, and cognitive control (Shenav et al, 2013, Kurzban, 2016), understand intentional actions as the result of cognitive cost-benefit computations which compare the pros and cons of different alternatives. As Sripada states, “when estimates of overall benefits exceed costs, or exceed them by a certain margin, a person-level action ensues: the person makes a decision to perform the action” (2020). However, in order for the idea of a cost-benefit computation to be more than just a metaphor, an agent’s cognitive infrastructure must include an action-focused common measure on which the computation can be

24 Sripada’s full quote refers to “accumulation”. I will analyse thresholds and the details of the accumulation to bound model in chapter 7.
performed. This measure defines the units of the computation, ensuring that the costs and benefits of competing action-plans are commensurable. And, for a decision of what to do to be an agent’s own decision, this common measure must represent the agent’s own subjective and context-specific predictions of the value of each alternative. When the question of what to do arises, relevant predictions of value must be activated and transformed into this action-focused common measure. This measure, the unit of computation for decisions about what to do, is conative value. Conative values in support of an action will be benefits, those against the action or in support of an incompatible action will be costs.

This suggestion is consistent with the previously mentioned finding within the field of neuroeconomics that there exists a “centralized neural mechanism that enables the a priori incomparable subjective values of multiple stimuli to be compared using a single scale” (Serra, 2021). It is also supported by the fact that conative value was originally proposed as a solution to the problem of navigating the affordance landscape, where different resources have different care values. I therefore conclude that cognitive cost-benefit computations and the motivational strength idea can be unified through the following statement:

The single, commensurable, computational unit of cost-benefit computations is the motivational force of activated, value-based motives.

Perhaps the word evaluation more clearly conveys the sense of the decision-making process than computation. A decision of what to do is made based on conative values. The strength of an agent’s conative attitude toward each plan is a subjective, contextually-modulated, representation of the plan’s value. The plan she values the most is the plan she is most motivated to execute, so that is the plan she chooses. In
situations where the pro and con attitudes are of similar strength, the agent will remain uncertain about what to do.

Famously, activated motives are dual-faceted. They combine a motivational force towards action with the phenomenological experience of being pulled or grabbed. As such, motives are the psychological phenomena at the core of decision making that integrate the physical and the phenomenological. An agent’s experience of a motive, the felt motivational pull, is a non-conceptual representation of the force being exerted by the motive.

The claim that motivational strength is the common unit of cognitive cost-benefit evaluations, if true, shifts the contemporary language of computations back towards more traditional understandings of action as being driven by the conative force of motives, which are felt. It also provides one pillar of support for the motivational strength idea. Conative value ensures commensurability, and gives affective “heat” to otherwise cold computations.

Before I finish, I should perhaps take a moment to forestall a potential worry about conative value commensurability:

It has become relatively standard amongst philosophers of mind to divide difficult choices into those based on indifference and those based on incommensurability (Kane, 1999, Holton, 2006, but note Holton’s caution about the word incommensurability in footnotes 23 and 24). In the first kind of choice, the alternatives are not incommensurable but equivalent, so an agent is forced to “arbitrarily pick”, between the two. Setting aside indifference is sensible, in my view, because even if random picking is a kind of choice, it doesn’t seem to be a paradigmatic or particularly interesting kind (Vierkant, 2018, 2022). However, Holton
describes the second, more interesting, kind of choice, as when ‘we do not know how to compare’ (p6). How can I respond to this? First, there is a difference between being functionally able to decide between alternatives, and consciously “knowing how” to compare them. The former relates to the functional strength of pro- and con-motives, which I argue are always commensurable, whereas the latter will either rely on somewhat opaque conative phenomenology which may give the impression of incommensurability (see chapter 3), or on consciously available reasons, which possibly are incommensurable. Indeed, Holton’s detailed analysis shows how choices can be made ‘in the absence of conscious judgement’ (p8). If Holton’s incommensurability is understood as “seeming incommensurability”, then our two accounts are very much aligned.  

But this leaves a diametric worry, for me if not for Holton. If motives are so commensurable then how can I explain uncertainty about what to do? A rough preview of my answer is that conative values are predictions, and as such they involve a risk-of-error. An agent is uncertain about what to do if she perceives that the risk-of-error is too high, and she deliberates further with the aim of reducing that risk. When an agent withholding a decision, it is because the difference between these pro- and con- conative values is small in the context of the possible impact of future evidence.  

I will therefore proceed on the basis that conative values are commensurable and that the existence of uncertainty about what to do does not undermine that claim. 

25 This is the conclusion I will reach in chapter 8, but I will diverge from Holton in chapter 9, when explaining the feeling of effort.  
26 I present the full argument in chapter 7.
Takeaways

This chapter was intentionally uncontroversial. Having said that, the analysis has laid some significant foundations which are intended to provide a solid grounding for the forthcoming arguments. Therefore, I hope that readers will accept, or at least be open to, the following propositions:

1. Conative value indicates the magnitude of motivational force. Unlike doxastic value, if an agent conatively values a plan of action then she is necessarily motivated to implement that plan.

2. Whilst doxastic attitudes are subject to a coherence requirement, conflict amongst conative attitudes can persist. This is managed within cognitive architecture through the inhibition of unchosen (and not yet chosen) motives.

3. The plan-of-action content of a conative attitude incorporates a schedule.

4. Metacognitive actions are driven by metacognitive motives, just as practical actions are driven by practical motives.

5. The cost-benefit model of decision-making requires a commensurable unit of computation on which the evaluation is performed. For practical decisions about what to do, the unit is conative value. The cost-benefit model is therefore equivalent to the motivational strength idea.

These propositions are all consistent with the motivational strength idea, which I will defend throughout the thesis. Proposition 2 suggests the importance of inhibition to any motivational account. Later chapters will more than confirm that suggestion.
Chapter 2

Intimations

Are Intentions Just Uninhibited Desire-Motives?

Introduction

In recent years it has become accepted that although the desire to A and the intention to A both essentially encompass the motivation to A, they are distinctive mental attitudes. According to this view, an intention is a sui generis attitude that comes into existence when the question of what to do is settled\(^{27}\). In cases where there is initially some uncertainty about whether to A, desiring to A and intending to A are separated by a mental event, a decision. In this chapter I argue that, although intentions are stable commitments to the execution of a plan, commitment merely reflects the fact that the attitude is uninhibited and stability is driven extrinsic factors, including what has been called the implementation mindset. This implies that the distinctive attitude account is mistaken. Intentions are just uninhibited desire-motives.

Mele describes intentions as ‘executive attitudes towards plans’ (2003, p27). In his account there are three features of intentions: First, they have representational content. Their content is a representation of a plan of action. Second, they have a conative psychological orientation. The agent is motivated to execute the plan. And third, intentions have an executive, or settledness, feature. ‘To have the intending

\(^{27}\) Following Mele, I assume that when an agent suddenly swerves to avoid a dog that jumps into the road, she does that intentionally. This implies that the definition of settled must be ‘thin’ (1992, p168), in the sense that it need not be the result of deliberation or any other extended mental action of settling-upon.
attitude towards a plan is to be settled (but not necessarily irrevocably) on executing it’ (ibid, p28). The first two features are shared by desires, but the final element distinguishes an intending attitude from a desiring attitude.

In his influential account, Bratman also differentiates intentions from desires. He calls the executive feature of intentions ‘commitment’ and argues that ‘intentions are, whereas ordinary desires are not, conduct-controlling pro-attitudes. Ordinary desires, in contrast, are merely potential influencers of action’ (1999, p16). In addition, he highlights a second feature of intentions, arguing that an agent’s commitment to a plan has a distinctive stability. An intention ‘resists reconsideration: it has a characteristic stability or inertia’ (ibid, p16). Finally, when an agent intends to execute a plan, not only is she committed to the plan’s execution, but the plan also serves to constrain her further practical reasoning. Intentions are subject to a coherence requirement.

Whilst I agree with the broad thrust of these thoughtful and influential accounts, my focus will be on a very specific question: Even if one accepts that intending to A implies a stable commitment to A-ing, does it necessarily follow that intentions and desires are distinctive mental attitudes? In order to answer this, I will analyse whether it is it correct to say that commitment and stability are intrinsic properties of intentions.

Lewis stated that a ‘proposition that ascribes intrinsic properties to something is entirely about that thing; whereas an ascription of extrinsic properties to something is not entirely about that thing, though it may well be about some larger whole which includes that thing as part.’ (1983, p197). Assuming that mental attitudes are only part of the larger cognitive environment, I am questioning whether it is correct to
ascribe the properties of commitment and stability to a particular type of attitude, an intention, or whether they are more accurately ascribed to the agent’s cognition, or state-of-mind, as a whole. I will argue for the latter and for the specific claim that intentions are desires which are a) uninhibited and b) situated within an implementation mindset. These two facts about the larger cognitive whole, within which a conative attitude operates, lead to commitment and stability respectively.

An analogy may serve to help. Imagine that a ranch is situated in the path of a hurricane, and that on the ranch is a tethered horse. The horse wants to, and tries to, run from the storm but it is inhibited from doing so. Seeing the horse’s distress, the owner cuts the rope, allowing it to escape. An intrinsic property of the horse is its desire to run away and escape. The fact that it is initially held is an extrinsic property of the larger whole.

I will argue that, during deliberation multiple alternative plans are considered and, for as long as deliberation is ongoing, all of these alternatives are inhibited. The decision event is the moment when one of the alternatives loses its inhibition. This uninhibited desire must then cohere with other uninhibited desires, and exhibits stability to the extent that the agent, post-decision, is in an implementation mindset.

Whereas the conative pull towards execution is an intrinsic property of an intention, being stably settled upon execution is an extrinsic property. The intrinsic properties of intentions are equal to the intrinsic properties of desires, so intentions and desires are not distinctive attitudes.

The structure of the chapter is as follows:

---

28 One could flip the argument and say that desires are just intentions that are a) inhibited and b) are situated within a questioning mindset. Although I think my proposal is more natural, I am neutral with regards the terminology.
In section 1 I analyse debates within the philosophical literature about the definition of plans and desires. I argue that much of the difference is terminological and suggest specific terms, plan-representations and desire-motives, to ensure clarity. I briefly discuss implications of the complex and dynamic nature of plan-representations.

In section 2 I detail the cognitive processes that inhibit desire-motives, in the context of the hierarchical architecture of the human brain. I reveal the link between inhibitability of attitudes and person-level actions.

With these definitions and findings in hand, in section 3 I consider Bratman’s three features of intentions: commitment, stability and coherence. I argue that they are not intrinsic properties of the intending attitude.

I summarise the findings in section 4. Intentions are just uninhibited desire-motives.

The analysis of this chapter is functional in nature. At times I will use words such as effort which could be interpreted in phenomenological terms, but I will not rely on any arguments based upon phenomenological analysis. I am, however, confident that the account presented is entirely compatible with, and supported by, the phenomenology that surrounds decision-making and intention implementation. Chapters 3, 4 and 8, which involve detailed phenomenological analysis, can be considered as further support for the claims presented here.

1 Terminological Clarifications

1.1 Plans and Representations of Plans
In the context of attitudes such as desires and intentions, there are two potential interpretations of the word plan, which relate to whether philosophers understand it as a noun or as a verb.

For Mele, plans are the non-conative, representational aspect of an attitude. They are what the attitude is about. As such, an agent can have a variety of attitudes towards a plan. She “might admire plan x, be disgusted by plan y and desire to execute plan z” (2007, p741). The plan is the content of the attitude. It sets out what A-ing involves. An agent can have a plan for A-ing without that implying that she plans to A. He gives an example: ‘I am now mentally reviewing a plan for making paper airplanes that I once committed to memory. Consequently, it is true that I now have such a plan. But I do not now plan to make a paper airplane.’ (1992, p136)

In contrast, Bratman understands plans as involving “an appropriate sort of commitment to action: I have a plan to A only if it is true of me that I plan to A” (1999, p29). Plans according to Bratman are not recipes or procedures but “intentions writ large” (ibid).

For Mele plans are hypothetical actions. They are schemes to be considered. Whereas for Bratman plans are categorical, although not irrevocable. They are commitments to be executed.

Given that the word plan is both a verb and a noun, it is impossible to resolve this terminological difference. The best way to proceed is cautiously, with clarity. I will attempt to achieve this by calling Melean plans plan-representations, and Bratmanian plans plan-commitments.
As I will generally be following Mele in separating the attitude from the content, my focus forthwith will be on plan-representations.

1.2 Complex and Dynamic Plan-representations

Plan-representations in philosophical texts, including this one, are often designated by a single letter, but that does not imply that they are either simple or static. On the contrary, they are usually complex and dynamic. This is necessary if an agent is to fully evaluate the implications of making one choice rather than another, and to adjust her behaviour along the way to maximise the likelihood of achieving her desired end.

This point is beautifully illustrated with Austin’s metaphor of a miner’s lamp, as repeated in full below:

Although we have this notion of my idea of what I’m doing —and indeed we have as a general rule such an idea, as it were a miner’s lamp on our forehead which illuminates always just so far ahead as we go along—it is not to be supposed that there are any precise rules about the extent and degree of illumination it sheds. The only general rule is that the illumination is always limited, and that in several ways. It will never extend indefinitely far ahead. Of course, all that is to follow, or to be done thereafter, is not what I am intending to do, but perhaps consequences or results or effects thereof. Moreover, it does not illuminate all of my surroundings. Whatever I am doing is being done and to be done amidst a background of circumstances (including of course activities by other agents). This is what necessitates care, to ward off impingements, upsets, accidents. Furthermore,
the doing of it will involve incidentally all kinds of minutiae of, at the least, bodily movements, and often many other things besides. These will be below the level of any intention, however detailed (and it need not of course be detailed at all), that I may have formed. (1966, p438)

The two issues of complexity and dynamism raise questions about intentional actions and intention stability, which I will consider separately:

A complex plan-representation, to which the agent has an attitude, includes multiple elements. Even if she has a positive conative attitude to the aggregate plan-representation, she may have a neutral or negative conative attitude to some of those elements when considered in isolation. Bratman calls this the ‘package deal’ effect (1999, p143).

It is generally agreed that for an agent intentionally to A, A-ing must be an element in the to-be-executed plan-representation29. The debate is over whether the agent must also have a positive conative attitude towards A-ing itself (or at least not have a negative conative attitude towards A-ing). This nuance was most famously captured by Knobe in his 2003 experiment:

*The vice-president of a company went to the chairman of the board and said,*
*“We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.” The chairman of the board answered, “I don’t care at all about harming the environment. I just want to make as much profit as I can. Let’s start the new program.” They started the new program.*

*Sure enough, the environment was harmed.* (p191)

---

29 This implies that an unexpected consequence of an action is not intentional. The agent may still be blameworthy, but for negligence.
Participants were asked whether the predictable environmental consequence of a chairman’s decision was intentional, and 82% stated that it was, even though the vignette suggested that the chairman did not have a positive conative attitude towards harming the environment. If a positive conative attitude towards A-ing is not necessary for behaviour to be intentional, then what Bratman called the ‘simple view’ (1999), that for an agent intentionally to A she must intend to A, is false.

There is a deep and subtle literature on this question, containing much empirical data. Philosophers and psychologists have focused particularly on the seeming normative dimension to attributions of intentionality. To remain neutral, I will generally use the terms willed or intentioned instead of intentional. Whilst intentioned is normally used as a suffix, meaning having intentions of the type mentioned, by not including a prefix I understand it to mean having an intention simpliciter. Thus, intentioned should be understood to denote the existence of a positive conative attitude. For all intentional acts that are intentioned acts, the simple view is true.

The second issue, relevant to discussions of intention stability in section 3.2, is the fact that plan-representations are always partial, and details are filled out, and changed, over time. Given that an intention is a conative attitude to a plan-representation, does that imply that when plan-representations are updated the intention changes?

---

30 A second vignette was word for word identical to the above, except that the word ‘harm’ was replaced with ‘help’ and the word ‘but’ after profits was replaced with ‘and’. There was a dramatic contrast in responses to the two vignettes. Only 23% of participants said that the chairman had helped the environment intentionally.

31 This may indicate the concept of intentional behaviour not only has a descriptive role, but also serves a moral and evaluative function, or it may indicate that observers believe norm-breaking behaviour is motivated, even if the agent in the vignette says they don’t care (Guglielmo and Malle, 2010, but see Cova et al., 2016).

32 I will retain the word intentional when describing the views of other philosophers whose analysis hasn’t required this distinction.
Bratman states that plans have a ‘hierarchical structure. Plans concerning ends embed plans concerning means…. I may hold fixed certain intended ends while deliberating about means’ (ibid, p29). This sounds right but presumably even once settled-upon, means remain more susceptible to change than ends. This suggests that stability is not consistent across intentions but is modulated with respect to position in the plan hierarchy.

Whilst complexity and dynamism can almost certainly be accommodated within the distinctive attitude framework, if intentions are neither distinctive nor intrinsically stable then that simplifies the model. Although I will not presume that these features of plan-representations shift the balance of argument significantly in my favour, I am confident that they do not support the case against.

1.3 Desire-Motives

Within philosophy there is some dispute over whether all desires encompass motivation for action, and whether non-desire-based motivation exists. To put it another way, is desire either sufficient or necessary for motivation?

With respect to the former, it might be argued that desiring outcome \( p \) is equivalent to approving of \( p \) or ‘representing-as-good \( p \)’ (Kriegel, 2015, p74). On the surface, approval need not necessarily be accompanied by motivation. In ordinary language, for example, an agent may state that she approves of, or has a desire-for, sunny weather. Although this statement is one in which the agent represents sunny weather as good, it will generally not be accompanied by a motivational force. This is because there is nothing a typical agent can do to affect the weather.
In line with this, Goldman has argued that ‘wishes can count as desires without one's having the ability, hence the disposition, to act on them. I wish for or want good weather tomorrow for my golf date, but I am not disposed to do anything about it’ (2010, p91). Wishes and hopes could therefore be thought of as non-conative desires-for.

Unlike a desire-for, a desire-to perform an action is always accompanied by motivation. Mele calls this an action-desire, which he defines as follows:

‘An attitude x is an action-desire iff there is some action-type A such that the agent’s A-ing is the focus of x……..and necessarily, if x were to function non-deviantly and effectively, x would contribute to its own satisfaction by inducing the agent to make a suitable attempt’ (2003, 29)

Consistent with this is the view that an agent who maintains a frustrated desire-to without being motivated to do anything about it is failing a norm of desire (Wall, 2012). Having a desire-to find shelter from the rain necessarily implies being motivated to find shelter. If an agent is not so motivated, then this is a sign that she doesn’t truly desire-to find shelter. While beliefs aim at truth, action-desires aim at satisfaction.

The most obvious way to avoid having frustrated desires is to act in a way that brings about what is desired. This is why desires are often described as having a world-to-mid direction of fit. They aim to bring about certain changes in the world. Although the pull to action may not be acted upon if it is outweighed by stronger conflicting desires, ‘one who desires to perform an act and has no conflicting desire [or combination of desires] of equal strength is in fact unable to refrain from that act’ (Looper, 2020, p1348). Anscombe states that ‘the primitive sign of wanting is trying
to get’ (1957, p68), but trying to get necessarily involves a plan-representation. This may be a simple reach-for, a more complex temporally extended sequence, or it may be exploratory. An exploratory plan-representation would be necessary in cases where an agent wants to A, but needs to research what must to be done to make A-ing possible. The plan-representation is the content of the desire-motive\(^{33}\), and if it includes synchronic action then the agent feels the motivation, the urge, to execute that action.

Many philosophers have argued that action-desires aim at a non-actual but possible world. For example, Velleman asserts that ‘one can desire that \(p\) only if \(p\) seems attainable’ (1992, p17). In contrast, ‘one can wish that \(p\) even if the truth of \(p\) seems quite impossible, and one can hope that \(p\) even when \(p\) already seems to be true’ (ibid). Attainability is linked to motivation in that, to be motivated to perform a specific action to bring about a desired end, an agent must believe that the action is appropriate to realize that end. If she believes there is no such appropriate action, because the desired world is actual or is impossible, then necessarily the agent cannot feel motivated to perform it. ‘What is clearly unattainable is not subject to consideration as an end, as there is no way of conceiving means to its achievement’ (Whitford, 2002, 339 discussing Dewey’s account of desires). Dewey therefore disagrees with Goldman: desires ‘are not wishes, but represent states that might be attainable’ (ibid).

Quite possibly the difference of opinion between Dewey and Goldman is purely linguistic. All parties agree that some forms of approval, such as wishes and hopes, are not necessarily accompanied by a motivational force. Whether this disqualifies

\(^{33}\) I assume, in line with Berridge et al’s research (2009) that a desire-to is a want-to, and that agential wants and likes can diverge.
them from being desires depends on whether one restricts the category to action-desires, or is willing to broaden it to include desires-for. Rather than resolving this dispute I will opt for clarity. In this chapter I will forthwith focus purely on desires-to that are accompanied by motivational force. If non-motivational wishes (or hopes) are also captured in a broader class of desires, then those wish-desires differ from action-desires specifically in that they lack what is most interesting to me, a motivational force.

In conclusion, I will use the term desire-motive\(^{34}\) for the conative attitude that is both necessary and sufficient for motivation. Elsewhere in this thesis\(^{35}\), I will defend what Mele (2003) has called the motivational strength idea. This accords with the common-sense notion that, when given the choice, an agent will choose to do what she most wants to do. For the purpose of this chapter, I will assume that it is true and applies to desire-motives:

When choosing between A-ing and \(\neg(A\text{-}ing)\), the agent will choose to A if and only if, at the moment the choice is made, the strength of desire-motives towards A-ing is greater than the strength of desire-motives towards \(\neg(A\text{-}ing)\).

Here, A is the plan-representation of an action and \(\neg A\) is anything other than A. \(\neg A\) includes doing nothing, an action B, multiple possible actions B or C or D etc, or multiple combined actions B+C+D etc. The total motivational force in one direction at any moment is the sum of all motivational forces in that direction. For the avoidance of doubt, a desire-motive is, by definition, activated and occurrent. Unactivated

---

\(^{34}\) In the broader thesis I use the simpler term, “motives”, but here I need to leave open the possibility that desire-motives and intention-motives are distinct. Once I have shown that an intention is just an uninhibited desire-motive, I will be justified in using the simpler language.

\(^{35}\) Chapters 3, 5 and 9.
standing or dispositional desires are not motivational and are therefore not desire-motives.

Although this discussion of the motivational strength idea has focused on the link between the strongest desire-motive and the chosen action, it is important to recognise the counterfactual: If the desire-motive to A had not been the strongest, then A-ing would not have been chosen. In this instance the desire-motive to A would have been inhibited by the desire to do otherwise.

Inhibition, in its different forms, will be central to the findings of this chapter, so is analysed in detail in the following section.

2 Inhibition

2.1 Lateral inhibition of unchosen desire-motives

The simplest cause of inhibition of the desire-motive to A, is B-ing, where A and B are mutually exclusive plan-representations. This is because a choice between two conflicting plan-representations exhibits a winner-takes-all dynamic, accomplished through mutual lateral inhibition (Mao and Massaquoi, 2007, Cisek, 2007). For example, imagine a child is told she is allowed a single scoop of ice-cream. By choosing strawberry ice-cream, any desire to eat an alternative flavour is necessarily inhibited. The desire for chocolate may persist, but if she is forced to choose only one, and the desire-motive for strawberry is strongest, the desire-motive to have

---

36 One can imagine a hypothetical being who resolved indecision by trying to enact both plans simultaneously, with the result being an average or blend. However, this could have disastrous consequences, as King Solomon highlighted with the proposed compromise of a divided baby.
chocolate cannot be implemented. Indeed, for any action to proceed, all conflicting alternatives must be inhibited.

Although the winner-takes-all dynamic is described as an inhibitory process, ‘it does not require any independent inhibitor—it inhibition is the necessary consequence of combining competition with capacity limitations, as in every neural system.’ ‘Inhibition is an emergent property of cognitive processing rather than a separable factor.’ (Hommel, 2015, 42-43, see also Friedman et al., 2008). The motivational force of the desire-motive to eat strawberry ice cream explains not only strawberry ice-cream eating, but also the inhibition of any weaker desire to eat an alternative flavour.

Such inhibition, which occurs solely as a consequence of an agent’s positive conative attitude towards a competing alternative, I will term lateral inhibition.

Assuming that this is uncontroversial, I consider three further classes of inhibition: deliberation-driven inhibition, metacognitive lateral inhibition, and willed inhibition.

2.2 Deliberation-driven inhibition

In Jean Buridan’s famous story, an ass is equidistant from two equally attractive bales of hay, between which it is unable to decide. Unfortunately, despite a growing hunger, the ass remains paralysed by indecision and eventually dies of starvation.

Assumed within the story is the fact that when hungry, asses desire hay. This is due to a simple and adaptive negative feedback loop (Cisek, 2019) or organic circuit (Dewey, 1896) in which animals desire that which will reduce a need. The hungrier the ass, the stronger both competing desire-motives but, unfortunately, they still balance out. Either desire alone would cause the ass to rush to the hay, but neither
is stronger than the other so the ass cannot decide. Both plan-representations remain hypothetical rather than categorical.

Some readers may believe that, because it doesn’t matter which hay-bale the ass chooses, eventually it will be able to pick. Perhaps it will use a technique that harnesses randomness (although hooves make coin flipping difficult), or perhaps a micro-perturbation in the environment will be enough to move the decision-making process away from what is an increasingly unstable equilibrium.

In response to the random picking argument against decision-paralysis, one could adjust the story in two ways:

First, the ass’ paralysis may be specified to be temporary. In this case the ass is initially unable to decide, but then somehow resolves the indecision despite ongoing uncertainty about which option is best.

Second, the scenario can be shifted such that the need to make a decision is less urgent, but the stakes are much higher. Sartre gives an example along these lines when he describes a student who cannot decide between two courses of action: Joining the resistance or looking after his elderly mother (1946). Like Buridan’s ass, both alternatives are highly valued but here the issue is the potentially significant harm associated with rejecting one of the options. The lack of increasing costs of indecision, when combined with the fact that choosing one option would not reduce the desire for the alternative, implies a much more stable equilibrium. We don’t know whether Sartre’s young man ever did make a decision, but we can be confident that for an extended period he was torn between the two options, unable to do either.

Irrespective of one’s view of the possibility of resolving indecision, it is clear that in cases of plan-representation conflict neither action can proceed until a decision is
made. Whilst indecision remains, both options are inhibited. Both alternatives remain hypothetical until the question is settled. This inhibition of both alternatives during deliberation, I will unimaginatively call deliberation-driven inhibition.

In a competition between plan-representations A and B, both are inhibited until indecision is resolved. After a decision to A is made, A is implemented whilst the desire-motive to B remains inhibited. The inhibited plan-representations, both before and after the decision, are the contents of desire-motives. The question of this chapter is whether the post-decision uninhibited plan-representation is also the content of a desire-motive or of a newly formed sui generis attitude.

2.3 Lateral inhibition of metacognitive desire-motives

Metacognition is the series of processes that monitor, evaluate and control cognitive activity. In human agents there are two kinds of metacognitive desire-motive. Most relevant for this chapter are attitudes about cognitive processes, such as the motive to focus attention or prolong deliberation. These metacognitive attitudes are generally an adaptive response to the fact that human cognitive and temporal resources are finite, so should be allocated with care. The second kind are attitudes about attitudes, such as the desire to control the desire to eat ice-cream. I will reserve analyse of these attitudes until section 2.4.

Meta-conative attitudes conflict with other meta-conative attitudes in a motivational contest. For example, the metacognitive desire-motive to prolong deliberation (in
order to reduce uncertainty) conflicts with the metacognitive desire-motive to stop deliberation (in order to allocate cognitive resources elsewhere)\(^{37}\).

To reduce the potential harm associated with error, humans monitor uncertainty, and ascribe value to information and uncertainty reduction. However, there is also value in appropriately allocating finite cognitive and temporal resources. ‘Sometimes continued deliberation is costly’ (Shepherd, 2015, p346). Spending too long focusing-on and deliberating-over a single issue has opportunity costs (Kurzban 2016, Musslick and Cohen, 2021). The greater the opportunity costs the greater the urge to curtail deliberation.

The conflict between decision uncertainty and decision urgency creates an explore-exploit tension (Hills et al, 2015) or speed-accuracy trade-off (Bogacz et al, 2010) that an agent needs to manage. The agent (usually subconsciously) evaluates these costs and benefits to determine whether to stop or prolong deliberation. This is a metacognitive evaluation, that I call evaluation M (Hall, 2023).

In summary, when making a decision about what to do, an agent performs two evaluations:

Evaluation E (cognitive): An evaluation of whether to perform action A or ¬A.

Evaluation M (metacognitive): An evaluation of whether to stop or prolong E-ing.

In chapter 9 I will propose a specific model for evaluation M, but for this chapter, its precise details are unimportant. What is important is that there is a cognitive process, which takes in relevant factors and determines whether to stop or prolong

---

\(^{37}\) I analyse the importance of metacognitive lateral inhibition for active decision making, as well as the phylogenetic evolution of the ability to stop or prolong cognition, in chapter 9.
the evaluative process E. When it comes to the question of whether to re-open a question, the same inputs are relevant. For a given set of evidence and (internal and external) environment, if the answer to the question of whether to stop deliberation is in the affirmative, then the answer to the question of whether to re-open the question is in the negative I will return to this point when analysing the (extrinsic) reasons for intention stability in section 3.

Although the desire-motives described are metacognitive, the inhibition is still lateral. The desire to stop deliberating is inhibited by the desire to prolong, and vice-versa. Lateral inhibition, metacognitive or otherwise, is the necessary consequence of mutually exclusive competition and there is no need for an independent inhibitor.

When indecision is resolved, the chosen motive is released from inhibition and its plan-representation begins to be implemented. In the case of prolonged deliberation however indecision remains and the lower-level, practical, desires-motives between which the agent is deliberating remain inhibited.

2.4 Willed inhibition

Human agents, due to metacognitive abilities, can explicitly decide to engage in inhibition of an unwanted response. For example, an agent may consciously decide to not move a muscle whilst playing hide-and-seek; or to resist the temptation to eat ice-cream; or to delay decision implementation until checking with the boss. In these cases, the agent doesn’t just have a conative attitude which naturally leads to inhibition. Rather she has a positive conative attitude towards the content inhibit A-ing (or delay A-ing). Her intention is to perform an act of inhibition. This I will call willed inhibition.
Whilst meta-conative attitudes towards motives are relevant for the thesis in general, and for chapter 5 on intentional synchronic self-control in particular, I can set them aside here. In this chapter the question is whether the intention stability results from factors intrinsic to the attitude. An explicit metacognitive attitude with a content along the lines of *inhibit all alternatives until implementation is complete*, might enhance stability of an intention but it would certainly not be intrinsic. Indeed, in section 3.2.2 I briefly consider explicit policies with respect to reconsideration, such as a resolution or an oath. I conclude that neither indicate that intentions have intrinsic stability. In fact, the use of explicit policies indicates that the agent believes that intentions are sometimes not stable enough.

2.5 Inhibitability and Person-level Actions

In his analysis of intentional agency, Bayne describes intentional actions as ‘actions that are carried out by agents themselves, and not by some sub-personal or homuncular component of the agent’ (2013, p162). But how should we understand the difference between person-level actions and sub-personal responses? For Bayne the differentiating factor is the level of integration within the broader cognitive economy. Whilst a sub-personal reflex-response is functionally isolated from the rest of an agent’s cognitive architecture, an action that is ascribed to the person herself must be motivated by an attitude that is cognitively integrated ‘in a flexible and appropriate manner’ (ibid, p164). I will call this requirement, of flexible integration within the cognitive economy, the Bayne requirement.

Bayne’s difference between person-level actions and sub-personal responses can be cached out in terms of inhibitability.
A knee-jerk response, once triggered, is uninhibitable, so can never be flexibly integrated. The agent as a person can have no influence over the reflex. A desire-motive, on the other hand, is susceptible to lateral, deliberation-driven and willed inhibition. No matter how strong an activated desire, it is always possible that the desired action will be rejected, if the agent is more motivated to do otherwise, or delayed, if deliberation is incomplete.

Any activated desire-motive may, at a particular moment in time, be inhibited or uninhibited, but desire-motives are always, by definition, inhibitable. This inhibitability is what underpins the flexible integration of desire-motives within the cognitive economy, and ensures that motivated actions satisfy the Bayne requirement.

2.6 Absence of inhibition

As described by Bargh and Chartrand, in their discussion of ‘the unbearable automaticity of being’ (1999), most behaviour is driven by automatic, non-conscious mental processes. In these cases, there is no extended consideration of the question of what to do, and no desire-motive conflict or inhibition of the sort outlined above. Action, assuming no external resistance, proceeds automatically and effortlessly because no alternative exists or comes to mind. This raises the question of whether such automatic actions are the product of intentions. The lack of deliberation seems to remove any moment of intention formation. Although I believe that it is possible to argue within the distinctive intentions account that intentions can be

38 Automatic actions should not be confused with the knee-jerk reflex responses, described in section 2.5. Only the former are inhibitable.
passively acquired (Mele, 1992) and/or that automatic actions can be intentional (Wu, 2013), I would like to propose a much simpler solution:

Intentions are just uninhibited desire-motives\(^{39}\).

If true, whether an action counts as intentioned is unaffected by whether it is preceded by deliberation. What is relevant is whether the action is motivated. An intentioned action is a motivated response to the current state of affairs. In automatic actions, the lack of a considered alternative course of action implies that the desire-motive was never inhibited. It was activated and immediately implemented. A desire-motive that is born uninhibited is a desire-motive that is born as an intention.

Having set out what is meant by inhibition, and why James was right to call it ‘an essential and unremitting element of our cerebral life’ (1890), I will now return to the main question of this chapter: Are intentions distinctive attitudes or are they just uninhibited desire-motives?

In the following section I will analyse Bratman’s three features of intentions, commitment, stability, and coherence. I will find that these features can be accounted for without the need for any additional intrinsic properties that differentiate intentions from uninhibited desire-motives. I will therefore conclude that intentions are not distinctive attitudes.

3. Analysing the Features of Intentions

3.1 Plan-representations and Commitment

\(^{39}\) I am not the first to make this argument. Lemaire (2012) suggested that ‘intentions are predominant uninhibited desires with plans as their content’. 
Goschke described an intention ‘as a state in which an agent has a mental representation of a goal (e.g. the light turns on) and an action plan suited to achieve the goal (flipping the switch) and is committed to executing the action’ (Goschke, 2013, p412). However, as Bratman describes, the plan-representation contents of intentions are often only partial: ‘As time goes by I fill in my plan with specifications of means, preliminary steps, and more specific course of action’ (1999, p29). The details of plans are filled out as needed. The intention to call my sister now necessarily includes a more detailed plan than the intention to visit her sometime next year.

Although intentions are ‘categorical’ whereas desire-motives are ‘hypothetical’ (Kriegel, 2015, p86), the partial plan-representation contained within a newly formed intention is necessarily the same as that contained within the winning desire-motive. To deny this would be to assert that an agent can form an intention to C, as the result of evaluating plan-representations A and B, but not C.

Of course, there are many scenarios in which C will emerge as an alternative during a process of deliberation that starts with consideration of only A and B. Perhaps C is a compromise or a version of A that maximises the chances of later B-ing. Perhaps the possibility of C-ing only came to mind later as a third alternative. Indeed, the process of deliberation, if extended, is a chance to interrogate, fill out and enhance partial plan-representations and consider further options. However, in all of these scenarios the desire-motive to perform action C is present in advance of the decision being made, and, at the moment of intention formation, is valued more highly than the alternatives. The plan-representation content of the chosen desire-motive necessarily becomes the plan-representation content of the post-decision intention.
Once the decision-making process is complete, implementation of the chosen plan-representation proceeds. Unless the agent reopens the question and reconsiders her intention, she is committed to performing whatever action is specified by the plan-representation, as well as finessing the details of the plan-representation as and when necessary. This allows that there may be an extended period of inaction post intention formation, but only if that is an element of the plan-representation.

However, even after the question is settled, an unchosen desire-motive may continue to exert a conative pull, which must be inhibited. Inhibiting or resisting such an obstinate desire-motive is experienced as effortful and is described by Massin as an effort of will (2017). It is the effort felt in resisting the ongoing desire to do otherwise.

The ability for an unchosen desire-motive to persist beyond the moment of decision, combined with the continuity of the chosen plan-representation, is consistent with the idea, raised in section 2.1, that the only difference between desire-motives and intentions is the on-off switch of inhibition. In this interpretation, hypothetical means inhibited and categorical means uninhibited. Being committed to executing plan-representation A means only that implementation of plan-representation A is not inhibited.

In some common language uses of the word, commitment has a strong meaning, implying that an agent is inflexibly bound to a decision. Committing to a partner in an Anglican marriage, for example, is famously accompanied by the vow to love and to cherish, for better, for worse, for richer, for poorer, in sickness and in health, till death us do part. However, this absolute inflexibility with respect to reconsideration is not a feature of intentions. Agents are very much expected to rationally reconsider
intentions as facts change. Indeed, public vows or private resolutions exist because intentions can sometimes be too susceptible to reconsideration.

Despite this, it is possible that there is a middle ground, in which intentions, whilst open to reconsideration, exhibit what Bratman has called stability. The moment of intention formation may involve not just the release of inhibition, but also the addition of an intrinsic partial resistance to reconsideration. This would be an argument for intentions as distinctive attitudes.

I will analyse the drivers of intention stability in the following section. However, before moving on it is important to note an initial finding: Bratman’s first feature of intentions, commitment, is not, in isolation from stability, an intrinsic feature which differentiate intentions from uninhibited desire-motives. Rather, it is an extrinsic feature of the wider cognitive infrastructure. The inhibition, or not, of conative attitudes is analogous to the extrinsic rope in the story of the frightened horse.

If commitment in the context of intentions, implies stability, then it remains possible that intentions are distinctive, but that distinctiveness will be a function of stability rather than of commitment per se.

3.2 Stability

A central element of any account of intentions as distinctive mental attitudes is that they have an intrinsic, albeit partial, stability which means they resist reconsideration.

I will argue that, although the observation that intentions are stable is accurate, this stability is not a function of anything intrinsic to intentions which differentiates them from desire-motives. Instead, intention stability is a function of three features of the
post-decision landscape. These relate to plan-representation implementation, evaluative consistency between pre-decision deliberation and post-decision reconsideration, and the competitive balancing of flexible versus focused attention. The latter has been called the implementation mindset (Heckhausen and Gollwitzer, 1987). I will consider these three features in turn.

3.2.1 Plan-representation Implementation

Once a decision is made, the chosen plan-representation begins to be implemented. This is a function of the commitment element of intentions, or in the language of this chapter, the fact that the chosen desire-motive is uninhibited.

Implementation of a plan-representation will often lead to the unchosen alternative becoming relatively less attractive, or even impossible. For example, if there are two similar routes home and a driver is at a junction, it may be a difficult choice as to whether to turn right or left. However, once she has chosen and made the turn then, other things equal, the further she drives the less motivated she will be to change her mind and return to the junction. Indeed, in certain decisions there is no way back. If an agent chooses to eat a cake rather than resist temptation, then once she has swallowed the cake it is impossible to undo her action. In that instance the chosen plan-representation becomes infinitely stable due to the impossibility of reversing time and doing otherwise.

Sometimes however, plan-representation execution will not increase stability. In the above example of the drive home, if the driver decides her route in advance of the junction, then execution of that plan-representation will only add stability once she passes the junction. Before reaching that point, plan-representation implementation
contributes no stability. Equally, if the agent under temptation decides not to eat the cake, then the opportunity of doing so remains. To increase stability, her plan-representation needs to be more than just “resist”. It must include actions such as leaving the room or throwing the cake in the bin.

The proposed link between stability and plan-representation implementation suggests two thoughts. First, this instance of stability exists along a spectrum and is a function of, amongst other things, changes in the external environment. It is not a binary all-or-nothing property of intentions. Second, the cake-eating example hints at an inverse correlation between stability and efforts of will. Actions such as eating the cake or leaving the room increase stability, but they also reduce the effort required to resist the unchosen desire. Plausibly this is also the case in the driving example. As she passes the junction, the driver may feel torn, but as the junction fades in the rear-view mirror so will the conative pull of the alternative option.

The inverse link between stability and efforts of will is explained by the impact of plan-representation implementation on the space of attainable possibilities. If an agent only desires what is attractive and attainable, then making something unattractive or unattainable removes the motivational pull. If efforts of will imply that an unchosen motive is being resisted, then the fading in strength of that unchosen motive will naturally reduce the required effort.

I have shown that some aspects of stability can be explained by considering the implications of plan-representation implementation, but I believe that there are two other complimentary explanations. I will first analyse the impact of what I call

---

40 Note that stability is positively correlated with cognitive effort, as discussed in section 3.2.3.
“evaluative consistency” between the processes that evaluate whether to allocate cognitive resources to pre-decision deliberation and to post-decision reconsideration.

3.2.2 (Meta-)Evaluative Consistency

Every day experience tells us that even if intentions resist reconsideration, this resistance is only partial. Consistent with this, all Bratmanian accounts agree that agents are able to rationally reconsider intentions as facts change. Whether or not to reconsider is a metacognitive question. It is the higher-order question of whether to re-open the lower-order question of what to do. 41

In the following analysis I will set aside scenarios where the change in the information set is so striking that it immediately causes an agent to change her mind about what to do. Like the question of whether to prolong deliberation, the question of whether to reconsider is only relevant when uncertainty exists. If there is no uncertainty about what to do then there is no need for either deliberation or reconsideration.

In section 2.2 (and Hall, 2023) I argued that humans perform a pre-decision metacognitive cost-benefit computation, which I called evaluation M. 42 This compares the value of uncertainty reduction to the cost of allocating cognitive resources in service of uncertainty reduction. In the pre-decision phase, evaluation M concerns whether to prolong deliberation or settle the question. I would like to suggest that this same evaluation methodology is parsimoniously applied to the post-

41 This metacognitive model is supported by Holton’s discussion of resolutions as involving ‘a conjunction of two simpler intentions, one first order and one second-order (ie an intention about an intention)’ (2009b, p26)
42 This will also be set out in detail in chapter 9.
decision phase, allowing an agent to (usually subconsciously) evaluate whether to
stick with the decision or reopen deliberation.

Imagine that the question of whether to reconsider is raised. Whilst the pre-decision
evaluation M questions whether prolonging deliberation is an appropriate use of
limited cognitive resources, a post-decision evaluation questions whether
reconsidering is an appropriate use of those resources. The pre-decision evaluation
concerns whether uncertainty has been reduced enough to settle the question, whilst
any post-decision evaluation concerns whether uncertainty has increased enough to
re-open the question. Using the same evaluative processes for these pre- and post-
decision phases is what I am calling evaluative consistency.

I will first argue that evaluative consistency itself causes intention stability, before
considering whether it is possible for something intrinsic to the intending attitude to
enhance stability without reducing consistency:

The evaluation M framework suggests three reasons for an agent to rationally
reopen a question: She might believe that (i) uncertainty has risen; (ii) that stakes
have increased, or were underestimated in the initial evaluation; or, less discussed,
(iii) that competition for cognitive resources has diminished, so it is a better time to
deliberate43. Although the weighting of these different reasons in different
circumstances may be debated, if reconsideration is driven by any of them, then it
can be considered rational reconsideration. It is noteworthy that none of these
justifications for rational reconsideration relate to the underlying question of what to
do. Instead, they relate to the meta-question of whether or not to reopen the
question.

43 If there is a fourth reason, then this does not undermine my argument.
In cases where all relevant factors, such as the level of uncertainty, are unchanged from the original decision, it seems clear that the answer to whether to reconsider should be in the negative. To make one (meta)-decision to stop deliberation, and then to make a second (meta-)decision to reopen deliberation, when nothing internally or externally has changed, would be irrational. Indeed, the general argument that an agent should not change her mind unless there is a change in her understanding of the relevant information set, is an argument for the application of a single evaluative methodology to pre- and post-decision phases. As well as being cognitively inefficient, an entirely different process for post-decision evaluation would increase the risk of a loop in which questions are repeatedly closed and then reopened. This would harm stability. In situations where the uncertainty and other factors are unchanged, evaluative consistency is not only rational and efficient, but it also enhances intention stability.

An additional implication of pre- and post-decision application of the same Evaluation M would be a correlation between how hard a decision is to make, and how likely it is to be reconsidered. If the value of uncertainty reduction is high in the initial decision, then not only will making that decision be effortful, but the increase in uncertainty required for rational reconsideration will be minimal. In contrast, easy decisions imply that circumstances have to change significantly in order for reopening the inquiry to be rational. Although I am not aware of any philosophical analysis of the claim that the stability of an intention is correlated with the initial difficulty in settling the question, intuitively it seems reasonable and consistent with contemporary neural models. For example, in the case of Sartre’s young man, the stakes are so high

---

44 See Friedman, 2019, for an argument that incessant checking is epistemically problematic.
45 Recent work combining human magnetoencephalography with behavioural and neural modelling suggest that confidence modulates the extent of neural post-decision processing, with higher confidence driving lower
that resolving indecision in the face of ongoing uncertainty is borderline impossible. If he was finally able to force himself to make a decision, then he would presumably continue to vividly feel the attraction of the unchosen option. The desire to reopen the question would be strong and the stability of the intention low. Equally, consider a student who finishes a maths paper with ten minutes to spare. Cognitive resources are now available. It would be natural for her to review the questions about which she was uncertain (but thought progress was attainable) rather than those she found easy.

As with plan-representation implementation, the stability that would result from a consistent evaluation M is neither all-or-nothing nor intrinsic to intentions. It is purely a function of the costs and benefits of using cognitive resources to reduce uncertainty. Between plan-representation implementation and evaluative consistency it is clear that some stability can be explained without the need to assume that intentions are anything more than uninhibited desire-motives.

However, this doesn’t rule out the possibility that there is some extra stability that remains to be explained. This is Holton’s view: ‘Stability can best be understood as a shift in the threshold of relevance of information: some information that would have been relevant in forming an intention will not be sufficient to provoke rational reconsideration once an intention has been formed’ (2009b, p18-19).

This raises the question of whether the shift can occur within the evaluation M framework.

changes of mind, and vice-versa (Atiya et al, 2019, Rollwage et al, 2020). This can be shown to be adaptive if (and only if) the evaluation of confidence in the initial response is correct (Rollwage and Fleming, 2021).
The answer is yes. As Bratman has argued, an agent can have a policy with respect to the question of whether to reconsider. For example, she might place a high value on consistency and believe that once a decision is made it should only be reconsidered in extreme circumstances\textsuperscript{46}, or she may have made an explicit resolution. Such a policy would add an extra desire-motive that weighed against reopening a question, post-decision, but did not weigh in favour of settling the question, pre-decision. Evaluation M is a cost-benefit computation. As such, consistent inputs will give a consistent output, but changes in the inputs may change the output. However, it should be clear that a stabilising policy is certainly not an intrinsic property of intentions. Rather it adds an additional desire-motive about intentions, that is incorporated into the (meta-) evaluation of whether to reconsider. This is the case if the policy is subconscious, for example if it was learned as a child and has been subconsciously implemented ever since, or if the policy is explicit, such as a resolution or an oath. Neither indicate that intentions have intrinsic stability. In fact, the use of explicit policies indicates that the agent believes that intentions are sometimes not stable enough.

It is likely, therefore, that in the quote above Holton had something stronger in mind: The idea that, even in the absence of an additional desire-motive, there is something about having an intention that shifts the threshold.

Two changes to the cognitive infrastructure could achieve such a result. The first relates to the post-decision evaluative process, and the second to intentions themselves:

\textsuperscript{46} ‘This lady is not for turning’ Margaret Thatcher, 1980.
A change in the evaluative process could make the evaluation of whether to reconsider, similar to but entirely distinct from, evaluation M. This would reduce evaluative consistency, and would be highly cognitively inefficient. However, an argument for evaluative inconsistency could be made if there was empirical evidence for distinct processes, and/or if inconsistency was necessary to explain outcomes such as intention stability. In the absence of evidence for either the mechanism or the phylogenetic and ontogenetic development of such a separate process, I will proceed, for now, on the assumption that the same evaluation M is used to consider whether to close the question and whether to reopen it. The rest of this chapter will confirm that no explanatory gap, which might undermine such an assumption, remains.

The second suggestion could be cached out in terms of a boost to motivational strength that occurs when an intention is formed. Due to this boost, an agent would be more motivated by an intention to A than by a desire to A. This would be consistent with the argument that stability is an intrinsic property of intentions and intentions are distinctive attitudes. It would lead to the possibility that an occurrent desire to ¬A could have a motivational strength greater than that of the desire-motive to A, but less than that of the intention to A.

Although plausible at first glance, the motivational enhancement idea raises three questions, only the first of which has a satisfactory answer:

First, how does increasing the motivational support for A-ing impact the meta-level evaluation of whether to reconsider? The answer is that the motivational boost will

---

47 In fact, evidence suggests a single process: ‘Higher confidence in an initial decision reduced the probability participants would change their minds in the face of new evidence (behavioural study 1: $\beta = -2.0$, $p < 0.0001$, see Supplementary Figure 5 C), while controlling for choice accuracy and reaction times’ (Rollwage et al 2020, supplementary note 5)
have an indirect rather than direct impact on evaluation M. Enhancing the motivational support for A-ing will, other things equal, reduce uncertainty about whether to proceed with A-ing, thereby suppressing the likelihood of reconsideration.

Second, how is the motivational strength enhancement sized? On the one hand it seems empirically incorrect to suggest that any enhancement is a constant across intentions, scenarios and time. On the other, proposing that enhancements vary with (external-to-the-attitude) context seems more consistent with the claim that stability is an extrinsic rather than an intrinsic property of the attitude⁴⁸.

Finally, is there any neurobiological evidence of motivational enhancement implementation? As far as I am aware, there is not. Neither leading accounts of the neuronal underpinnings of motivation (Berridge, Robinson, and Aldridge, 2009), nor neuroeconomic accounts of decision making (Paolo-Schioppa et al., 2011, Rangel and Hare, 2010) propose any such enhancement. In fact, Bronfman et al.'s computational modelling of post-decisional effects found that ‘a model that postulates a shift of value for the chosen alternative is unable to account for several of the observed phenomena’ (2015, p8). For the claim that cognitive infrastructure implements motivational enhancement to be justified, it would need to be supported with evidence.

Although not necessarily wrong, both proposed changes to the cognitive architecture have weaknesses, not least the lack of evidential support. Perhaps more importantly, neither are necessary. The strengthened stability that Holton highlights can be explained without the need for motivational enhancement or new evaluations. Instead, intention stability is a natural consequence of an agent’s cognitive resource

⁴⁸ Of course, motivational strength can change as details of the plan representation are filled in and updated, but the question here is whether the strength of the attitude changes even with no change in content.
limitations. When in an implementation mindset the agent uses her limited resources to focus attention on execution and therefore has less available for gathering new information and considering alternatives. This reduces the likelihood of the increase in uncertainty necessary to provoke reconsideration.

Having argued that evaluation consistency partially explains intention stability when the agent was confident in her initial decision, in the following section I analyse the claim that this stability is enhanced by the implementation mindset.

3.2.3 The Implementation Mindset

To support the view of a threshold shift, Holton relies on the empirical finding that agents can be differentially impacted by changes in the external environment depending on whether they are in the pre- or post-decision phase of cognitive activity.

Limited cognitive and temporal resources mean that humans are unable to incorporate every standing belief and desire into every deliberation. Instead, consistent with the motivational strength idea, the decision-making process limits itself to motives that are occurrent\(^49\). Evolution has honed and stabilised the process which selects for and activates context-relevant motives.

What is context-relevant depends not only on the external environment and the agent’s capabilities, but also on what the agent is doing. When the agent is deliberating, she is (Heckhausen and Gollwitzer, 1987) focused on assessing and comparing the value and likelihood of alternative courses of action, and receptive to

\(^{49}\) The Stanford Encyclopedia contrasts standing and occurrent desires by stating that the latter are playing some role in one’s psyche at the moment.
incoming information. In the implementation phase, however, focus should, and does (ibid), shift towards the procedural considerations of implementation. Cognitive resources are funnelled towards execution of the plan-representation.

As Holton notes, there are two possible interpretations of Heckhausen and Gollwitzer’s findings:

‘The first is that stability is a feature of intentions, but the effect of contemplating an intention is to cause agents to impute the stability to other states.’ ‘The second interpretation is that stability is not a feature of intentions themselves, but is solely a feature of the mindsets that agents bring to intentions’. ‘I don’t know which interpretation is right.’ ‘I will go on speaking of intentions as stable, meaning this to be understood as either that they have a degree of intrinsic stability, or that they engender a mindset that treats them as stable’ (Holton, 2009, p22)

Recent research supports the idea of intention stability, but explained by Holton’s second interpretation rather than his first. One example is the previously discussed observation, that focused attention lowers the ability for quick and flexible reconfiguration to a new task when the environment changes (Musslick and Cohen, 2021). This assumption, that focused cognition has an opportunity cost, is an integral part of Kurzban et al.’s theory of cognitive effort (2013) and of evaluation M.

Focus on implementation biases attention towards information relevant for execution and shields plan-representations from the activation of competing motivational tendencies. This targeted focus makes an agent less receptive to peripheral information, which naturally reduces the likelihood of plan-representation shifting.
Due to the specifics of their experimental setup, Heckhausen and Gollwitzer (1987) discovered that, although motivational thoughts declined in favour of procedural thoughts once the participant was in the implementation mindset, some motivational thoughts remained. They argued that, because their experimental task required ‘relatively little procedural thought, it seems justified to assume that the capacity still available for other concerns was quite extensive’. This spare capacity allowed subjects to continue ‘pondering their decision at least during the initial stage of the post decision period.’ (p111), as well as to engage in task-irrelevant thoughts.

This suggests a second link between plan-representation implementation and stability. Not only does implementation increase stability when unchosen alternatives become less attractive or attainable, but also when complex tasks demand significant cognitive resources. Stability here arises because cognitive capacity is limited and execution of the complex task is starving background monitoring tools of resources. In contrast, less demanding tasks require less resources and therefore leave greater capacity for consideration of alternatives.

In moderation, intention stability is adaptive, as it increases the likelihood and efficiency of plan-representation execution. Even if a task is not cognitively demanding, it may still make sense to down-weight background distractions. In the extreme, however, an agent’s focus on implementation can make her dangerously unresponsive to changes in the environment. Single-minded focus on finishing an essay is beneficial when the background noise is a party upstairs, but not if it is a fire alarm.

In a dangerous environment an agent might be willing to accept poor execution in return for greater vigilance. Even if the chosen task is complex, and would normally
be allocated significant resources, in some contexts this allocation should be resisted.

This need for balance reveals that it is wrong to assume that the demand for cognitive resources in service of implementation will always be met, and that only what remains will be available for pondering. Instead, the allocation process is competitive. Whilst resource-demand does increase with implementation complexity, resource-supply will be limited by competing demands. If the context suggests that resources are better used in horizon scanning, then it is implementation focus that will be starved.

The cognitive challenge is to find the appropriate balance between two adaptive but antagonistic control functions - goal-shielding and background monitoring - each of which have costs and benefits (Goschke and Dreisback, 2008). The lower the susceptibility to interference the greater the risk of overlooking significant information, and vice-versa.

Adding complexity is the fact that the appropriate balance is not a constant but is specific to a given task and situation. This suggests what Hommel has called a ‘Meta-control State Model’ (2015) which, based on the internal and external context, aims to optimise the cognitive environment within which an agent pursues intentioned action. Empirical data has shown that immediate situational cues such as context and performance-contingent reward do indeed bias control towards persistence or flexibility, but that more global factors such as affect also have an impact (Fröber and Dreisback, 2018, Dreisback and Fröber, 2021).

Further work is necessary to fully understand the nuances and implications, but the model that is emerging, along with evidence that changes in control states can be
sticky and persist from one task to another (Hommel, 2015) supports the claim that meta-control determines the mindset, which impacts intention stability.

If true this justifies the claim that motivational attitude stability can increase post-decision, even if the evaluation process is unchanged.

Evaluation M will give the same result on the same information. If the implementation mindset starves monitoring tools of resources or down-weights sensitivity to subsequent information (Bronfman et al 2015), then this reduces the likelihood that new information influences the cognitive process, thereby decreasing the chance that the output will change. The link between evaluative consistency and intention stability is increased further when monitoring tools are starved of the ability to capture new sources of uncertainty when they arise. Of course, one hopes that if there is a dramatic change in the environment (a fire alarm!) then this will be picked up no matter what the mindset. If so, the new information will feed into the evaluation and cause even the most dedicated student to reconsider.

Although Holton is right to argue that once an action is settled upon there is a ‘shift in the threshold of relevance of information’ (2009b, 1p8), this shift is driven by the mindset. It does not imply that intentions themselves have intrinsic stability.

Having argued that neither commitment nor stability are a function of factors intrinsic to the intending attitude, I now turn to the coherence requirement for plan-representations to which the agent is committed. This discussion will be brief.

3.3. Coherence
Whilst doxastic attitudes are subject to the coherence requirement, conative attitudes can persist even if they conflict with one another. When two directly conflicting motives persist within a single agent, they ready her to implement two inconsistent actions. Conative architecture controls the transmission from readiness to action with inhibitory gates. Implementation only begins for motives when the gate is lifted: when they become uninhibited desire-motives.

Bratman stated that any individual intention ‘becomes part of my web of intentions and plans, a web subject to the plan-type demands of consistency and coherence’ (1999, p32). Intentions are therefore different from other conative attitudes in that they are subject to consistency constraints, just like doxastic attitudes. But this coherence is not dependent on something intrinsic to the intending attitude. Rather, the winner-takes-all model, implemented via inhibitory gates, is itself an incoherence filter. This is because when a conative attitude towards a plan-representation is released from inhibition the plan is immediately implemented in accordance with the plan-schedule. Whilst an inhibited desire-motive is a “want to do”, an uninhibited desire-motive is an “am doing”, and an agent knows that she physically cannot perform two inconsistent actions at once.

Instrumental rationality refers to the ability of a motive towards an end to transmit its force across the means-end relation so that the agent is motivated to execute the means. Let means-end rationality can be expressed as follows:

If an agent is motivated to $E$, and knows that $M$-ing is a means to $E$-ing, then the force of her motivation to $E$ is transmitted towards $M$-ing

For example, if an agent wants to complete a triathlon, but can’t swim, then her desire to complete a triathlon will transmit its motivational force towards the desire to
learn how to swim. Importantly, means-end rationality applies equally, but in the opposite direction, to harmful actions:

If an agent is motivated to $E$, and knows that $H$-ing harms the chances of $E$-ing, then the force of her motivation to $E$ is transmitted towards refraining from $H$-ing

Thus, if an agent is aware that it is physically impossible to implement two plan-representations towards which she has positive conative attitudes, then this creates a direct motivational conflict. This is the case even if the conflict is not immediate but at some future time $t$. The agent has an open question to answer: What to do at time $t$? And that question is answered through a motivational conflict. Until a decision is made, both alternatives are hypothetical. The near-term, non-conflicting elements of both plans may be uninhibited, but the longer-term conflicting elements cannot both be implemented so an agent cannot be committed to them both.

The coherence of uninhibited desire-motives is a function of commitment, realised through the release of inhibition, and rationality. Although intentions are subject to the coherence requirement, this does imply the existence of any previously unexplored extra intrinsic feature of intending attitudes.

4. Conclusion

In this chapter I analysed the popular view that intentions are distinctive attitudes. I confirmed that Bratman’s three key features of intentions - commitment, stability, and coherence - are extrinsic rather than intrinsic properties of the intending attitude. I therefore concluded that intentions are just uninhibited desire-motives.
This was the first major argument of the thesis and was necessary for three reasons.

First, any analysis of agency must engage with the concept of intentions. A reader getting to the end of a thesis that purely focused on desire-motives would be forgiven for thinking that one of the most important elements had been ignored. By showing that an intention is merely an uninhibited motive, I simplify the model, and the writing. From now on I will use the word motive rather than desire-motive.

Second, a central pillar of this thesis is the motivation strength idea, and intentions are clearly motivational. The fact that intentions are just ordinary motives is important because it removes any pressure to adjust the motivational strength idea to accommodate an intrinsic property other than strength. I can proceed, for now at least, on the assumption that a model of decision-making can treat all activated motives equally, in the sense that the only factor relevant to a motivational contest is motivational strength.

Finally, by highlighting the importance of inhibition, I began to set up my account of conative phenomenology that will be developed in part II.
Introduction

In this chapter I defend the motivational strength idea in the context of what I have called the puzzle of relative strength. This can be set out as three seemingly inconsistent claims:

1. When making a choice, an agent always chooses to do what she is most motivated to do (The motivational strength idea)
2. The phenomenological intensity of an unchosen motive can be greater than that of a chosen motive (Experience)
3. The phenomenological intensity of a motive is a non-conceptual representation of its functional strength (Full Correspondence)

Through phenomenological analysis I reveal that phenomenological intensity and functional strength do have a correspondence relationship, but it is conditional. Premise 3 should be replaced with premise 4:

4. The phenomenological intensity of a motive is a non-conceptual representation of its functional strength if and only if it is inhibited (Conditional Correspondence)
This resolves the puzzle as 1, 2 and 4 can be simultaneously true. As such I remove a potential argument against the motivational strength idea.

1. The phenomenological intensity and functional strength of motives

David Chalmers (1996, Chapter 1) argues that some mental terms lead a double life, employed within two different conceptions of mentality: Phenomenological and psychological. In the first conception, mental states are characterised first-personally, in terms of their subjective feel, whilst in the second they are described third-personally, in terms of their causal and functional relations. Whereas a phenomenal mind is experiential, a psychological mind is functional.

Despite this cleaving at the joints, phenomenal and psychological properties co-occur. Conscious experience seems to be somehow tied to cognitive processing. Chalmers gives the example of a feeling of happiness: ‘Whenever one has the conscious experience of happiness, the functional role associated with happiness is generally being played by some internal state’ (p22). Perhaps experience arises from within the processing and/or is what Wegner described as ‘a window into the lovely machinery’ (2018, xvi).

Although not analysed in detail by Chalmers, motives are prime candidates for dual-concept status. These mental states, which Mele has called ‘action-desires’ (2003, p16)\textsuperscript{50}, play a key functional role in decisions and actions and also have a distinctive conative phenomenology, usually described as a pull, a draw, or an urge.

\textsuperscript{50} Action-desires always motivate the agent to engage in an action right now.
In what follows I will use the word motive to indicate that I am capturing all occurrent conative mental states, including intentions\(^5\), and excluding non-motivational wish-desires or dispositional desires. I assume that a motive is a conative attitude to a plan-of-action. The plan, which may be complex or extremely simple, detailed or partial, is the content, and a pro-attitude represents-as-good the plan.

The core idea, that I defend throughout this thesis, is the motivational strength idea, that a contest of motives is decided based on the strength of the competing motives. This can be expressed in slogan form: *when making a choice, an agent will always choose to do what she most wants to do.* Here is the same idea expressed more formally:

> When making a choice between A-ing and ¬(A-ing), the agent will choose to A if, and only if, at the moment the choice is made, the strength of her motivation towards A-ing is greater than the strength of her motivation towards ¬(A-ing).

For the avoidance of doubt, this allows for the possibility that, over an extended period of deliberation, the motivational balance will shift. In that case, the decision will be in favour of the strongest motive at the time of the decision, even if that was the weaker motive at an earlier stage of deliberation.

A wider model of agency combines the motivational strength idea with an assumption that motivational strength reflects the subjective value an agent attaches to a plan-of-action. Agents are motivated to achieve valued ends and the greater the value the greater the motivation. Pre-theoretically, the functional strength idea of motives seems to be supported by the fact that the motives are felt with differing

\(^5\) I argued in chapter 2 that an intention is an uninhibited desire-motive
intensity. An optimist might think that a unification of the phenomenological, functional, and valuational aspects of motives is within reach. I will call the model in which the functional strength and phenomenological intensity of motives both represent the subjective value of a plan-of-action, the unified model.

Unfortunately, after reflection and introspection, many philosophers have concluded that the functional strength of an action-motive is not revealed to the agent through its felt intensity. Indeed, Mele, arguably the pre-eminent thinker in the field, stated in 2003 that no recent advocate of the motivational strength idea (“MSI”) endorses such a view. Rather, “friends of MSI claim that a phenomenologically intense motive may be motivationally weaker than a competing motive with little or no phenomenological intensity” (2003, p162). This conclusion was reached after consideration of a number of examples along the lines of those set out below:

Take the case of Hubbard, who crossed the road on her way to the shops. In functional terms the story is clear. Hubbard was motivated to go to the shops because her cupboard was bare. To get to the shops, she needed to cross the road, so means-end rationality entailed that the motivational base of her motive to go to the shops was transmitted to a motive to cross the road. The fact that she did cross the road indicates that her motivation to cross was indeed greater than her motivation not to. However, the phenomenological picture doesn’t seem to align. Did Hubbard feel the motivational pull of a road-crossing desire? Quite possibly not. At most her felt experience lacked intensity. Under the unified view this would imply that the motive to cross the road was very weak. If so, even the slightest motive to do otherwise would have hijacked Hubbard’s journey to buy food. Although not impossible, this does seem at odds with lived experience. Agents do consistently
manage to cross the road to get to the shops, even in the absence of phenomenologically intense motives.

Arguably, even more powerful counterexamples to the unified model come from cases which highlight a seeming asymmetry between the phenomenology of a motive to succumb to temptation and that of a motive to do the right thing.

Consider Chloe, who is on a diet. Although she may really want to stick to her diet, and is in fact able to do so, this functionally strongest motive isn’t felt with the same phenomenological intensity as the motive to eat a pastry from the local patisserie. There is a mismatch between the relative functional strengths of the two motives and their relative phenomenological intensities.

These examples seem to indicate that the strength of a motive is not a unitary concept that is reflected across domains, so the unified model is false. However, all is not lost. Accepting this does not necessarily imply that the motivational strength idea is false. As the quote from Mele implies, friends of the motivational strength idea can believe that there is more than one dimension of motive strength. Believing that phenomenological intensity is separate and distinct from functional strength, and that the former is not necessarily an indicator of the later, allows friends of MSI to focus on functional strength and ignore the phenomenology. Nonetheless, the lack of unification is a setback, as it opens the question of whether the concept of motivational strength is vacuous. Mele suggests a response to this worry (2003, p164-5), but as I explore below, an extreme version of this line of reasoning leads to a conclusion that seems implausible from an evolutionary perspective.

To avoid vacuity, the functional motivational strength relevant to the motivational strength idea must relate to something real in the world. But how does one get a
handle on that if it is decoupled from conscious experience? Merely stating that the strength of a motive is given by the strength of an agent’s inclination to execute the plan runs the risk of circularity. An obvious answer is that there is a neuro-physiologically implemented context-modulated representation of the conative value of a considered plan of action. This valuation manifests in the functional strength of a motive.

However, granting the existence of a physiological realization of functional strength only serves to amplify the seeming unreasonableness of a separate and distinct phenomenological intensity. Is there also a neuro-physiological representation of phenomenological strength? If not, phenomenological intensity becomes untethered from physical reality. But if so, what benefit justifies the cognitive costs of sustaining these two measures? For there to be an adaptive benefit there must be some kind of correspondence relationship between phenomenology and function. The relationship can be biased, inverted, conditional or complicated, but not random. It therefore seems evolutionarily implausible that the human brain is structured to sustain two unrelated measures of value, one for the purpose of decisions and the other for the purpose of feelings52. Pushing this view to the extreme raises the question of whether it even makes sense to talk of functional strength and phenomenological intensity as two dimensions of the same mental state. Are they, in fact, single measures of the strength of two different mental states, functional motives and phenomenological motives? If so, then only confusion can follow from trying to learn about the former through introspection of the latter.

52 See introductory chapter (i) for a more comprehensive analysis.
Following the above reasoning from the premise that functional strength and phenomenological intensity are separate and distinct seems to lead to two conclusions. First, motive is not a dual-concept but two concepts, the latter of which is either untethered from physical reality or an evolutionarily expensive addition with no obvious benefit. Second, phenomenology is a fool’s game for those interested in the functional impact of motives on behaviour. Despite agreeing that the functional strength of a motive is not always experienced by an agent, I will reject both of those conclusions. I will do so by showing that functional strength and phenomenological intensity are not separate and distinct. Rather, phenomenological intensity is a non-conceptual representation of the functional strength of a desire motive if and only if a certain condition is met.

Phenomenological intensity is a window into functional strength, but a window which is sometimes blocked by external-to-the-attitude aspects of the cognitive machinery, and sometimes made opaque by other elements of the phenomenological landscape. This is a re-unification view in that (i) there is only one neuro-physiological representation of value at any moment in time and (ii) phenomenological intensity and functional strength, when they exist, are both reflections of that value. It implies that introspection on phenomenological intensity can give an insight into this value, as can empirical data from decisions made and actions pursued, but care should be taken, as the window only offers a conditional view (and empirical data is not always replicable and generalizable)⁵³. In some conditions the felt intensity of a motive is a reliable indicator of its strength. In other conditions, a motive can be functionally strong but phenomenologically silent.

---

⁵³ A full understanding is enhanced by combining all perspectives, as in the parable of the blind men and the elephant.
Elucidating the condition precedent for phenomenological reliability, as well as any nuances that impact an agent’s ability to isolate intensity, is the major aim of this chapter.

I will claim that the functional strength of a motive is revealed by its phenomenological intensity if and only if it is resisted, such that the source of resistance creates a spatial or temporal misalignment between behaviour and the plan content of the motive. I will then narrow this claim further, arguing that the source of resistance is within an agent’s own conative architecture. The functional strength of a motive is revealed by its phenomenological intensity if and only if it is inhibited. Successfully transitioning from a rough outline to a fully threshed out re-unified theory will achieve four things:

First, and most importantly, it will ensure that the re-unified model of motivation is aligned with everyday experience, and in particular the case studies which acted as counterexamples to the simple unified theory.

Second, the re-unified theory will remove an argument against the motivational strength idea.

Third, it will show that an agent is able to make correct inferences from her experience. This has adaptive benefits, enhancing a diachronic agent’s fine-grained understanding of her own flexible navigation through a landscape of affordances.

Here is a non-exhaustive list of benefits:

1. Monitoring of the strength of an inhibited motive over time: Even if a change in functional strength is not enough to change behaviour, it may still be valuable for an agent to know that the intensity of her desire to eat, for example, has increased over time. This is especially the case when changes
in intensity are not driven by something as simple and predictable as hunger. They may reflect unexpected developments to which a sudden change in the strength of desire could draw attention.

2. Comparing two inhibited motives at one point in time: Not only does this give an agent a window into her own motivational contest during intention formation, it also allows her to understand which of two inhibited motives would win in a hypothetical contest.

3. Remembering the strength of previously experienced motives, particularly those that were intense: Depending on the context, this may drive future behaviour towards or away from situations where this experience will be repeated.

4. Activating other attitudes that were previously dormant: These secondary attitudes may be supportive or conflicting. The activation may be automatic and/or part of a process of intentional self-control\(^54\).

The general idea is that although feelings don’t play a synchronic functional role, they do make a difference to a diachronic agent. They allow information to be stored in memory, to trigger other responses and to be used for prediction. Whilst one might be able to conceive of a zombie (Chalmers, 1996) that functions just like a human agent despite feeling nothing, this would only be possible if the zombie was able to collect and store this information about motivational strength. What seems highly implausible from an evolutionary perspective is for an agent to feel, collect and store information about conative phenomenological intensity that is unrelated to functional strength.

\(^54\) See chapter 5
Monitoring, remembering, and comparing one’s own shifting attitudes to a specific content, across time and context, are prerequisites for metacognitive control (Vierkant, 2022). This is because they enable an agent to understand that having a strongly pro- conative attitude at one point in space-time, doesn’t imply that she will endorse that same attitude from another vantage point. She may regret a previous motive, or dread a future one. Only an agent who is able to evaluate her own motives, and deem them sometimes inappropriate, will have the (metacognitive) desire to exert control over those motives. Metacognitive monitoring and evaluation are necessary for the exercise of intentional synchronic or diachronic self-control. This monitoring-evaluation-control feedback loop would be harmed, perhaps irreparably, if the functional strength of a motive was decoupled from its phenomenological intensity. Against this harm, there are no clear benefits of a cognitive architecture in which phenomenological intensity and functional strength are unrelated. A decoupled phenomenological intensity would only serve to confuse. It is not obvious why it would exist at all.

Finally, the re-unified theory makes it possible to address the why question of (conditional) conative phenomenology\(^ {55} \). While functional analysis focuses on the choice made, conative phenomenology offers insights into the value of inhibited alternatives. On a practical level this means that conative phenomenology is additive to an agent’s understanding of her own decision-making process, over and above any functional analysis. On a conceptual level the re-unified theory forces a gestalt shift in our understanding of conative phenomenology. Conative phenomenology is

\(^{55}\) Without having anything to say about the hard problem of consciousness.
silent with respect to seamless intention implementation but loquacious in the rough and tumble of motives that are frustrated, obstructed, and inhibited.

The rest of this chapter will be structured as follows:

In section two, I will consider and reject some alternative conditional models of phenomenological intensity which are based on elements internal to the attitude.

In section three I will propose and defend two propositions, which are integral to the relationship between phenomenological intensity and functional strength. The first, and most important, relates to whether action-implementation is fully aligned with the action-plan content of the motive. A motive that is being implemented seamlessly is not felt. Incorporation of this simple fact addresses most counterexamples to the simple unified model. The second proposition relates to the content of the motive. The overall phenomenology of an attitude to a plan-representation is impacted by the type of plan that is represented. A motive to grab a donut feels different to a motive to deliberate further or a motive to stay silent in class. Arguably, although the attitude and the content are conceptually distinct, it is practically difficult for an agent to isolate the intensity of the attitude from the other action-readiness aspects of the phenomenology. Ultimately, my argument will stand and fall by the truth of the first proposition. The second illustrates an openness to the idea of further complexity and opacity, but may also hint at an explanation for the primary proposition. Accepting the truth of the primary and secondary propositions allows one to make a further claim, that phenomenological intensity, when present, is a reliable if somewhat fuzzy indicator of the functional strength of an inhibited motive. The phenomenology of motivational pull offers a window that is translucent rather than transparent or opaque.
In section four I will consider the implications of the fact that, whilst functional analysis celebrates the victory of the strongest motive (indicating only that it was stronger than its competition), conative phenomenology reveals the strength of its weaker cousin, the inhibited motive. This phenomenological inversion encourages a new perspective on two paradigmatic experiences of agency: alternative possibilities and control. Conative phenomenology, the intensity of which reveals the subjective value of the road not taken, indicates that not only do attractive alternative possibilities exist but that the agent is pre-disposed towards them. A state of action-readiness is entered which, unless inhibited, would naturally lead to an attempt to realise those possibilities. The control revealed by conative phenomenology is control over those unchosen options. The agent is free to do what she most wants to do only because her pre-disposition towards (less) attractive alternatives is inhibitable.

Humans are flexible agents, motivated but not compelled by inhibitable motives. Conative phenomenology is evidence of this lack of compulsion. When conative phenomenology is intense it is evidence of the inhibitability of even a strong motive. Conative phenomenology is proof of the ability to say no. However, one must be cautious. This ability to say no is not due to the existence of a Cartesian homunculus, but due to the existence of a stronger motive.

But I am getting ahead of myself. Before getting excited about the implications, I will analyse and reject the idea that conative phenomenology is a function of something internal to the motive. In the following section, I consider three possible sources of internal-to-the-attitude conditionality. I will conclude that none of these proposals can explain the phenomenology.
2. Against internal-to-the-attitude conditionality

In section 3 I will argue that rather than being decoupled, phenomenological intensity and functional strength have a conditional relationship, where the relevant condition is whether the motive is inhibited. Only a motive that is inhibited has a conative phenomenology. Some might worry that the premise that phenomenological intensity and functional strength are separate and distinct was always a strawman, and that no philosopher denies some kind of conditional relationship between the two. If true, that is somewhat helpful to my case as it makes the idea much less controversial. Rather than having to convince readers of the existence of conditionality, my task is merely to uncover what form the conditionality takes.

In section 3 I will make my positive proposal, but before doing so I will analyse and reject some alternative conditional models which may, at first, seem appealing. The general idea, against which I will be arguing, is that whilst the umbrella term motive might be appropriate for discussions of functional strength, it is too broad for analysis of phenomenology. In this view phenomenological intensity reveals functional strength only for a subset of motives, and the concept of motives should therefore be carved in two. These might be called hot and cold motives, with only the former being phenomenologically accessible. There are three elements of a motive on which such a division could be based: its neuro-physiological implementation, its content, and its attitude. Could phenomenological intensity be conditional on any of

---

56 In reviewing this possibility, I will only consider motivational attitudes. If an agent believes that she should perform a certain action, but is unmoved by that fact, then this non-motivating reason for acting will not exhibit any conative phenomenology.
these aspects? If so, my proposal that conative phenomenology is conditional on inhibition, would be undermined.

2.1 Against neuro-physiological and content-based views of conditionality

One potentially intuitive idea is that only phylogenetically old, simple and/or appetitive motives (or cravings) are felt, whilst more sophisticated deliberative motives are not. This would be consistent with either a neuro-physiological or a content-based view of conditionality. From an evolutionary perspective, however, this proposal is not uncontroversial. If conscious awareness of the strength of a motive has any kind of adaptive value, then one would expect that phenomenological element to continue to be included within more phylogenetically recent implementations. Although not conclusive, this perspective justifies some scepticism.

More importantly, introspection reveals that simple/appetitive desires are neither necessary nor sufficient for conative phenomenology. Complex and phylogenetically recent motives can be felt. Consider the nagging feeling that one needs to get back to writing one’s essay, the pull of peer-pressure, or the grip of an oath. Equally, simple appetitive motives are not always experienced. When an agent reaches for food, the functional strength of her desire may not be accompanied by a phenomenological intensity. Sometimes she eats “on auto-pilot”. When Bargh and Chartrand speak of ‘the unbearable automaticity of being’ and argue that most of a person’s everyday life is determined ‘by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness’ (1999, p462) they are not solely, or even predominantly, talking about deliberative motives.
Perhaps one could take an almost opposite perspective and claim that there are certain kinds of motives that evolution or experience has routinised and made habitual, and these are not felt. This seems more plausible until one considers what happens when an agent encounters an obstacle. An agent who is on time when driving her usual route to work may experience no phenomenological pull, but if she gets stuck in a traffic jam then the conative pull will rise until she is up in her seat and shouting at the cars ahead. Her need to be on her way becomes more and more of a felt urge as it becomes more urgent.

There seems to be a serious problem for the neuro-physiological and content-based views of conditionality: The felt intensity of an attitude with a single content or physiological implementation can shift depending on the context. A single motive can be phenomenologically cold when things are proceeding seamlessly, but intensely hot when there are obstacles and/or time pressure. A defender of the neuro-physiological or content-conditional view would have to argue that phenomenology is conditional on these aspects and something else that is related to obstacles.

To be clear, I am not suggesting (and do not believe) that conative phenomenology as a whole is unaffected by the content of the motive. Here I am focused purely on the impact of content on phenomenological intensity. Whilst, as a practical matter, it may be difficult for an agent to compare intensities across different contents, as a conceptual matter intensity is orthogonal to content so one can never be impacted by the other.

In fact, there is only one case in which content and intensity are linked: When the content is null. An invisible colour does not have an intensity and neither does a motive with no content. I will return to this idea in section 3 as a potential explanation.
for my proposed solution. This isolated case, however, does not help the defender of the neuro-physiological or content-conditional view of intensity. I therefore conclude that these views are false and turn to attitude-based views of conditionality.

2.2 Against attitude-based views of conditionality

An attitude-conditional view of intensity would suggest that there are some mental attitudes that functionally motivate behaviour but exhibit no phenomenology. The most obvious attitudes to fill this role would be all-things-considered judgements and/or moral judgements (“A&M judgements”).

An initial concern with this view is that judgement attitudes have the wrong direction of fit to be motivational. For an A&M judgement to be intrinsically motivating it would have to include both truth-seeking and motivation-encompassing elements within a single non-decomposable attitude. Let’s call this kind of attitude a ‘besire’ (Kriegel, 2011). A besire would have to simultaneously represent-as-true (a plan of action being good) and represent-as-good (the plan of action). But there is a complication. Even if the two senses of good were aligned, the two representational attitudes have subtly different contents. The proposed besire attitude is not equivalent to representing-as-true-and-good (a plan of action). Rather, the two attitudes refer to two contents, and each attitudinal component must refer to the correct content\(^{57}\). The internal architecture necessary to ensure the appropriate attitude-to-content connectivity would effectively divide the besire into a compound of a desire and a belief, making the two theoretically separable. As such, there is no theoretical benefit

\(^{57}\) The wrong motivational response would follow from representing-as-good (a plan of action being good) and representing-as-true (the plan of action).
which would justify the existence of combined besieres over and above separate beliefs and desires.

In addition, it does seem to be a matter of fact that the cognitive and conative elements of A&M judgements are separable. As Mele argues (2003, p111-130, see also Carruthers, 2018), a listless person can be persuaded that she is morally obliged to perform an action but, because of her depression, still fail to be motivated to do so. A&M judgements are not sufficient for moral motivation. And, as implied by the etymology, listlessness impacts desires. This suggests\textsuperscript{58} that perhaps A&M judgements influence practical decisions only via antecedent motivations, such as the desire to do what one morally ought, and it is these antecedent motivations that are targeted by listlessness. In this proposal, a specific moral action is a means to the end of being a moral person. This externalist model has the benefit of efficiency in that antecedent motivations are presumed to exist in this and many other forms. Many agents do have the antecedent desire to do what they morally ought, as well as other antecedent motivations such as to adhere to social norms. In these cases, motivating A&M judgements would be duplicatory.

In conclusion, besires, if they exist, are (at least theoretically) separable attitudes, that are neither necessary nor sufficient for motivation.

This brief discussion makes the case that A&M judgements need not be intrinsically motivational. They can impact behaviour via ordinary motives. However, if it turned out that A&M judgements were directly motivational, would these besieres be functionally strong but phenomenologically cold? That seems unlikely to me. Part of the intuitive support for motivation-internalism about A&M judgements is precisely

\textsuperscript{58} It is only suggestive because an internalist can argue that, in the case of listlessness, besires vanish and are replaced by simple cognitive judgements.
their motivational “oomph”. Whilst evidence for internalism may sometimes take the form of functional third person analysis, motivation-internalists surely also rely, at least implicitly, on the support of conative phenomenology. If a moral agent sees morally repulsive behaviour, would she not feel a strong conative pull towards intervention? Imagine you see a person on the other side of a busy street, urgently in need of help. As the cars drive past would you not feel a desperate need to cross, an urge that would make you willing to risk your own life in the traffic? Obstructed besires, if they exist, have no less an intense phenomenology than ordinary motives.

It seems therefore that A&M judgements are either purely cognitive, in which case, from a directly motivational perspective, they are neither functionally strong nor phenomenologically intense, or they are besires, in which case they are both functionally strong and phenomenologically intense. A&M judgements do not support the attitude-conditional view of intensity.

All the talk of possible new attitudes has perhaps drawn attention away from a prime candidate that is hidden in plain sight, explicitly captured under the broad term motive. These are intentions. I argued in chapter 2 that there is nothing intrinsic to the intending attitude which differentiates it from other motives. An intention is merely an uninhibited motive. The claim I will defend in section 3 is that only a motive that is inhibited has a conative phenomenology. If both statements are true then that conative phenomenology attaches to all motives except intentions. But this seems a controversial claim. Uriah Kriegel, perhaps the leading contemporary thinker in the field of phenomenology, reached the conclusion that: ‘The fundamental form of our conative experience is the proprietary phenomenology of deciding-and-then-trying’ (2015, p72). Are these positions opposed? I believe not for two reasons:
First, Kriegel’s analysis refers to the moment of initiation, when the formation of an intention “fires up” a trying. An intention being implemented is temporally subsequent. It is the ongoing alignment of action with the plan content of the intending attitude. When alignment is achieved, the conative pull disappears. Kriegel agrees: ‘phenomenologically, the exercise of the will is not exhausted when a decision has been formed—only when the process of realizing the decision is underway’ (2019, p88). Kriegel’s phenomenologically silent ‘process of realizing a decision’ is my intention being implemented.

Second, Kriegel’s analysis is broader in scope than mine. He is interested in conative phenomenology as a whole, whereas I am focusing purely on the positive phenomenology of motivational pull and setting aside until chapter 4 the aversive experience of effort that can accompany a trying. There I will argue that, whilst a phenomenological pull or urge represents-as-valuable a hypothetical plan-of-action, the feeling of effort represents as costly (or controversial, or unsustainable) a trying. This second kind of conative phenomenology, whilst relevant for Kriegel’s analysis, is beyond the scope of this chapter. My narrow aim is to elucidate the relationship between the phenomenological intensity and functional strength of motives.

This analysis does raise questions about the “acceleratory” moment of intention formation, and what happens between the release of inhibition and plan-alignment, which I will analyse in section 3. For now, however, the important conclusion relates to attitude-based conditionality: There is nothing intrinsic to besires or intentions that explains the shifting phenomenological of motivational pull.

I therefore conclude that the conditional relationship between phenomenological intensity and functional strength is driven by something external to the attitude. The
question then becomes, what external factors are relevant? The answer, hinted at in the analysis and vignettes above, but explored in more detail in the following section, is that the relevant factor is whether the motive is inhibited.

3. The conditional model of phenomenological intensity and functional strength

3.1 The primary proposition: The phenomenological pull of a motive is absent when action-implementation is fully aligned with the content of the motive.

Simple motives are famously felt as “pulls” or “grabs”. As I will analyse in detail with respect to the secondary proposition, these feelings don’t just have an intensity, but also a directionality. An occurrent motive is “an affectively charged motivation toward a certain object, person, or activity” (Kavanagh, Andrade, & May, 2005, my italics). This directionality is a function of the plan-of-action that is the content of the motive. A motive for a donut exerts a pull that feels related to the action of grabbing a donut, a motive to run from a snake exerts a push away from the snake.

The correlation between conative phenomenology and the considered action-plan can potentially be explained by the fact that being motivated to perform an action is not theoretical but practical, in the sense of indicating a preparation or readiness for action\(^{59}\). The magnitude of what Kruglanski et al. call motivational readiness ‘generally tends to increase with the magnitude of an individual’s want state’ (2014, 261). This is consistent with a parallel rather than serial model of decision making, as proposed by Cisek\(^{60}\) (2007). Action systems are not engaged only after decisions. Rather, choices are made between plans-of-action that are readied in parallel. A

\(^{59}\) For a vivid illustration of physical action-readiness imagine a dog’s anticipatory saliva.

\(^{60}\) Cisek calls his model the affordance competition hypothesis. I will discuss affordance in more detail in section 2.2
decision releases into execution an action that has already been, at least partially, prepared. This model is not only evolutionarily plausible\textsuperscript{61} and consistent with the widely accepted view of decisions as the output of competitive accumulation to bound models\textsuperscript{62} (Smith and Radcliffe, 2004), but importantly for our purposes, explains the directional phenomenology.

However, once an agent has settled a question and is seamlessly executing the chosen action-plan, she no longer feels the pull associated with the winning motive. A relevant analogy is the pull of a rope on a person. Imagine a mechanism that pulls a rope at 5mph. An agent would feel a tug if she held the rope and tried to stand still, but not if she walked towards the mechanism at 5mph. The claim made in the primary proposition, is that whilst conative phenomenology of a motive rises with action-readiness, it disappears with action-plan implementation. But is this claim correct?

Introspection indicates that there is certainly a ring of truth: Empathise with Chloe, the dieter, faced with temptation. As she sits there, arms folded, resisting the appetising pastry, she will clearly feel the pull. If she didn’t like pastries then she would not, but she loves pastries and is hungry. However, if she gave in, reached for the pastry and was about to take a bite, the phenomenology would shift. In the process of executing the action of pastry-eating, the phenomenological pull is no

\textsuperscript{61} Evolutionary plausibility is supported by both considerations of phylogenetic continuity and adaptive performance: ‘Abilities such as sophisticated cognitive decision making did not appear from thin air, complete with appropriate anatomical connections and a full developmental schedule. They evolved within an ancestral context of real-time interactive behaviour’ (Cisek, 2007, p1595) and ‘with a set of such potential actions partially specified, the animal is ready to quickly perform actions if circumstances demand’ (p1586)

\textsuperscript{62} “These competitive populations of neurons are hypothesized to be motor patterns, which means they are the very same populations that are involved in the execution of a movement” (Bruineberg & van den Herik, 2021, p6)
longer present. This explains why Hubbard, on her way to the shops, didn’t feel an intense motive to cross the road. Motive’s that are in the process of being successfully executed are not felt. If the traffic was so bad that she got stuck on the pavement whilst car after car went past, then she might feel a growing motive to cross. She might feel torn between his motive to stay safe and his motive to cross, eventually throwing caution to the wind and making a dash for it.

Three clarifications should be made immediately to forestall objections:

First for the feeling of motivational pull to be entirely absent there needs to be a perfect alignment between the action that is implemented and the plan-of-action that is the content of the felt attitude. If Chloe wants to grab the pastry and devour it in one go, but has been brought up to behave with decorum, then she may slowly, one might say intentionally, reach forward and take a small bite. This well-mannered behaviour would not remove the phenomenological pull of its ill-mannered cousin. The rope analogy may again be helpful here. Walking towards the rope-pulling mechanism at 5mph removes the felt tug. Walking towards it at 3mph does not. The ongoing monitor-and-control loop of intentional action will likely contribute to a phenomenological background hum, rising in intensity in the face of a significant implementation challenge.

Second, the conative phenomenology of motivational pull is only one aspect of an agent’s conscious experience. She will also experience feelings of effort, of (un)confidence, of emotion, of pleasure or pain, and these will not be static. Depending on other aspects of her psychological make-up, these might trigger a significant shift in her evaluative landscape. After giving into a craving, for example, an agent’s feeling of desire might be replaced by horror, guilt or disgust. She might
then feel a rising urge to throw the rest of the tempting item out, or to make a resolution never to give-in again.

Third, intensity is only one dimension of conative phenomenology. This sets up the secondary proposition, which I will consider in some detail:

*The secondary proposition: An agent may struggle to isolate intensity from other aspects of her conative phenomenology.*

The overall conative phenomenology of each motive differs as a function of its content. The motive to eat a donut feels different from the motive to fall asleep, as it does from more complex motives such as the motive to maintain a resolution or to behave in ways that one believes to be socially acceptable. This link between conative phenomenology and the type of action being readied has been explored most actively in the affordance literature, where for the environment to ‘afford’ something to an agent means to provide or supply it (Gibson, 1979, p127). Under this framework trees afford climbing, amongst other things, and balls afford throwing. Affordances are relational, in that they depend on the abilities of the agent. A tree is climb-able, or a ball throw-able, only to an organism capable of climbing or throwing. Heft has argued that ‘part of what makes the notion of affordance so compelling is the phenomenology it presents’ (2001, p114) This is expressed clearly by McClelland:

‘The most vivid cases of affordance perception are those in which we experience an action not just as *available* but as positively *solicited*. Consider the way in which we experience a ringing phone as demanding to be answered, an itch as demanding to be scratched or a cake as demanding to be eaten’ (2019, p407).
Although it is conceptually possible to differentiate between the intensity of an attitude and its content, if the content presents itself as an element of the phenomenology, then introspectively isolating the intensity may become practically difficult. If my conative attraction to ball throwing and tree climbing have similar functional strengths, then I may struggle to introspectively order them by phenomenological intensity. It would be analogous to trying to determine whether a red or green colour was more intense. Of course, the relative functional strengths will be revealed if a decision is made. The point is that this functional outcome, observable in the third-person, may not always be predictable from a phenomenological, first-person, perspective.

This translucency would plausibly worsen in cases where the alternative is between physical and mental actions. Is an agent able to successfully rank the phenomenological intensity of the motive to climb a tree vs the motive to solve sudoku puzzles? Maybe the answer is yes if the difference in intensity is high, but not if it is marginal. Finally, there may be a difference between the felt intensity of positive vs negative motives, the motive to kick a ball vs the motive to refrain from doing so. If, as McClelland argues, affordances prepare or potentiate actions, how easy is it for an agent to compare the phenomenological intensity of a potentiated ball-kick from that of the motive to refrain? At the extreme, the intensity of an actionless motive may be as impossible to determine as the intensity of an invisible colour. If so, then the secondary proposition may contain an explanation for the primary proposition. I return to this idea in “explanation 1” of section 3.2.

The nuances of implementation alignment, motive content and the broader phenomenological landscape make the window into motivational strength translucent rather than transparent. However, they do not undermine, the primary claim I wish to
defend: That the phenomenological pull of a motive disappears when action-implementation fully aligns with the content of the motive. Given that a motive inclines an agent to execute the action-plan, then a lack of alignment must be a consequence of resistance. If there was no resistance, there would be no conative phenomenology.

In his 2020 paper, ‘Atoms of Self-Control’, Sripada introduced the idea of a brief and simple state called a response pulse, where ‘the psychological functional role of a response pulse is broadly akin to what philosophers call an “action-desire”, a desire to perform some action straightaway’ (p804). He then made the cautious claim that the phenomenology of response pulses ‘shows up most clearly under conditions of conflict’ (p805). I am merely strengthening this claim. Without conflict there is no felt pull. An unresisted motive is a motive that, from the perspective of motivational pull, is not felt.

If the primary and secondary propositions are true this opens the possibility of a further claim: When a motive is resisted, its phenomenological intensity reflects the functional strength of the motive. If true, this would support the motivational strength idea and remove the need for two decoupled neuro-physiologically implemented measures of value – an idea that, although possible, seems implausible.

In support of the conditional recoupling of phenomenological intensity and functional strength, consider the following scenario of a resisted motive that changes in phenomenological intensity over time:

Imagine an individual who is seeing how long she can go without food. In the beginning it is easy. Her motive for food is low and she feels little or no conative pull towards food eating. Over time she gets hungrier and hungrier, until eventually she
gives in, stops the clock and has something to eat. Introspectively, it is clear that the intensity of her felt desire for food over that period would have increased, until finally it was overwhelming. This increase may not have been linear. Perhaps just before the desire to eat became unbearable, she was able to distract herself for a while by listening to a song from her favourite band. But then suddenly the song finished and the phenomenological intensity of the motive came flooding back into her consciousness. Over that period, what was happening to the functional strength of her motivation to eat? It seems clear to me that it was not increasing in a way that was decoupled from the felt intensity. She was not at risk of being functionally overwhelmed by the strength of the motive whilst the song was playing and the phenomenological intensity was low. The introspectively supported and parsimonious solution is that the functional strength rose and fell entirely in line with the phenomenological intensity.

Equally supportive is the case of a competition between two resisted motives. Imagine a dieter who is faced with not one but two popular dishes. Perhaps due to the talents of the chef, one of the prepared dishes triggers something in the dieter and she feels an intense urge to pick up her chopsticks and tuck in. The other dish, despite being popular amongst her peers, leaves her unmoved. In this case, the intensity of feeling does seem to reflect the strength of the relative motives. If the dieter does give in to temptation, then it will be in favour of the motive that she feels most strongly.

Thus, it appears that there is a relation between phenomenological intensity and functional strength. The strong claim I am making is that this relationship is not coincidental or random. Rather, the relationship is conditional: If a motive is resisted then phenomenological intensity is a felt representation of functional strength. If not,
if there is full alignment between the action-implementation and action-plan content of the motive, then the felt intensity is zero. Cognitive architecture seems to have implemented the following “IF function” of motives:

\[
\text{Phenomenological intensity} = IF (\text{resisted}, \text{functional strength}, \text{zero})
\]

Some readers may find the introspective support for this theory sufficient, but others may be concerned with the question of why the mind would implement such a function. In section 4, I will discuss the benefits of such a conditional output, but first I will consider whether an acceleratory understanding of force supports the idea, and then tentatively suggest three possible explanations of the implementation. If true, these will provide further support for the primary proposition and the conditional model of unification.

3.2 The Acceleratory Model of Motivational Force

In understanding the conditionality of phenomenological intensity, it is helpful to clarify that the term motivational strength refers to the strength of motivational force, and to recall from Newtonian physics that forces are acceleratory (F=ma). When an agent is driving and wants to go faster, she presses the accelerator. If she wants to slow down, she hits the brake, which is a decelerator. Although the driver’s ultimate aim is a location in space and time, and at any point she is travelling at a certain velocity, her interventions are acceleratory.

Before proceeding it is necessary to acknowledge a significant difference between the experience of (external) acceleratory forces and the conative phenomenology of (internal) motivational pull. In the former it is the acceleration itself which is felt, whereas in the later it is the inhibited acceleratory readiness. To use the metaphor of a drag car, the conative phenomenology of motivational pull equates to the tyres’
screeching and smoking, while the brakes are on, rather than the acceleratory push in the driver's back once the brakes have been released. Positive conative phenomenology is in this sense counterfactual or predictive. The agent becomes aware of the acceleratory force that would be unleashed if she were to yield to her desire.

In order to transpose the idea of inhibited acceleratory readiness to conative attitudes, I must first analyse an element of a plan that has hitherto been ignored – its situation in time.

When an agent has a plan towards achieving a certain state of the world, it explicitly or implicitly addresses the question of when. She may want a pastry right now, or to speak to her boss before the end of the day, or to run a marathon in under three hours. If it ends up taking her a week to speak to her boss or eight hours to finish the marathon, she will not have aligned the world with her plan-of-action. As with other elements of a partial plan, the time specifications may be rough and filled out later, flexible and of secondary importance, or they may be a very specific primary focus. However, a plan with no specified timeframe for any element is not the kind of thing that can be the content of a motivational attitude. It is more like the content of a wish. As self-help books encourage, an agent who wants to achieve a goal someday should identify what she can do today in order to make progress towards that goal. This motivates her to do something specific in the short-term and increases her likelihood of achieving her long-term goal.

By locating a plan-of-action in time – by incorporating a schedule – the plan automatically includes a rate of progress. It is no longer just a plan to change the world, but a plan that incorporates a rate of change. To achieve valued ends, an
agent needs to change the world, but to do so she must have the ability to modulate the rate of change. That is what (acceleratory) motivational force allows her to do.

The link between motivation and the perceived current rate of progress, provides a conceptual bridge between the sometimes long-term nature of plans and the straightaway-ness of action-desires. Occurrent motivation is a response to a negative answer to the question of whether progress is being made as planned. It inclines an agent to do something now to get things back on track. If, on the other hand, perceived progress is in line with planned progress, then there is no acceleratory force needed. If an agent has agreed to call her father at 10pm then she feels no motivational pull towards the phone until that time. This suggests an explanation for the primary proposition: A motive is phenomenologically silent when progress is aligned with the plan because there is nothing (acceleratory) to do.

In simple cases where an agent has not yet started attempting to implement a plan it is obvious that an acceleratory motivational force is necessary. If Chloe is staring at a pastry and she wants a pastry, then the motivational force of her pastry desire attempts to shift the rate of progress towards pastry eating up from zero. This motive-consistent acceleratory force can be felt. At the other extreme, if pastry eating is proceeding flawlessly then no intervention by a motivational force is needed, and there is no felt intensity of pastry motive.

Nuances arise when implementation is proceeding, but not at the rate specified in the plan. This captures the case of well-mannered Chloe, eating her pastry slowly and politely, whilst feeling the motive to accelerate the process towards grabbing and devouring. This ill-mannered motivational force will, unless inhibited, accelerate her progress.
Progress misalignment is also relevant to the case of Hubbard and her journey to the shops. Assume that her schedule was relatively relaxed and she had made no explicit allocation of time for road-crossing. If the traffic on the road had been as expected and she had crossed without too much trouble, there would have been no (felt) motive to change the rate of progress. However, if the road was busy then, as time passed, her divergence from the implicit schedule would have started to make itself known. Hubbard will have felt an increasingly strong motive to cross, inhibited by her motive to stay safe. The less relaxed her schedule, the greater the opportunity cost of waiting further, and the greater the likelihood that the motive to cross would overwhelm her natural caution.

Readers who are otherwise sympathetic to the idea that motivation is importantly impacted by the temporal elements of a plan-of-action, may feel some concern about my openness to a lack of precision in the schedule. In particular, what does it mean for a temporal element of a plan to be implicit, and how can an implicit temporal element be motivational? Is a defender of the link between motivation and rate of progress not committed to the stronger claim that the plan-of-action must include explicit deadlines or milestones? It is tempting to accept this revision. If an agent has made a plan to call her elderly father before he goes to bed at 10pm, then if she suddenly notices that time has flown and it is 9:45pm, then she will jump up, innervated by an urgent desire to call her father right now. If, on the other hand, the time is 3pm then she would feel no such urgency. Her desire to call her father right now would be much more susceptible to being defeated by more pressing needs. Whilst this example supports the link between motivation and explicit temporal elements of a plan, does it undermine the case for implicit schedule? Here, reasonable people may disagree, and I encourage further analysis, but my intuition
is that there is at least one understanding of the word implicit in which the temporal element of a plan can be implicit but still motivating. This is when the temporal element of a plan is implicitly constrained by other background concerns of the agent.

What I have in mind is inspired by Bratman’s proposal that previously endorsed intentions and plans provide a ‘background framework’ (1999, p33) which constrains future intentions. Imagine that at the time Hubbard set off for the shops, her plan was partial in the sense that it did not include a schedule. If she had run into a friend on the way she would have been happy to stop and chat. Assume that she did so. Eventually she went on her way and go to the main road, but, as it was rush-hour, there was an endless stream of traffic. As she waited, she thought of her dog at home and started to feel a greater and greater urgency to make a dash for it. Although she had not explicitly thought about it in advance, her background concern about leaving the dog unattended, imposed a time constraint on the shopping trip. When a slow rate of progress starts to conflict with a background concern, the latter begins to constrain the plan-of-action.

Plausibly, the constraint doesn’t need to be as specific as an unattended dog. Human agents seem to have an inbuilt aversion to wasting time. From an evolutionary perspective, time and energy are finite resources that must be allocated wisely, so it is unsustainable to fail to make progress for an extended period. This perhaps justifies a universal background constraint of sustainable progress. If this threshold is not met then the agent should try harder.

63 There are multiple studies showing that people will choose unpleasant experiences or effortful tasks if they are bored (see Wu, Ferguson & Inzlicht, 2023)
64 I analyse the positive feeling of a high rate of progress in chapter 8.
I acknowledge that further work is needed to fully elucidate the implications of Bratman’s framework for conative phenomenology, and I remain open to the idea of a compromise position. Perhaps background concerns only impact phenomenology when they become explicit and/or time constraints are better described as indirect rather than implicit.

Sometimes progress will be misaligned with a plan-schedule because it is unachievable. From a rational and evolutionary perspective an agent’s motives should be modulated by attainability constraints. Sustaining an unachievable motive is not only a waste of resources but also a cognitive failure: it indicates a false belief that the plan could be realised. A rational agent’s plan-schedule should adjust to fit what is possible. Ricouer calls this ‘consent’: ‘converting within myself, the hostility of nature into the freedom of necessity’ (1950, p346). But human agents are not perfect beings, sometimes they will keep trying even if something is unachievable.

Imagine a skinny weightlifter. She may not be able to bench press 150kg, but she can always try. If she wants to (try to) lift the weight and is trying as hard as she can to lift the weight then she will feel no motivational pull, even if she fails to lift the weight. I will consider the implications of this example for plan-alignment and resistance:

Plan-alignment: Is the skinny weightlifter plan-aligned? Well, that depends. If her plan-of-action is to lift the weight, then her progress is very obviously misaligned, and will remain so until she gives up. Alternatively, if she has a more rational plan of trying to lift the weight, then, by trying her hardest, she has done her bit. She has discharged her decision to try and lift the weight. This second explanation is more consistent with the affordance literature, in which affordances are relational and
depend on the abilities of the agent. The weight is not lift-able, but it does afford trying-to-lift. Personally, I am neutral as to which view is correct. The former would imply that plan-alignment is sufficient but not necessary for an absence of motivational pull. The latter, if generalisable, would imply plan-alignment is necessary and sufficient. Kriegel has stated that trying is ‘the mental core of action’ (2015, p89). What the example of the weightlifter indicates is that it is this mental core that is relevant for feelings of motivational pull. This makes sense, as both tryings and conative phenomenology are entirely mental phenomenon, whereas action is not.

Resistance: It is natural to suggest that the skinny weightlifter faces external resistance. The downward force from the 150kg weight is ultimately what is stopping her from raising the bar. Having said that, this external resistance does not directly stop her from trying, nor does it cause a feeling of motivational pull. Her motive is psychologically uninhibited, and is phenomenologically silent. This can be contrasted with well-mannered Chloe, whose ill-mannered desire was felt because it was inhibited by its psychological cousin. Combining these two vignettes leads to the conclusion that the resistance that is necessary and sufficient for the experience of motivational pull is internal: It must originate from within the agents own conative architecture.

In section 2 I argued that motivational pull is dependent on external-to-the-attitude factors. I then argued that for a motive to be felt it must be resisted. Now I have clarified that the source of resistance, whilst external to the attitude, is internal to conative architecture. The functional strength of a motive is felt as motivational pull if and only if it is inhibited.
Despite this, a failure to implement the plan-schedule of motive is not necessarily phenomenologically silent. First, the skinny weightlifter will feel effort. Second, non-conative phenomenology may be powerful. A swimmer who is racing his hardest but can see victory slipping away, will feel unhappiness, anger, and the desire to train harder in future. Finally, as an agent implements an intention she will constantly recalibrate. If progress is misaligned, she may feel a rising desire to accelerate progress, or she may adjust her schedule, consenting to the new reality. This will be particularly relevant immediately after intention formation as predictions meet reality and, more generally, if there are any implementation surprises. In all cases, however, motivational pull is felt only when an agent has a motive, new or updated, that has not yet been released from inhibition.

For the rest of this chapter (and thesis) I will proceed on the basis that phenomenological intensity of motivational pull is zero when progress is motive plan aligned, and that misalignment, to the extent it is relevant from conative phenomenology, is a consequence of inhibition. This conclusion supports both the primary proposition and the conditional model of unification.

Taken together, the preceding analysis allows me to tentatively suggest four convergent explanations for the existence of the IF function of motives:

Explanation 1: Phenomenological intensity is directional and acceleratory. It is a measure of the strength of the readiness to implement a specific action plan. If the plan-of-action is entirely aligned with current actions, then, from an action-readiness perspective, it appears empty. The intensity of the motive has no change in action.

65 As a schoolboy swimmer I remember crying underwater. If only I had felt motivational pull instead!
66 Another example of a phenomenologically silent “empty” motive would be an incredibly strong motive to execute an action plan with a schedule that starts tomorrow. There is nothing to do now, so nothing is felt.
to which it can attach. Talk of the intensity of a motive to change nothing is nonsensical. An uninhibited motive to maintain alignment will never have a phenomenological intensity.

Explanation 2: Phenomenological intensity is experienced at the agential level. It therefore reflects not the potential impact of a motivational force on any opposing motive but the potential impact on the agent. The motive to pursue an alternative possibility demands an acceleratory agential change, whereas the motive to hold the course does not. This is analogous to the earlier example of the rope. The agent only feels the pull of the rope when her speed is misaligned with that to which the rope’s mechanism is set. An aligned motive will never have an agentially experienced intensity.

Explanation 3. An agent’s decision-making architecture implements a ‘winner takes all’ model, which naturally involves inhibition of any unchosen motive. This cognitive architecture implies the existence of neurally implemented inhibitory “gates”. Phenomenological pull indicates that presence not of a motive in isolation but of an interaction between a motive and an inhibitory gate.

Explanation 4. The phenomenology of motivational pull would have no value to an extremely simple agent, responding inflexibly to environmental changes. However, with the addition of inhibition and the ability to flexibly optimise current behaviour, comes the cost of untaken opportunities. The experience of motivational pull alerts the agent to resisted affordances and ungathered conative value. This information, if stored, has value for a diachronic agent.
Further work is encouraged to confirm whether any, or all, of these explanations has merit. For now, I will assume that the remaining uncertainty does not undermine the truth of the following statement:

*The functional strength of a motive is revealed to an agent through phenomenological intensity if and only if it is inhibited.*

In the following section I explore the implications and adaptive benefits of the inhibition-based, conditional model of unification.

4. The inversion of conative phenomenology.

The analysis of section 3 supported the claim that a motive’s strength is revealed phenomenologically to an agent if and only if it is inhibited. This can be broken down into two independent elements which allow an agent to make inferences from her conscious experience. The first element allows an agent to infer the existence of an inhibited motive from her conative phenomenology. The second allows her to infer the (approximate) functional strength of an inhibited motive from its phenomenological intensity.

Although the first element is perhaps the weaker claim, it has significant implications for the phenomenology of agency. Taken seriously it implies a wholesale inversion of how philosophers understand the experience of agency. Conative phenomenology is not the phenomenology of implemented motives and intention alignment, but the experience of frustrated desires and of (as yet) unfulfilled plans. It is the phenomenology of inhibition-driven plan misalignment.
The affordance literature (McClelland, 2019) talks of aspects of the environment making *demands* on an individual. A ringing phone demands to be answered, an itch demands to be scratched. But conative phenomenology is not the phenomenology of agential acquiescence to demands. It is the phenomenology of demands resisted. Action-plans may have been prepared or readied or potentiated, but conative phenomenology highlights those plans that have not been implemented, at least not yet, or not in full.

The contrast with reflex-responses is stark. These “knee-jerk” responses are irresistible, but their very irresistibleness removes them from the conative landscape. The responses are functionally strong, and drive behaviour, but are never felt.

In his analysis of intentional agency, Bayne describes intentional actions as ‘actions that are carried out by agents themselves, and not by some sub-personal or homuncular component of the agent’ (2013, p162). For Bayne the differentiating factor is the level of integration within the broader cognitive economy. Whilst a sub-personal reflex-response is functionally isolated from the rest of an agent’s cognitive architecture, an action that is ascribed to the person herself must be motivated by an attitude that is cognitively integrated ‘in a flexible and appropriate manner’ (ibid, p164). Conative phenomenology, by inverting its focus, captures only mental states that are flexibly integrated within the cognitive economy.

Motives are inhabitable. No matter how strong an activated desire, it is always possible that the desired action-plan will be rejected if the agent is more motivated to do otherwise, or delayed if deliberation is incomplete. As Ricouer states, ‘The very ability to decide … is always to some extent a refusal, a discarding of rejected motives’ (1966, p14). If an activated motive is inhibited, even temporarily, then it is
felt and this experience confirms to the agent that she has flexible agency, that her response to the demand of a motive is not fixed. The control revealed by conative phenomenology is control over unchosen motives. The agent is free to do what she most wants to do only because demands can be inhibited and pre-dispositions towards (less) attractive alternatives can be inhibited. Agential control, as revealed by conative phenomenology is the ability to flexibly navigate an affordance landscape, motivated but not compelled by inhibitable motives.

In addition, the inversion of conative phenomenology is relevant for questions of seemings-based justification for libertarian beliefs.

The most obvious result is that the conative phenomenology of motivational pull, is consistent with the motivational strength idea. The phenomenology does not justify the belief that motivational contests are decided on anything other than functional strength, or that motives can be divided into different “hot” and “cold” kinds based on internal-to-the-attitude factors. All conative attitudes towards an unrealised plan-schedule are felt if they are inhibited.

However, the fact that it is subjectively difficult to isolate the phenomenological intensity of a motive from its content, does create a nuance. It implies that when an agent is torn between two alternatives of similar functional strength but different plan-contents, she may find it difficult to order them by intensity. Although she will always decide based on relative functional strength, this functional outcome, observable in the third-person, may not always be predictable from a phenomenological, first-person, perspective. She will not have full first-person access to her own psychological level decision-making process.
Motives are commensurable because motivational contests are based purely on functional strength, but they may seem to be incommensurable due to other aspects of conative phenomenology, such as action-readiness. The ‘window into the lovely machinery’ (Wegner, 2018, xvi) is translucent rather than transparent.

**Conclusion**

The phenomenological intensity of a motive is a representation of its functional strength if and only if it is inhibited. This resolves the puzzle of relative strength which can now be rewritten as three compatible statements:

1. When making a choice, an agent always choses to do what she is most motivated to do (The motivational strength idea)
2. The phenomenological intensity of a motive is a non-conceptual representation of its functional strength iff it is inhibited (Conditional Correspondence)
3. The phenomenological intensity of unchosen motives can be greater than those of chosen motives. (Experience)

This finding is important in the context of the thesis, because a) it supports the motivational strength idea and b) only once the correspondence relationship is known can the full power of phenomenological data be unleashed, to interrogate and enhance functional theories of the mind. In the next chapter I will analyse a second, complimentary form of conative phenomenology: the feeling of effort.
Chapter 4
Conative Phenomenology: Effort

In this chapter I develop a unified representational theory of the feeling of effort, starting with the idea that the feeling of effort is the feeling of trying-against-resistance. This raises the question of whether, for all three sub-species of effort (effort of will, cognitive effort and muscular effort) the resistance takes a common proximate form. After analysis revealing the link between cognitive cost-benefit computations and contextually activated, value-based, motives, I show that, for all sub-species of effort, a trying is resisted by the conative force of a conflicting motive. As conative states indicate (positive or negative) value-commitment, I conclude that the feeling of effort represents-as-costly a trying, where cost is understood in the technical conative sense.

This is important for the thesis for three reasons. First, alongside the analysis of chapter 3, it confirms that conative phenomenology is consistent with the motivational strength idea. Second, it indicates that the feeling of effort is a marker of intentional action. Third, it reveals that wherever there is effort there is motivational conflict. Motives for and against the effortful action must exist. This will be particularly relevant for my analysis, in chapter 9, of the act of making a difficult decision.

1. Introduction – Feelings of effort
Feelings of effort are central to the experience of agency. They are a marker of intentional action. Whilst effort is generally viewed as praiseworthy and adaptively beneficial, it is ‘typically experienced as aversive’ (Massin and Bermúdez, 2023). Despite this, the experience of effort has been described as the ‘experience of the self as a source of force’ (Bayne and Levy, 2006) and the experience of control (Pacherie 2008; Lukitsch 2020). It is a characteristic feature of intentional acts of self-control (Holton 2009, Sripada 2021), and, in the debate about active agency, it is the phenomenology of effort that is proposed to make agents feel that decisions between alternative possibilities are ‘up to me’ (Mele, 1995, p135).

Surprisingly, there is no general and satisfactory theory of the phenomenology of effort. To be general, the theory must explain, in a unified framework, all subspecies of effort. To be satisfactory it must answer the question posed by Bermúdez in his recent paper: ‘What is the feeling of effort about?’ (2023). I focus my analysis on Bermúdez’ paper and present my answer to his question.

Bermúdez starts by taking what he calls a ‘feeling-first approach’:

An action is an effort iff its execution is accompanied by a feeling of effort (p3)

This approach is a launchpad for a unified theory of effort, but it raises the question of what, if anything, the feeling represents? The feeling-first approach, on pain of circularity precludes the intuitive idea that the content of a feeling of effort is an effort. What is known as the feeling of effort is therefore a feeling of <X>, where X is yet to be defined. In this chapter I will attempt to define X. By doing so I am implicitly rejecting a non-representational view of effort.
When considering the different kinds of effort, I will follow Massin’s (2017) helpful categorisation between three subspecies: Muscular effort, effort of will and cognitive effort. Whilst effort of will is felt by an agent trying to act in accordance with a formed intention, cognitive effort relates to the cognitive process itself, and is felt, for example, when focusing attention or making a difficult choice. This raises the possibility that there are three separate X’s, one for each subspecies of effort, which would imply that ‘there is no lower-level unification of efforts in terms of mechanisms, resources or forces’ (Bermudez, 2023). If, however, there is a single X that applies to all the subspecies, that would allow unification at both the higher- and lower-level.

I propose a unified theory in which the feeling of effort, for all three subspecies, “represent-as-costly” a trying, where the term costly should be understood as having a technical conative meaning, consistent with the representational attitude framework. Whereas a positive conative attitude, such as a desire, represents-as-valuable a plan-of-action, implying a motivational pull towards implementation, the feeling of effort represents-as-costly, implying a motivational force against what the agent is trying to do. It indicates that by proceeding with the intended course of action, the agent is forgoing some alternative action which has conative value. The feeling of effort is therefore a non-conceptual representation of the conative costs of a pursued course of action. It indicates that the attempted action is controversial or unsustainable or costly.

In chapter 2 I argued that an intention is merely an uninhibited motive and that implementation could be delayed if (and only if) that was specified in the plan-schedule. Here I will assume that a trying is also an uninhibited motive, but one that
the agent is attempting to implement right now. A trying is the attempted implementation of the immediately pressing, occurrent element of an intention.

2. Trying against resistance: The assimilation problem

In his detailed and perceptive analysis of conative phenomenology, Kriegel argues that ‘the fundamental form of our conative experience is a proprietary phenomenon of deciding-and-then-trying’ (2015, p72). In his account, ‘the mental core of action is captured by the notion of trying’ (p89), and, as such, trying is the answer to Wittgenstein’s famous question: ‘What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?’ (1953, p621). Using this language, we might say that an intentional action is a successful trying. According to Kriegel, tryings are also highly relevant to discussions of effort: ‘There is an intimate connection between the phenomenology of trying and the phenomenology of effort’ (2015, p90).

By choosing to define an intention in terms of a trying rather than an action, Kriegel accommodates the fact that intentions can be frustrated. They can fail in the face of resistance. Even if the resistance is not great enough to cause a failure, it can make the execution of an intention hard. ‘Trying involves the experience of mobilizing force in the face of resistance.’ Trying ‘mobilises effort’ (ibid). Supporting the importance of resistance to any account of effort is the fact that an unresisted trying is not effortful. As Massin states: ‘The intensity of effort… corresponds to the degree to which the exerted force is counteracted by the resistive force”. (2017, p244). I will use Kriegel and Massin’s suggestions as a starting point for my analysis:
The feeling of effort is the feeling of trying against resistance.

An effortful action is therefore a resisted but ultimately successful trying, whereas an effortless action is an unresisted succeeding. A failed attempt to act is a trying-but-failing, which is effortful for as long as the trying against resistance is sustained. To be clear, a trying is always ‘categorical’ (Ricour, 1950, p70) in that it is committed-to and uninhibited. Feelings of effort don’t arise from resistance to ‘hypothetical’ (ibid), not yet decided upon, motives.

This framework suggests that the feeling of effort is not driven solely by a trying, nor solely by the existence of an obstacle to a valued end, but by a clash of the two. It is driven by a trying-against-resistance relation. Although helpful, this immediately raises the question of whether the ultimate source of resistance needs to be in a certain “proximal” form in order interact with a trying in a way that induces a feeling of effort.

To understand the difference between ultimate and proximate resistance, imagine that Dario is moving forward at high speed in a car, and the car hits a barrier. The barrier is the ultimate source of resistance to his forward motion, but it does not necessarily directly decelerate Dario. If he is not wearing a seatbelt, then the car will stop but he may not. The locking of the seatbelt is the mediating mechanism which transforms the ultimate resistance of the barrier into proximate resistance to Dario’s forward motion.

Similarly, whilst external resistance may result in a failure of implementation, it does not necessarily mean that the trying will stop. An agent can keep trying indefinitely.67

---

67 She can try as long as she is alive, even though that trying will become less and less effective.
A skinny weightlifter under a 150kg bench-press, may not be able to lift the weight, but she is always able to try. The fact that external resistance cannot directly influence trying follows from the broader fact that the external environment cannot directly influence non-sensory states of the mind. Any influence from the external world to conative states is mediated by sensory states, just as any influence from internal states to the world are mediated by tryings. Tryings and sensory states are sometimes called (Markov) blanket states as they separate the adaptive system from the external environment (Parr, Pezzulo and Friston, 2022).

Figure 2. Illustration of blanket states. Adapted from Parr, Pezzulo and Friston, 2022

The challenge, but also the opportunity, of a model based upon a trying-against-resistance relation, is to answer the following two questions: First, for each

68 Blanket states are necessary for flexible agency: They ensure ‘that an adaptive system’s internal states are autonomous from environmental dynamics and can therefore resist their influences’ (p44). At the other end of the spectrum is a reflex-response. Here the environment can cause the movement directly, and internal states cannot intervene.
subspecies of effort, is there something identifiable as a source of resistance, which the agent is ultimately trying to overcome? This is easy for muscular effort and efforts of will but harder for cognitive effort. Second, can each ultimate source of resistance directly interact with a trying in a way that induces a feeling of effort? This is easy for efforts of will, harder for cognitive effort and, given that external resistance cannot influence tryings directly, a potential worry for muscular effort. If the feeling of effort is only induced by a trying-against-proximal-resistance, then finding an ultimate source of resistance is not enough. I must also detail the mediating pathway that produces a proximal source of resistance. The wrong kind of resistance causes what I will call an assimilation worry: That there is a trying and a resistance but no obvious relation between the two. Even those who are less concerned by the assimilation worry, should see the benefits for unification of a proposal in which the characteristics of proximal resistance are elucidated, and the mediating pathways from ultimate to proximal resistance are specified. If this can be achieved then the phenomenology of effort, for all three subspecies of effort, can be explained in a single unified theory. This is what I will attempt in this chapter.

In sections 3-5 I analyse the three subspecies of effort in the context of ultimate and proximate resistance. I start with a brief analysis of efforts of will, for which the right kind of resistance is most easily identified. The harder analytical work will come in sections 4 and 5 where I unify cognitive and muscular effort under the same framework.

---

69 An alternative strategy might be to argue that agents monitor the tryings and external states separately and somehow combine them to model (rather than observe) the trying-against-resistance relation. In my opinion, the external world is compared to the plan-of-action to generate a feedback signal that more (or less) force is needed, but this comparison does not explain ongoing feelings of effort. I will not explore this possible explanation further as I believe it to be a) unlikely to explain the phenomenology, b) unnecessary given my alternative explanation, and c) inconsistent with a unified approach (as it would apply to muscular effort only).
For cognitive effort, the question is whether there is any resistance at all. Contemporary theories claim that the feeling of cognitive effort is the experienced measurement of the costs of cognition (Kurzban et al., 2013). At face value this cost of using cognitive resources doesn’t seem to be a kind of resistance. I analyse Bermúdez’ suggested solution, that cognitive costs output impulses against mental action, which need to be resisted. Although a neat solution, I argue that this proposal leads to problems that I call *cost duplication*. My alternative proposal is that the impulse to stop the action is a motive, and is therefore an *input* to the computation. Although this difference may seem small, what I am suggesting is a full inversion of philosophical understanding of the phenomenology of effort. Feelings of effort do not make an agent want to stop, but reveal to her that she (already) wants to stop. The motive to stop the action, which was defeated in the motivational contest, makes its (ongoing) existence known to the agent through the conative phenomenology of effort. On pain of duplication, this feeling cannot then synchronically trigger further motives against action.

For muscular effort I argue that although the ultimate source of resistance is external, this only causes feelings of effort indirectly, as an input into the motive to stop. I outline an adapted version of Bermúdez’ theory which accommodates this fact. The impact of external resistance can, via pain and fatigue, for example, indirectly activate an impulse to stop. Unlike cognitive effort this does not lead to cost-duplication as the external resistance is not a direct input to the cost-benefit computation. Physical trying is not stopped directly by physical resistance, but by the fact that the agent’s motivation balance has shifted to favour stopping.
If this is true then all three subspecies of effort involve a relation of conflict between two conative forces, which take the general form of the intention to proceed against the impulse to stop. This implies that the content of a feeling of effort has a common form for all three subspecies of effort, so unification is possible.

In section 6, I introduce Kriegel’s analysis of representational attitudes and their contents. Extending this framework, I argue that the feeling of effort represents-as-costly a trying. Finally, in section 7 I discuss whether feelings of effort as detailed can have any purpose within the cognitive economy.

In what follows I gather all activated conative forces under the umbrella term motive. This term should be understood to capture all motivational forces that move an agent to act, including tryings, urges, cravings, impulses, wants, action-desires, and any other sources of motivational force. There are two important facts about motives. First, they are dual-faceted, combining the conative force with the phenomenological experience of a pull to action. Second, they are always inhibitable by other conative forces. As such they are flexibly integrated within the cognitive economy. The inhibition relevant to this chapter is lateral, in that it occurs solely as a consequence of an agent’s positive conative attitude towards a competing alternative.70

Not captured by the term motive are non-motivational states, such as wishes and unactivated dispositional desires, or uninhibitable ‘knee-jerk’ reflex responses which are functionally isolated from the rest of an agent’s cognitive architecture.

70 See chapter 2 for analysis of lateral and deliberation-driven inhibition.
3. Efforts of Will

An agent feels an effort of will when she has formed an intention to perform a certain action but is subject to a motive, such as a craving or an impulse, to do otherwise. Trying to act in line with an intention is only effortful when it faces resistance from the force of a conflicting motive. For example, if an agent has committed to dieting, her trying to diet will be harder if she has an urge to eat. If she has no such urge, then dieting will be easy. In this scenario, the source of the resistance-to-be-overcome is an eating-focused urge, but the specifics can be abstracted away. All that is relevant is that there is a conflict between conative forces.

The general case of a successful effort of will is a trying-to A overcoming resistance from a motive against A-ing (or towards stopping trying-to A). In a contest between motives, both conative forces are psychological and are flexibly integrated within the same agent’s cognitive economy. As such, there is no assimilation worry. The trying and the motive to stop trying interact in a way that allows one to inhibit the other and, because conative forces are dual-faceted, that interaction is felt. If the feeling of effort is a trying against a certain kind of resistance, then an internal conative force seems to be the right kind of resistance.

I will proceed on the basis that the feeling of effort in an effort of will results from conative conflict. The feeling is of a trying-against-resistance where the resistance comes from the motive to stop trying. For efforts of will the ultimate and proximate forms of resistance are the same. They are the psychological impulse to stop trying, which is a motive. I now turn to cognitive effort.
4. Cognitive Effort

Kurzban et al. (2013) proposed a theory of cognitive effort which has been broadly endorsed. They argued that decisions about whether to persist in, or abandon, a cognitive task are the result of cost-benefit evaluations, and that the mental effort felt when engaged in cognitive activity is ‘the conscious, experienced measurement of the costs’ (p662) of ongoing deliberation. This raises two related questions: How can something as cold as a cost be experienced as a feeling? And how, if at all, is this feeling of effort linked to resistance?

Bermúdez suggests that both questions can be answered if one makes a simple adjustment to the theory. He argues that ‘cost-benefit computations output, not the feeling of effort, but an impulse against performing the action in question at the specific level of intensity – in short an impulse to stop the action’ (p20, my italics).

Just as in the case of the tempted dieter, there is a conflicting motive to be overcome and therefore ‘the feeling of effort represents the currently executed action as involving a resistance of an impulse to stop the action in question’ (p25).

If true, the source of the resistance to be overcome in cognitive effort would be a psychological impulse, just as it is in an effort of will.

Although an attractive solution because of its unifying properties, the initial question, of how a cost-benefit computation can output a feeling, has been transformed into one of how a cost-benefit computation can create an impulse. And, in addition to the question of how, one might add the question of why? The cost has already been incorporated into the cost-benefit evaluation of what to do, so an impulse against the action would be either epiphenomenal or duplicatory. Bermúdez’ elegant solution
has considerable merit, but the link between costs and impulses needs to be clarified.

A potential way forward is suggested by analysis of an implicit assumption of cognitive cost-benefit computations: They rely on the existence of a common unit on which the computation can be performed. I turn to that now.

4.1. A common computational unit

For the idea of a cognitive cost-benefit computation to be more than a metaphor, an agent’s cognitive infrastructure must include an action-focused commensurable unit on which the computation can be performed. Additionally, for a decision of what to do to be an agent’s own decision, the common unit of the cost-benefit computation must be a contextually modulated manifestation of the agent’s values. These values may be innate or learned, intrinsic or extrinsic, subconscious or explicitly endorsed.

Ends that are consistent with an agent’s values are ends to which she is implicitly or explicitly committed. To be committed implies being motivated, in appropriate circumstances, to act in ways that achieve states of affairs consistent with those ends. When the question of what to do arises, relevant values must be activated and transformed into this action-focused common-currency. Values in support of an action will be benefits, those against the action or in support of an incompatible action will be costs.

Over the last decade, empirical data in the appropriately named field of neuroeconomics has confirmed the idea of a common-currency: ‘Neuroeconomics is now anchored in a new theoretical framework that posits the existence of a
centralized neural mechanism that enables the a priori incomparable subjective values of multiple stimuli to be compared using a single scale’ (Serra, 2021, p47).

On a more abstract level, philosophers have historically talked in terms of belief-desire calculus and motivational contests. The motivational strength idea, defended throughout this thesis is that, when an agent acts intentionally, she will always do whatever she is most motivated to do. For the purposes of this chapter, I will assume that the below is true:

When choosing between A-ing and ¬(A-ing), the agent will choose to A if, at the moment the choice is made, the strength of motives towards A-ing is greater than the strength of motives towards ¬(A-ing).

This suggest that cognitive cost-benefit computations and the motivational strength can be unified through the following statement:

The commensurable unit on which cost-benefit computations are performed is the motivational force of contextually activated, value-based motives.

Perhaps the word evaluation more clearly conveys the sense of the decision-making process than computation. The decision of what to do is made based on values, and the strength of an agent’s conative attitude toward each plan is a subjective, contextually modulated, representation of the plans value. The plan she values the most is the plan she is most motivated to execute, so that is the plan she chooses. In situations where the pro and con attitudes are of similar strength, the agent may remain undecided.

As Kriegel states: ‘The value-commitment built into conative states is attitudinally encoded…. we could demarcate the conative in terms of representing-as-valuable’
In addition, activated motives are dual-faceted. They are the psychological phenomena at the core of decision making that integrate the physical and the phenomenological.

The claim that dual-faceted motives are the common unit of conative cost-benefit evaluations, shifts the seemingly cold language of computations back towards more traditional understandings of action as being driven by the conative force of motives, which are felt.

**4.2. A revised solution**

Bermúdez stated that cost-benefit computations ‘output… an impulse to stop the action’ (2022, p20). What the above analysis shows is that impulses are not separate and additional outputs that guide an agent towards stopping. Rather, impulses are the costs. The impulse to stop is the input into the cost-benefit computation on the negative side. Equally, the agent doesn’t need to engage a further resistance-overcoming force to offset the output of the cost-benefit computation. The resistance-overcoming force just is the motive to proceed, which is the input into the cost-benefit computation on the positive side. The feeling of cognitive effort is a non-conceptual representation of the (input) cost of ongoing deliberation.

Some readers may wonder if the particular costs outlined by Kurzban et al. can really be thought of as motives. I believe the answer is yes, albeit metacognitive motives. The cost-benefit computation discussed in their paper relates to the question of whether to stop or prolong deliberation. This is a metacognitive question so the
relevant motives are metacognitive. The motivational contest is between the motive to reduce uncertainty and the motive to conserve cognitive resources.

In cognitive cost-benefit computations the inputs are pro- and con- motives. This applies equally to the previously discussed efforts of will. The urge to eat is not only experienced as a motivational pull but, if strong enough, is able to influence behaviour. This is because it is an input into the cost-benefit computation determining whether to diet or to eat.

This unifies efforts of will and cognitive efforts as tryings-against-resistance, where the resistance is a conflicting conative force, the psychological impulse to do otherwise.

A trying is a chosen motive, whose maximum motivational force is given by the subjective, contextually modulated, value of the desired end. In the case of an endogenous struggle the feeling of effort arises due to the resistance provided by a conflicting motive which must be inhibited. In contrast, in the case of muscular effort the ultimate source of resistance to a trying is an external force or obstruction. Whether there is an intermediate, proximate source of resistance, I consider in the following section.

5. Muscular effort

When analysing cognitive effort, there was a worry about whether the kind of costs highlighted by Kurzban et al. could be a source of resistance to tryings. Bermúdez suggested a solution, that cost-benefit computations output an impulse which must be overcome. I argued that motives, such as impulses, are the inputs to cognitive
cost-benefit computations. This explains the presence of proximal resistance-to-be-overcome, whilst avoiding any risk of cost duplication.

With respect to muscular effort, there is a symmetrical worry. Although it is obvious that external forces can provide resistance to the output of tryings, it is less clear that such resistance can be incorporated into internal evaluations. There is a potential assimilation worry which needs to be resolved.

One possibility is that, beyond impacting motives indirectly via beliefs about attainability, external resistance doesn’t need to be incorporated into internal evaluations. The agent’s intensity of trying will rise as needed until it hits some maximum. If that amount of trying is sufficient to engage the muscular force required to overcome the resistance, then she will succeed. If not, she will fail. At first glance, this scenario might seem consistent with simple and short-lived muscular efforts, such as that required to unscrew a tight jar lid.

However, the situation gets more complicated when one considers temporally extended efforts such as endurance running. Here the athlete is (almost) always able to run one more step, but that doesn’t mean she is willing to do so. In this scenario there is a much clearer case for the existence of an internal cost-benefit evaluation which compares the desire to keep running with the desire to stop running. The latter is informed by levels of pain and/or fatigue as well as beliefs about external factors such as the steepness of the hill, but these factors influence action indirectly via a desire to stop.

The transmission channel could be described as leading from external resistance, via monitoring mechanisms such as affect and sensorimotor feedback, to an inhabitable conative force towards stopping. Under this model, fatigue or pain can
lead to stopping behaviour, but only indirectly via a stopping-impulse, in the same way that hunger and eating are mediated by the desire to eat. This mediation is necessary to allow agents to flexibly respond to the environment. Intentional actions are driven by inhibitable motives rather than inflexible reflex responses.

Revisiting the jar opening scenario, a closer inspection reveals a similar dynamic but over an accelerated timescale. The fact that the engaged force is not enough to undo the lid will not, by itself, cause the agent to stop trying. Theoretically, she could continue trying-but-failing indefinitely. Stopping trying can only follow a removal of the desire to undo the jar, or a predominant motive to stop. Arguments for the former are undermined by the fact that the agent will often try again, perhaps with the help of an external aid. If a motive to stop is at least part of the cause of the stopping, then the feeling of effort might indicate the presence of this proximate resistance.

A possible objection might be that there are situations when an action is effortful, even when there is no motivation to stop. For example, imagine that Malcolm is walking to meet some friends and suddenly the wind and rain pick up. It is hard going, and he has to physically lean into the wind to keep making progress. Is there any moment when he feels effort but there is no motive towards stopping? Set aside the fact that it is physically hard, and he is feeling fatigue, as effort is different from fatigue. Some readers might think that he does feel effort without wanting to stop, but would presumably agree that there is a point at which the agent would stop walking and seek shelter. At that point he must be motivated to stop, otherwise he would not do so. My claim is that the motivation towards stopping doesn’t suddenly jump from zero once some threshold is hit. Rather it increases in a continuous manner and is revealed to an agent through the feeling of effort. Of course, for as long as Malcolm
keeps going, he doesn’t most want to stop – stopping is not his predominant motive - but his growing feeling of effort does indicate a growing motive towards stopping. This is supported by the fact that the timing of when he stops walking and seeks shelter is dependent on the strength of his motivation to keep walking. If he is in no rush or if he is not bothered about these particular friends, then he will seek shelter earlier than if he is late for an important date.

This suggests that an adaptation of Bermúdez’ earlier proposal could apply to muscular effort:

External resistance triggers, not the feeling of effort, but (indirectly, via sensory states) an impulse against performing the action in question at the specific level of intensity – in short, an impulse to stop the action.

An example might serve to illustrate the point: When pushing at the absolute limit a cyclist may describe his legs as screaming at him to stop. Now, when anyone or anything screams at you to stop, it is adaptive that this triggers an impulse to stop. However, the impulse does not guarantee stopping behaviour. Fortunately for competitive cyclists, impulses are inhabitable. They can be defeated by a predominant desire to go on. Professional cyclist Jens Voight’s famous quote captures the dynamic: 'When my legs hurt, I say: “Shut up legs! Do what I tell you to do!”'

There are a number of attractive features of this account of feelings of muscular effort. First, and most importantly, it resolves the assimilation worry. Second, because motives (including impulses) are the common units of cost-benefit computations, it explains why an agent will sometimes stop trying even when she...
has the capacity to go on. Third, it enables unification of all three subspecies of effort: They all indicate conative conflict, which is felt.

Perhaps more controversially, this account drives a wedge between feelings of effort and the magnitude of the external resistance. The former can vary even if the latter is unchanged. Psychological differences, cognitive fatigue, perceptions of progress, and even acts of mental control can modulate feelings of effort independently of external forces. This phenomenon will be familiar to cyclists who maintain a fixed power output throughout a session but feel wide variations in perceived effort. Equally, there may be times when an athlete feels no effort, despite the fact that she is exerting power and her activity is objectively fatiguing her muscles.

Although the ultimate source of resistance is external, the intensity of the feeling of effort corresponds to the magnitude of the proximate (psychological) resistance – the motive to stop.

This is not to suggest that the link between external resistance and feelings of effort is completely broken. In some cases, depending on the sensory experience and the mediating trigger, it remains strong. Sometimes, for example when an agent feels extreme pain, the pathway from external resistance to feelings of effort may be fast, compelling, and relatively inflexible, whereas at other times, when an agent is fatigued after a long day, it may be more open to influence from other psychological factors. Although I recommend further interrogation of the triggering of psychological impulses, the nuances will not impact the central claim of this chapter, that, for all three subspecies of effort, the feeling of effort is a feeling of:

A trying-being-resisted by a conflicting motive.
By understanding that feelings of effort, of all kinds, can be expressed in terms of the conative force of motives, and combining that with the fact that dual faceted motives integrate the physical and the phenomenological, it is possible to unify all feelings of effort in a single account.

Progress has been made, but I have not yet elucidated the precise content of the feeling of effort. In the following section I will do so, using a framework that differentiates between a representational attitude and its content.

6. Representational guise

In a recent paper Kriegel (2023) argued that phenomenal properties vary with attitude as well as with content. He suggests that when an agent fears a snake, for example, fear is the mode of representation rather than an element of the content that is represented. ‘Plausibly then the danger commitment is ‘attitudinally encoded’: [the agent’s] experience does not represent the snake as dangerous, but represents-as-dangerous the snake’ (p175). Content can therefore be represented under different guises depending on the agent’s internal state. A dog can be represented-as-dangerous or represented-as-lovable.

This framework can be applied to emotions, which have historically proven problematic for representationalists due to a lack of obvious content. Kriegel’s account marries a specific attitudinal guise to a generalized content. ‘We might hold, for instance, that depression represents-as-dull the world, anxiety represents-as-threatening the world, and so on’ (p175).
Finally, and less controversially a desire represents-as-good, or represents-as-valuable, its content. This is consistent with the guise-of-the-good thesis (Tenenbaum, 2007) and explains why the phenomenology of conative desires differs from that of cognitive beliefs, which represent-as-true their contents. Although I will not further defend Kriegel’s thesis here, I believe it is plausible, and could be extended to the feeling of effort in a way that is consistent with the findings of this chapter.

The first step in this process is to argue that the content of a feeling can itself be a trying. Although human agents do have the capacity for full-blown analytical meta-representation, that is not necessary here as it is broadly accepted that humans and other animals can integrate signals about cognitive states even if they don’t conceptualise them as such (see, for example, Dokic, 2012, and Proust, 2013, on procedural metacognition, and Carruthers and Williams, 2022, on model-free metacognition). Of course, some animals don’t have or don’t feel such signals. The point is that for any animal that does feel effort, the content of the feeling can be a psychological state even if it cannot meta-represent. There seems little reason to believe that the content of the feeling of effort cannot be a mental state such as a trying.

The second step is to find a representational guise that captures not only the direction of the phenomenology but also the relation of conative conflict. After all, an unresisted trying has no phenomenological signature. An initial suggestion, therefore, might be that the feeling of effort represents-as-resisted a trying. However, what has been uncovered in this chapter is the specific conative nature of the
resistance. When combined with the fact that conative forces are value-based and conative phenomenology represents-as-valuable its content, this suggests that:

The feeling of effort represents-as-costly a trying.

As costs are negative, this correctly captures the fact that feelings of effort are notoriously aversive\textsuperscript{71}, even though effortful action is often described as praiseworthy or agentially affirming. The trying is the content, so the fact that the trying is towards a valued end is irrelevant to the phenomenology of effort. This is consistent with Kurzban et al.'s description of the feeling of mental effort as an experience of the costs of ongoing deliberation. The greater the costs, the more intensely negative the feeling. If, on the other hand, there are no costs (i.e. no opposing conative forces) then the trying will not be represented-as-costly. Although benefits exist, they are not felt. From a phenomenological perspective a costless trying is silent.

Although the feeling of effort is a consequence of conative conflict, as is the feeling of motivational pull, the directionality is flipped. The phenomenology of motivational pull represents-as-valuable, whereas the phenomenology of effort represents-as-costly\textsuperscript{72}. To the agent an unchosen action-plan seems to emit a sirens call, whilst a resisted trying is accompanied by a phenomenological warning that it is controversial. In addition, there is a subtle difference between the two contents. Whilst the content of the feeling of motivational pull is a (hypothetical) plan-

\textsuperscript{71} Effort itself can be valued, but that is a higher-order feeling about effort not a feeling of effort (which is about a trying). The feeling of flow is arguable a positive feeling about (the progress of) a trying. I analyse flow in chapter 8.

\textsuperscript{72} Note that the presence of motivational pull, does not always imply the presence of effort. In the case of deliberation-driven inhibition, competing motives may both exert a motivational pull - such that the agent feels torn – but there is no (practical) feeling of effort. There may of course be cognitive effort. This is consistent with the theory because the agent is trying to deliberate, not trying to “inhibit both alternatives”.

189
representation, the content of the feeling of effort is a (categorical) plan-commitment. It is feeling about the occurrent element of a settled-upon, plan-of-action: a trying.

Having said that, inhibition is central to both feelings. Feelings of motivational pull only refer to inhibited motives and feelings of effort only refer to inhibiting motives – motives that are laterally inhibiting their opponents. The intensity of both feelings reflects the motivational strength of the inhibited motive.\(^{73}\)

7. The purpose of conative feelings

This proposal outlined in this chapter is controversial because it implies that the feeling of effort is a window through which the agent can view an input into the decision of what to do. As such, on pain of duplication, feelings of effort cannot further bias the agent against the chosen action. But does this imply that the feeling of effort is little more than a steam whistle on the mind-engine (Huxley, 1874)? If so, that might seem to be an argument to reject, or at least be wary of the account proposed. I offer a few brief thoughts:

First, windows and whistles serve a purpose. The feeling of effort is a person level phenomenon, which broadcasts information about an element of the cognitive machinery to the broader workspace (Barr, 2005, Dehaene & Changeux, 2011). Second, the intensity of felt effort is dynamic. Through awareness of the level of

\(^{73}\) This is consistent with motivational intensity theory (Brehm and Self, 1989, Gendolla, 2004) which states that needs, incentives and instrumentality appraisals do not directly determine effort. Instead, these variables ‘exert an indirect influence by setting an upper limit of motivation, or the peak of what individuals would be willing to do in the service of motive satisfaction. What should directly determine effort is what individuals believe that they must do to satisfy a motive’ (Wright, 1996, p428).
effort, human agents can monitor changes. Any fast or surprising change may trigger a review of the situation, even if the effort is not at a level that would normally equate to a decision to stop. Third, when an action is costly, awareness of this fact may activate a search for alternative means to the same end. Finally, even if the feeling of effort cannot impact the current decision, it can be valuable to a diachronic agent. She can remember historic efforts or imagine future efforts and incorporate any lessons learned into current decisions or plans. For example, if an agent has had experience of temporarily effortful actions leading to positive outcomes, then she may be more willing to keep trying in the future despite short term pain. Alternatively, an agent who has quit in the past may prepare herself for the next event by collecting reasons to persist, which can be rehearsed if necessary. At the extreme, agents might even begin to seek out the feeling of effort, or praise it in others. Because the feeling of effort provides a fine-grained insight into evaluations that would not be available to those only able to observe behavioural responses, it facilitates a more precise calibration of future decisions.

The feeling of effort is aversive, so it is generally perceived to be action guiding in that it ‘biases agents toward effort avoidance’ (Massin and Bermudez, 2023). However, the findings of this chapter imply a restricted understanding of this claim. When a feeling of effort arises, it implies that the current (chosen) course of action involves costs. Awareness of this fact does not imply that the agent should give up on their goals (after all, the benefits outweighed the costs) but it may activate a search for alternative means to the same end. If an agent can achieve the same beneficial goal with less conative cost, then that increases her overall (predicted) sustainability.
Finally, if something is effortful it means that the alternative unchosen action-plan is still being maintained as a live, or activated, alternative. Whilst this causes the feeling of effort, it also facilitates ease of switching. Plan-representations are partial so, as levels of effort rise and switching becomes more likely, the details need to be filled out. Whereas, from the perspective of opposition the main function of a conflicting motive seems to be to act as resistance against the current trying, as a change of mind becomes more likely it is adaptive for the agent to be fully prepared to efficiently implement the new decision.\textsuperscript{74}

Taken together, feelings of effort support the flexible, adaptive behaviour of a diachronic agent even if, as non-conceptual representations of input costs, they cannot directly and synchronically further undermine a chosen course of action.

**Conclusion**

This analysis of effort suggests two insights which will be carried forward within the thesis:

First, care must be taken when using phrases such as “effort is costly” or “agents are biased towards effort avoidance”. Taken at face value, these statements risk getting the directionality wrong. A feeling of effort reveals to the agent that a chosen action is controversial and involves costs, but it does not add a further cost that must be synchronically incorporated and resisted.

More parochially, the feeling of effort is highly valuable for philosophers in that it indicates the presence a motive that is being resisted. As a general point, this supports the motivational strength idea and confirms that feelings of effort are

\textsuperscript{74} There is a nice analogy here with opposition parties of government.
sufficient (albeit not necessary) for intentional action attributions. More specifically, and perhaps more interestingly, it challenges philosophers of mind to identify two motives for each feeling of effort. If an action is effortful then motives for and against that action must exist, and if theories of mind cannot explicate what those motive are, then they are incomplete. The most obvious case of this incompleteness is found in the making of a difficult choice. Answering the question of how a choice can be an act has been described by Vierkant as the agency problem of compatibilism. The understanding of the feeling of effort which has been developed here will equip us to finally solve that problem in chapter 9.
Chapter 5
Intentional Synchronic Self-Control
Overcoming the Mismatch Problem

Introduction

In this chapter I continue my defence of the motivational strength idea by analysing the challenge from intentional acts of synchronic self-control. These raise a question, which Altehenger (2021) calls the mismatch problem:

How is it possible for a single, instrumentally rational agent, whose motivational balance is in favour of a wayward action, to simultaneously be motivated to engage in a control action? How can the three statements below all be true?

1. When making a choice, an agent always choses to do what she is most motivated to do (The motivational strength idea)
2. An agent can intentionally shift her current motivational balance, changing what she most wants to do (Intentional synchronic self-control)
3. If an agent desires an end and is considering an act that will harm that end, then the motivational force of her desire will be transmitted in full into desire to refrain from the harmful act. (Instrumental rationality)

Sripada and Altehenger have argued that the existence of intentional synchronic self-control implies that the mind is divided and/or that an agent can overrule her strongest motives by using a separate and distinct source of “willpower”.

194
I deny that conclusion, and argue that intentional synchronic self-control follows naturally from the existence of metacognitive motives, acting in accordance with the motivational strength idea. Indeed, a nuanced understanding of instrumental rationality as it relates to metacognitive motives, implies that there is no puzzle at all. Premises 1-3 can be simultaneously true.

Although the analysis of this chapter is functional rather than phenomenological, support for premise 2 is based on phenomenology. Intentional synchronic self-control is the paradigmatic effort of will, which is felt. The functional analysis of this chapter, when combined with the phenomenological analysis of chapter 4, explains not only how intentional synchronic self-control is possible, but also why it feels the way it does.

1. The Puzzle of Self-control

On the one hand it seems to be a truism that, when an agent acts intentionally, she will always do whatever she is most motivated to do. I have called this the motivational strength idea. On the other hand, it does seem that if an agent feels or believes that her strongest desire is inappropriate then she is able to do, or at least try to do, something about it. She can resist temptation by engaging in synchronous self-control. This is experienced most clearly when effortfully inhibiting a wayward desire.

Whether, and how, these statements can be simultaneously true has been called the puzzle of self-control (Alston, 1977, Mele, 1987). Proposed solutions (Kennett and
Smith, 1996, Mele, 1997, Sripada, 2014, Altehenger, 2021) take the following general form:

1. Wayward desire W is motivationally stronger than conflicting virtuous desire V (at time $t_1$)
2. Something happens that shifts the motivational balance
3. The motive towards V-ing is now greater than the motive towards W-ing
4. Wayward desire W is defeated (at time $t_2$).

In solutions of this form, there is no conflict between stage 4 and the motivational strength idea. At the time of the final decision, $t_2$, the agent most wants to V. The puzzle arises in stage 2. What is it that happens and how can that something, which either enhances V or inhibits W, happen at a time when desire W is stronger than desire V?

To understand the puzzle, we must first set aside situations in which the something that happens is random or is driven by the action of someone else. This is neither self-control, nor a puzzle. Perhaps more interestingly, we can also set aside situations in which the agent herself has taken an earlier action which provokes the self-control when required. This form of self-control is not synchronic, but diachronic. The earlier action can be assumed to have taken place at a time when the desire to V was stronger than the desire to W. Having set aside those two scenarios, the something that happens at stage 2 is *synchronic self-control*. As Sripada states, ‘synchronic self-control is directed at attenuating the [relative] strength of a wayward desire that is active at the very time that self-control is deployed” (2014, p2).

---

75 Sripada would argue that, in his account, W is still the strongest *desire*, but as Barnes (2019) has shown Sripada’s arguments are consistent with claim that W no longer has the strongest motivational base.
Kennett and Smith (1996, 1997) proposed that synchronic self-control is triggered by a non-random but non-actional response to the situation. Just as an agent is about to act on a wayward desire, non-conative but modifying thoughts (or attitudes or feelings) can arise. As a result of this non-actional response, the motivational balance can shift. For example, imagine that an agent believes she has been slighted by an enemy and desires retribution. Her strongest desire is to harm her enemy, but every time she gets the opportunity to do so an image arises in her mind. She sees herself in prison and her children alone. Although the mental image does not have any intrinsic motivational power, it triggers a shift in the motivational balance of her desires, such that she holds off acting (for now).

Non-actional responses such as thoughts or feelings are not governed by the motivational strength idea, so can arise despite the strength of a wayward desire. And although the response indirectly undermines the wayward desire there is no direct conflict between the response and the desire. I will assume that at least some synchronic self-control is triggered in this way. However, intuition, backed by conative phenomenology, suggests that synchronic self-control is not always non-intentional and non-actional. Often self-control feels like an effortful agential act.

Although Kennett and Smith’s solution is ingenious, what I will call the puzzle of intentional synchronic self-control remains. The rest of the chapter is concerned with this narrow version of the original puzzle.

In section 2 I analyse the core of the puzzle, drawing heavily on the work of Altehenger (2021). This is partly because her elucidation of the divided-mind solution is incredibly clear and easy to work with, and partly because she identified a ‘mismatch’ that needs to be explained: How can a single unified agent desire more to
engage in self-control than to refrain from doing so, when her wayward desire, which is by assumption her strongest desire, should produce an instrumental desire to refrain? I argue that, even under the assumption of instrumental rationality, it is at least theoretically plausible that Altehenger’s mismatch can be explained.

In section 3, I propose a hierarchical solution to the mismatch problem, based on metacognitive desires.

In sections 4, I interrogate the divided-mind solution, as advocated by Sripada and Altehenger, and in section 5 I suggest that the most promising version of the divided-mind solution is hierarchical.

In sections 6 and 7 I compare the unified- and divided-mind hierarchical models and highlight two major areas of divergence. From the analysis of these sections, I conclude that the unified model has greater theoretical and empirical support.

Throughout the chapter I will sometimes use the word “desire” in accordance with the self-control literature, but I take a desire to be equivalent to what I have called a motive: a conative attitude that is both necessary and sufficient for motivation. I will proceed on the basis that to desire-to, to want-to and to conatively value have the same meaning, and that, when activated, they definitionally imply a conative motivational force towards achieving the valued end.

2. Intentional Synchronic Self-control and Instrumental Rationality

Intentional action is motivated action, and, in the case of pro- and con motivational forces, the motivational strength idea requires that the action is chosen through a
motivational contest. The contest from which intentional synchronic self-control results is a contest between the desire to perform a control action, which shifts the motivational balance, and the desire to refrain from such an action. For synchronic self-control to proceed, the motivational force of the former must be greater than the latter. Stage 2 from the general solution identified in section 1 is therefore split into four parts as detailed below.

1. Wayward desire W is motivationally stronger than conflicting virtuous desire V (at time $t_1$)
2. a) The question of whether to exercise self-control is (non-actionally) prompted  
b) The agent evaluates whether to exert control to decrease the relative motivational strength of W  
c) The desire to exert control is greater than the desire to refrain from control.  
d) The agent engages in synchronic self-control, shifting the motivational balance from W to V  
3. The motive towards V-ing is now greater than the motive towards W-ing  
4. Wayward desire W is defeated (at time $t_2$).

The addition of stage 2c) creates what Altehenger (2021) calls a mismatch problem. Choosing to engage in synchronic self-control implies a problematic mismatch between the relative motivational strength of desires W and V, and the relative motivational strength of the desires to exert or refrain from self-control. Her claim is that under a unified- rather than divided-mind view, the combination of 1 and 2c) should be impossible. This mismatch is at the core of the puzzle of intentional synchronic self-control.
Altehenger’s case hinges on an assumption that a unified mind is subject to, and constrained by, instrumental or means-end rationality. This states that if an agent desires an end, then the motivational force of her desire for that end should be transmitted towards means to that end. If an agent wants a cup of tea, for example, she must therefore be motivated to fill and switch-on the kettle. Being motivated towards an end without any motivation towards means to that end would be conatively incoherent. Equally, under means-end rationality, if an agent desires an end and she is considering an act that will harm that end, then the motivational base of her desire will be transferred in full to the motivational base of the desire to refrain from the harmful act.

Suppose that an inventor with a strong environmental conscience is considering developing plastic bags. He thinks of the fact that the bag will not decay and imagines the waste that can be reduced, if individuals use the same bag for years. There will be no need to constantly cut down trees for paper bags that tear and get soggy and need to be replaced76. His motivation is strong until a strange character claiming to be a time-traveller explains that most bags will be thrown away after only a single use. Because they never decay, thrown away bags will pile up in mounds of waste and clog up the oceans. They will become an environmental disaster. The inventor is instrumentally rational, so his motivation will immediately shift. He values the environment and will therefore resist anything that causes harm to the environment, including his own idea.

76 This was apparently the belief of the inventor of plastic bags, Sten Gustaf Thulin.
In the above story, the inventor was instrumentally rational for the entire time, but initially had a false belief about the consequences of his proposed actions. This failure of orthonomy (Kennett and Smith, 1996) was corrected by the time traveller.

In cases where intentional self-control is being considered, however, we can rule out failures of orthonomy because the agent is fully aware, and highly focused on, the consequences of targeting a desire. If instrumentally rational she will therefore be motivated to refrain from self-control. Although it is possible that agents may not be fully instrumentally rational, for the purposes of this chapter I will accept Altehenger’s challenge by assuming that they are, and that the motivational base of the desire to $W$ is fully transferred to the desire to refrain from exercising self-control.

The mismatch problem can be expressed in terms of a question: How can the three statements below all be true?

1. When making a choice, an agent always choses to do what she is most motivated to do (The motivational strength idea)
2. An agent can intentionally shift her current motivational balance, changing what she most wants to do (Intentional synchronic self-control)
3. If an agent desires an end and is considering an act that will harm that end, then the motivational force of her desire will be transmitted in full into desire to refrain from the harmful act. (Instrumental rationality)

I propose a simple hierarchical solution. Statements 1-3 can all be true because the question of whether to engage in self-control is metacognitive. When the question of

---

77Mele argues that self-control is made possible because motivational strength is not smoothly bestowed across the means-end relation (1997, p122). Altehenger counters that without ‘an account of the psychological mechanisms which cause the supposed mismatch... it will appear unacceptably ad hoc’ (2021, p248). I will argue that Mele’s ancillary action solution works even in cases of full instrumental rationality.
self-control is raised, metacognitive motives become relevant. These conative attitudes towards or against practical desires can be specific, such as the desire to suppress the desire to eat, or general, such as the desire to always remain in control. If strong enough, they can initiate a control action which will attempt to shift the balance of practical desires even on the assumption of full instrumental rationality.

Mele (1987, 1997) argues that there is an important distinction to be made between actions that directly conflict with the wayward desire, and ancillary actions. An ancillary action is one that can be taken at the same time as the primary action without conflict. For example, an agent’s strongest desire may be to read the paper, but that doesn’t preclude her from also acting on her desire to have a cup of tea. Tea-drinking and paper-reading do not conflict and can therefore be undertaken simultaneously. The only relevant inputs to the question of whether to drink a cup of tea are the desire for tea vs the costs of tea making.

Mele believes that certain ancillary actions can be used to shift the motivational balance relating to a primary action. The example he gives is of Ian:

Ian has just finished eating and he is thinking that he ought (all things considered) to get back to work now. However, he is enjoying the golf tournament on TV and he remains seated. He tells himself that he will watch the match until the next commercial break; but the commercial comes and goes and Ian is still in front of the set. Thinking that he had better drag himself away from the television now, Ian utters a self-command: “Get off your butt, Ian, and paint that shed!” Ian turns off the set, picks up his painter’s cap, and walks into his backyard. (1987, p. 69)
Ian’s self-command is clearly theoretically possible, as there is no direct conflict between speaking out loud and watching TV. However, when Ian utters the command, he does it specifically with the aim of stopping watching TV, so the command is a means to an end that conflicts with his strongest desire. Although the motivational base of Ian’s desire to paint the shed (to please his wife) is transferred to the desire to issue a command, we know that this is smaller than the motivational base of the desire to watch TV, which is transmitted in full to the desire not to issue a self-command. Instrumental rationality seems to undermine the distinction between indirect and direct conflict. Does this mean that Ian cannot perform an act of self-control? I think the answer is not necessarily for two reasons that relate to the fact that the acts being considered in the two separate motivational contests are not identical:

First, one way that a successful act of self-control can be different to V-ing is if it is easier than V-ing. Although motivational contests are generally thought to be between two alternatives, each alternative will itself have costs. The motivations on each side of the scales are net motivations, after taking in to account the costs. And, of course, instrumental rationality does not imply that the direct costs of performing one action are transferred to an alternative with the same end. If an action has high costs and there is a chance of achieving the same end with low costs then the net motivation towards the latter is greater than the former. In Mele’s language, the fact that two desires have the same ‘positive motivational base’ (1987, p69) doesn’t mean that they have the same total motivational base.

For example, imagine Vicky is walking in a Scottish valley and trying to decide between eating a sandwich from her pack and going to a great pub that is over the
hill. She really wants to go to the pub but decides it is not worth the physical effort. However, just as she is about to bite into her sandwich, a hiker explains that there is a safe and well-lit tunnel that would take her straight to the pub. She puts on her pack and set off through the tunnel. The gross motive towards the pub is unchanged but the net motivation increases as the hill-climbing cost is removed.

A central aim of many of the tips and tricks of self-control, such as issuing self-commands or engaging in reappraisal, is to make them as achievable as possible. Any reduction in costs will shift the motivational balance towards self-control. This shift may be quite small, in which case it will only make a difference in situations where the initial balance between $W$ and $V$ is close to 50:50, but it is an explanation for a mismatch between the relative motivational strength of $W$ and $V$ and the relative motivational strength of engage self-control vs refrain.

Mele’s vignette was inspired by a quote from Skinner that highlights this point: ‘In getting out of a bed on a cold morning, the simple repetition of the command “Get up” may, surprisingly, lead to action. The verbal response is easier than getting up and easily takes precedence over it.’ (1953, p236). Saying “get up” does not involve the costs of engaging weary muscles or getting cold, whereas climbing out of bed does.

To summarize, if there are a number of different actions that serve the end of self-control, then it is uncontroversial to state that costly actions may fail whilst easy actions may succeed. Switching to the easiest action with the least costs will maximise the chance of initiation and hopefully success.

Second, because the act of self-control is different to V-ing, there may be elements of the motivational contest between engage-in or refrain-from control, that are not
relevant for the competition between W and V. If these are powerful enough, or the motivational ratio of W to V is close to 50:50, inhibition of W is possible. And there is an obvious category of motivations that fulfil the requirement of only applying to the control question. These are desires about desires, enabled by metacognitive capabilities. I will analyse these metacognitive motives in the following section.\(^7\)

3. The Hierarchical Mind

How does the question of whether to maintain or inhibit a desire to W differ from the question of whether to W or not? The obvious answer is that the former question is metacognitive, where metacognition is the ability to monitor and control one’s cognitive processes. When a question of whether to act on a desire is raised, there are directly relevant pros and cons that do not apply to practical actions. These capture any desires about desires that an agent may have, and any costs of considering or taking metacognitive actions.

In contemporary discussions of metacognitive actions there has been extensive discussion of the cost of exerting control. In its simplest form metacognitive control requires the allocation of cognitive resources and this has a cost, either in terms of a depletion of future abilities in the case of cognitive fatigue (Muraven and Baumeister, 2000), or in terms of the opportunity costs of shifting resources from elsewhere.

\(^7\) Note that although I view this chapter as a defence of Mele’s position vs Altehenger and Sripada, and I will focus on their accounts from now on, Mele did briefly consider and reject the higher-order motivation approach in his 1987 paper (section 5). He argued that Ian was motivated to issue a command, whereas in my account Ian is motivated to engage in a control-action (defined in section 4.1, below), which is metacognitive. We have a difference in attitude contents. My personal view is that both accounts work together, with higher-order motives only sometimes being necessary. In Ian’s case perhaps they were not. After all, it is Mele’s example! (p74)
(Kurzban, 2016). This motivates conservation of cognitive resources to the extent that use of those resources is effortful. I argued in chapter 4 that the feeling of effort is the feeling of resisting or inhibiting an ongoing desire. In the case of cognitive effort, the desire that is being inhibited is the desire to conserve cognitive resources.

Less contemporary perhaps, but actively discussed in the philosophical literature has been the existence and impact of metacognitive or second-order desires. Frankfurt (1988) made the difference between first- and second-order desires central to his argument for free will. ‘The agent's role, according to Frankfurt, is to reflect on the motives competing for governance of his behaviour, and to determine the outcome of the competition, by taking sides with some of his motives rather than others’ (Velleman, 1992, p476). Whatever one’s views with regards the agent’s role, Frankfurt’s premise that human agents can have desires about desires is uncontroversial.

Frankfurt described creatures that acts only according to their first-order desires as wantons. Without metacognitive understanding it would be impossible to be anything else. However, human agents are fortunate enough to have metacognitive understanding and this opens the possibility of meta- or second-order attitudes. Humans can and do have beliefs and desires about their current first-order attitudes.

Akrasia is traditionally defined as acting against one’s best judgement. A metacognitive agent can form the belief that she is behaving akratically because her proximal desires are inconsistent with her all-things-considered best interest. It is an obvious step from believing that desires are inappropriate to wanting to change them. What this implies is that if an agent is aware that a practical desire is inappropriate, and believes that self-control is possible, then she can have a motive
to exert self-control, that is separate and distinct from her practical desires. This has a motivational power that only applies to very specific acts, acts of metacognitive control over practical desires (See Gordon, 1986 for a similar argument).

The desire to exert self-control, may be directed at a specific wayward desire, for example if an agent is trying to control her eating ahead of a summer holiday, or it could be more general if, for example, she tries to exert control in any situation where she becomes aware of motivational conflict. Indeed, the desire to exert control may become excessive as in cases of pathological control (of others) or anorexia. As Kotabe and Hofmann describe, ‘one of the axioms in psychology is that people aspire to be effective and in control. Versions of this axiom have been referred to as striving for self-efficacy (Bandura, 1977), valuing control (Higgins, 2012), the need for competence (Deci & Ryan, 1985), and effectance motivation (White, 1959). The underlying principle is that people are chronically motivated to feel effective (competent, in control, etc.)’ (2015, p625). This basic motive (Fiske, 2003) may be independent of, and pursued independently to, any more specific goals in conflict with the wayward desire. For example, a dieter may aspire to control food desires both because she wants to lose weight, and also because she values control.

Bringing this all together, and setting aside any costs associated with performing the first order acts, the motivational contest, described as stage 2c) above, lines up as follows:

On the side of refrain from control is the (instrumentally transmitted) motive towards the wayward desire W as well as the motivation to conserve cognitive resources. On the side of engage control is the (instrumentally transmitted) motivation towards the virtuous desire V as well as the motivational power of the metacognitive desire to
control the lower order desire W. Therefore, in order for an agent to try to perform an act of synchronic self-control, her desire to exert control must be great enough to overcome the combination of the net motivational balance of the lower order contest and the costs of control.

In situations where the first-order contest is finely balanced, and the costs of control are relatively low, then as long as her metacognitive desire is strong enough, she will be able to exert control, even if the motive towards W-ing is greater than the motive towards V-ing.

Metacognitive desires have a role to play in deciding metacognitive questions. This should not be surprising. When deciding whether to perform an act A, the most obviously relevant inputs are desires about A-ing, and the costs of action A. This doesn’t change in the specific case when A is a desire-controlling action. The primary inputs to questions regarding whether to act on desires are desires about desires, and the costs of any action on desires. Whilst Altehenger is right to remind us of instrumental rationality and its particular application to acts of synchronic self-control, that does not mean that the primary inputs should be forgotten.

Some readers might wonder whether the existence of the metacognitive motive shifts the practical motivational balance against the wayward action, even before a control action has taken place. If it did, this would imply that the motivational shift was not achieved through intentional action. There are two pathways by which this could occur:

First, the metacognitive desire, or contemplation of the control action, might trigger a shift in the agent’s motivational balance, without any further action. I agree that this is possible, but it is certainly not guaranteed. For example, imagine that an agent is
in a bath. She wants to stay in but is contemplating pulling the plug. Pulling the plug is a control action, which will shift her motivational balance as the bath empties. But desiring or contemplating plug-pulling will not have the same impact. She must perform the control action to shift her motivational balance. Other control actions such as reappraisal or distancing are the same. They only influence the agents motivational balance after they have been implemented.

Second, instrumental rationality might imply that the force of the metacognitive motive is transmitted to the motivational contest which decides the practical question. In that case, the metacognitive question and the practical question would be decided by motivational contests which incorporated the exact same motives. But usually\(^\text{79}^{80}\), there is no means-end transmission of the metacognitive desire to the practical motivational contest, because performing the wayward action neither helps nor harms an agent’s ability to execute the control action. Staying in the bath does not harm the agent’s ability to pull the plug, so instrumental rationality does not entail that the metacognitive motive weighs against the staying in the bath. Whether she stays in the bath is solely dependent on her practical motives.

In summary, the hierarchical model implies that for metacognitive agents there are two motivational contests: one for the control decision and another for the practical decision. The metacognitive motive is only relevant for the former, so the two contests have different inputs. If the inputs are different then the outputs can diverge

\(^{79}\) In the unusual case where executing the wayward action does rule out any future control, then the motivational force of the metacognitive desire is transmitted towards refraining from the wayward action. In that case the motivational inputs to the practical decision and the control decision are the same, so it is impossible for an agent’s total motivational balance to favour (simultaneously) the wayward action and the control action.

\(^{80}\) Another unusual case is what Sripada calls ‘full-blooded willpower’ (2014, p48). In this case the control action provoked by the metacognitive desire is not ancillary but simply ¬W. I return to why this is particularly effortful in section 7.
– there can be a mismatch – even under the assumption of instrumental rationality. If the motivational force of the metacognitive motive is strong enough, an agent’s motivational balance can favour (simultaneously) a wayward action and a control action. She can simultaneously stay in the bath and pull the plug.

This solution explains intentional synchronic self-control via metacognition, an accepted and widely studied phenomenon that enables acts of control directed at mental attitudes, and preserves means-end rationality. The mismatch problem is resolved, and metacognitive contests are subject to the motivational strength idea, just like practical contests. This is both efficient and evolutionarily plausible.

Some readers will be convinced by the arguments thus far, but others may feel that there is a better alternative in the divided-mind solution proposed by Sripada and Altehenger. Still others may wonder whether the hierarchical model proposed is just a version of a divided-mind solution, and therefore adds little. To ensure everyone is covered, I will analyse the divided-mind solution in section 4 and then contrast it with my hierarchical proposal in section 5.

I am tempted to say that the divided-mind proposal is an unnecessarily complex solution to a problem that has now been diffused.

4. The Divided-mind Solution

The model I have proposed is hierarchical in nature and includes two motivational contests. This raises the question of whether hierarchical is another way of saying divided? To answer this question, it is necessary to outline the proposal set out by
Sripada (2014) and Altehenger (2021a, 2021b) and determine the extent of the divergence between the two accounts.

4.1 Introducing the divided-mind

The key assumption of the divided-mind approach, is what Altehenger calls ‘motivational partitioning’ (2021b, p547) which can be understood in terms of two claims:

Claim 1. Divided Motivational Architecture: the human mind contains (at least) two distinct motivational systems, M1 and M2.

Claim 2. Proprietary Action-Thesis: self-control is a “proprietary action” of the M2 motivational system. Only M2 can initiate (and sustain) exercises of self-control. (p548)

Sripada and Altehenger think of M2 and M1 as roughly capturing the Platonic divide between Reason and Passion, describing them as ‘the deliberative motivational system and the impulsive motivational system’ (ibid). I will therefore begin by assuming that M1 desires emerge non-reflectively, whereas M2 desires are responsive to deliberative judgements. M1 and M2 desires can diverge, as in the case of an agent who wants a drink, but judges that she shouldn’t have one. Intriguingly Altehenger (2021a) does suggest that ‘one could instead postulate a constitutive link between M2 and an agent’s (i) higher-order desires, (ii) long-term goals, or even (iii) resolutions’ (p258, my italics). I will set this possibility aside for now but return to it in section 5.
A further element of the divided-mind proposal is that there are two separate motivational contests. The first resolves questions about whether to engage in an ordinary ‘overt’ non-control action, and the second resolves questions about whether to engage in a control action, where ‘control actions are a class of actions that serve the function of altering an agent’s current motivational situation’ (p252). Control actions are therefore actions taken to regulate or inhibit the motivational force of desires.

Having specified, in line with claim 2, that only M2 desires can motivate control actions, Altehenger (p255-256) suggest two potential models for how the lower-order non-control action contest is resolved: In the A-version only M1 desires compete with each other, whereas in the B-version M1 and M2 compete directly to motivate overt action.81 In the B-version deliberative motives can overcome impulsive motives in a simple motivational contest, and self-control powers are only initiated when the M2 desire is weaker than the M1 desire. In the A-version self-control is used every time there is a clash. When M1 and M2 are passion and reason, it seems to me that the B-version is more intuitive but I remain neutral on that point.

Before moving on it is worth clarifying a nuance with respect to control actions and their initiation. This will forestall a worry about partition permeability, and indicate the implicit importance of metacognition to the divided-mind solution.

4.2 Resolving the worry of partition permeability

---

81 In the A-version these control actions can down-regulate M1 motives whereas in the B-version they can both down-regulate M1 motives and up-regulate M2 motives.
A potential worry for the divided-mind account arises from the fact that actions used for control are also ordinary actions. Shouting a command, pulling the plug, or calling one’s mother, of course, but also shifting attention, appraising, and even folding one’s arms and doing nothing are all actions that are taken in non-self-control contexts. As such, M1 motivations will influence whether these actions take place. This is not a controversial point: As Sripada states, ‘desires draw one’s attention towards desire-relevant situations and prospects (Scanlon 1998). They also enhance certain hedonic attributes of prospects, such as their expected pleasure, pleasantness, or appeal (Strawson 2009)’ (2014, p56). The question of whether to engage in an ordinary action, including an ordinary action which can be used for control purposes, is decided by a contest which involves both M1 and M2 motivations.

Thus, a motivational partition cannot be understood as indicating that an action such as focusing attention can only be motivated by M2 desires. Instead, it is focusing attention for self-control reasons that can only be motivated by M2 desires. Control actions are desire-controlling actions\textsuperscript{82} which result from desire-controlling motives. These motives are by definition metacognitive. The motives necessary for control actions, in both my hierarchical model, and the divided-mind model are metacognitive.

However, this, by itself, does not dispel the worry of partition permeability. An ordinary action executed for control reasons is still an ordinary action, so an agent who want to execute such an action must overcome ordinary practical motivations against doing so. The necessary move, is to claim that, conceptually and neuro-

\textsuperscript{82} ‘Those actions which constitute exercises of self-control’ (Altehenger, 2021b, p548)
physiologically, there is an intermediate step between the decision to engage self-control action and the execution of the ordinary action for control reasons. This is the step of *activating regulatory powers*. As Sripada states, ‘one’s practical desire provides motivation to engage the regulatory systems that execute willpower. Once triggered regulatory systems operate to suppress the wayward desire’ (p56). Whilst the motivational partition cannot stop impulsive desires from competing to drive action, it can stop impulsive desires from having a say in whether regulatory power is engaged. The initiation of regulatory power in the divided-mind account is an action from which the influence of M1 desires is blocked. When Sripada states that ‘the exercise of willpower is a proprietary action’ (2014, p57), the action to which he refers is the act of initiating regulatory powers. Here is Sripada’s account of the sequence:

1. A is the person’s motivationally strongest M2 desire, and A is opposed by B, an M1 desire that is the persons overall strongest.
2. A provides the motivation for exercising willpower against B, and an exercise of willpower ensues.
3. The R-powers of regulatory systems that implement willpower are sufficient to defeat B.
4. B is defeated (p57)

Step 2 is not an unnecessary explanatory flourish, but is the key additional element of the divided-mind account. The act of initiation *must* be isolated from the act of

---

83 Sripada: ‘Let us review these steps to ensure their cogency. Step 1 presents the standard account of what happens during temptation and is largely uncontroversial. Step 2 is possible because the divided mind view treats willpower as a proprietary action, and its motivational base is exclusively the motivation-encompassing attitudes in the deliberative compartment. Even if one’s practical desire A is motivationally weaker than one’s
control that it causes, for the partition idea to work. If, and only if, this understanding is accepted then the worry of partition permeability can be dismissed.

Both the hierarchical and the divided-mind solution explain an agent’s ability to shift her motivational balance. The former relies on the existence of metacognitive desires, while the latter proposes splitting desires into different kinds, creating a motivational partition, and the existence of a separate and distinct source of “willpower” that gives an agent regulatory power. But which model is correct? In the following analysis I will focus on the argument that only M2 desires can initiate acts of control. In section 7 I will return to the possibility of separate regulatory willpower.

5. The hierarchical divided-mind

5.1 Are desire-controlling motives always deliberative motives?

In the divided-mind account the M1 motivational system is not able to initiate acts of control, and this is perceived to be a positive element of the proposal. As Sripada states, ‘one’s wayward desire does not exercise willpower [control] in order to suppress the motivations that arise from one’s practical judgment. This observation is perhaps so obvious, it might strike one as strange that anyone should take time to draw attention to it’ (2014, p67). However, if we set aside the somewhat divisive word willpower and replace it with control, then it no longer seems so obvious. What about the cheating husband who is with his lover, and turns the picture of his wife

wayward desire B, A can still initiate an exercise of willpower against B. Step 3 is granted by hypothesis. Step 4 follows naturally from steps 1–3.’ (p57-58)
and family towards the wall? Or the addict or anorexic who exerts control against her own best interest. Are their control actions solely driven by, and in accordance with, their deliberative judgement or are they sometimes driven by impulse, habit, or emotion?

More generally, restricting the initiation of regulatory powers to deliberative motives removes an entire spectrum of plausible triggers for control actions, such as non-actional thoughts (Kennett and Smith, 1997), the kind of routinised habits that can be learnt and practised (Gillebaart et al, 2020), and emotions such as guilt or pride (Kotabe, Righetti, & Hofmann, 2019). In their recent review of models of self-regulation, Inzlicht et al state that ‘the role of emotion has been underdeveloped, if not caricatured… we advise more research on how emotion is integrated with and perhaps sometimes facilitates control’ (2021, p337). Is the driver who finds the seatbelt warning tone so irritating that she resists the impulse to buckle-up driven by deliberation or emotion? Finally, ‘people are chronically motivated to feel effective (competent, in control, etc.)’ (Kotabe and Hofmann, 2015, p625). It is possible that, in many instances, self-control can be traced back to an impulsive desire to assert one’s own agency, rather than a deliberative judgement.

I acknowledge that Sripada and Altehenger could perhaps go through the examples above and claim that each one is either deliberatively motivated or not a control action under their definition. This may deflate our disagreement into a terminological difference. Whilst I would be defining a control action as any action that results from a desire to change a desire, they would be defining it as any action that results from a deliberative desire to change a desire. Of course, at that point the motivational

---

partition would become unnecessary: it is not needed to ensure that only actions caused by deliberative desires are caused by deliberative desires!

There is however, a second argument for the view that deliberative motivations are not the only relevant factor in decisions about control. This is based upon analysis of the costs of cognitive control and the case of cognitive fatigue. I will turn to that now.

5.2 Expected Value of Control

Although Altehenger and Sripada focus most on the M2 support for an act of self-control, I will assume that their proposal factors in the costs of such an action, such that self-control only proceeds if the expected benefit of the proposed control is greater than the cost. Indeed, Sripada suggests this in his 2020 paper, highlighting the growing consensus of support for Shenhav et al’s expected value of control model (2013).

Importantly, these costs of control, and cost-benefit computations more generally, are non-deliberative. As Sripada states, the 'calculations are performed via subpersonal routines….in a rational person, these [practical] judgments will likely be an important informational input to the processes that tabulate the benefits and costs of control' (footnote13). If that is the case, then M2 judgements are not technically the sole proprietor of self-control actions, although they may be the sole instigator. Complicating matters further is the fact that the costs of self-control acts are not constant over time. They may fluctuate with cognitive fatigue (Luethi et al 2016, Stewart et al, 2009, Hurley, 2023) or opportunity costs, which can be context-specific (Kurzban, 2013).
I have argued throughout this thesis that the common unit of cost-benefit computations are motives and that agents are (non-deliberatively) motivated to conserve cognitive resources. For any act of self-control that demands cognitive resources, this motive sits on the cost side of the cost-benefit evaluation of whether or not to engage in the control action, and must be defeated by a stronger desire to exert control.

If both the pros and cons in the expected value of control computation are motives, then in a standard motivational contest they would be commensurable. This is in line with the hierarchical model but not the motivational partition model where M1 and M2 refer to the deliberative motivational system and the impulsive motivational system.

At this point, defenders of the divided-mind theory have three alternatives:

They could reject the input of non-M2 costs into expected value of control computations and state that control is initiated purely on the basis of the deliberative motivational system, with no regard for costs. As this goes against the consensus behind the EVC model and Sripada’s own (2020) views, I set aside this possibility.

Alternatively, they could accept the integration of costs of control, but deny that any (other) non-deliberative motives can influence control decisions. Presumably this would be justified by the claim that costs of control are very different from impulsive desires. Although this sounds intuitive, it seems to me to be inconsistent with Kurzban et al.’s widely endorsed theory of cognitive effort, in which they state that the motive to do otherwise is an opportunity cost of ongoing cognition: “Our model predicts that doing the math problems in the presence of the smartphone will be perceived as more effortful than when the smartphone is absent because the
opportunity costs are higher” (2013, p666). Kurzban’s model was also endorsed by Sripada in his 2020 paper.

The third alternative is to drop talk of impulsive and deliberative desires and embrace Altehenger’s postulation of a constitutive link between M2 and an agent's higher-order desires, understanding higher-order to mean metacognitive. In that case the desire to conserve cognitive resources would itself be a metacognitive desire so would be part of M2. The higher-order contest would then be between the motive to conserve cognitive resources and the motive to use cognitive resources to regulate the wayward desire. The expected value computation would be a natural consequence of the contest of M2 desires. Presumably this interpretation of the divided-mind model would be implemented along the lines of Altehenger’s A-version, where only metacognitive M2 motives are involved in motivating metacognitive control actions, and only practical M1 motives are involved in motivating non-control actions. When a control-action has an expected value greater than its costs, then the agent will (try to) regulate or inhibit M1 motives.

Accepting such a restatement would have the advantage of simplifying the theory. The old proposal that metacognitive actions are driven by deliberative desires, would be replaced with the intuitive idea that metacognitive actions are driven by metacognitive desires. Equally, it could be accepted by all parties without the need to litigate the question of whether impulsive desires can be metacognitive. For example, if Altehenger and Sripada believe that the set of metacognitive motives only includes deliberative motives and costs of control, then they can embrace the

\[\text{Note that I am assuming here that higher-order desires are metacognitive desires, rather than what Kotabe and Hofman call higher-order goals which are ‘a relatively ‘cool’ cognitive construct’ (2015). Higher-order goals, unlike metacognitive desires, are captured under the previous discussion of deliberative desires.}\]
new language without risking pollution from impulsive desires. On the other hand, those who believe that the desire for control can arise from non-deliberative means (Kotabe and Hofman, 2015, Inzlicht et al., 2021) would lose one reason to resist the divided-mind proposal.

The proposal that M2 equals metacognitive suggests significant convergence between the divided-mind model and the hierarchical model. In the following section I will highlight the remaining differences.

6. Comparing the Unified and Divided Hierarchical Models

The preceding analysis suggests that the strongest version of the divided-mind proposal is hierarchical. In this model, M1 captures all motives, whether deliberative or impulsive, about practical, non-control actions and M2 captures all metacognitive motives, whether deliberative or impulsive about metacognitive control actions. M2 includes any motive towards regulating practical motives, and any motives about the allocation of cognitive resources. Only creatures with metacognitive monitoring capabilities can have such ‘managerial’ motives (Hieronymi, 2009).

In the divided-mind hypothesis, M1 and M2 would be incommensurable and there would be no means-end rationality between the two contests. Synchronic self-control would be attempted if net M2 was positive and successful if the causal powers of the actions initiated by M2 were great enough to shift the balance of M1 desires.

In contrast, in the unified model, means-end rationality remains. This makes the hurdle for an act of control higher. It will only happen if the net motivation to exert control is greater than the net motivation to act on the lower-order desire. On the
side of engage control is the motive towards virtuous V-ing as well as the metacognitive motive towards control. On the side of refrain from control is the motive towards wayward W-ing as well as the motivation to conserve cognitive resources. Therefore, for an agent to try to perform an act of synchronic self-control, her desire to exert control must be great enough to overcome the combination of the net motivational balance of the lower order contest and the costs of control.

In the unified model, an unbalanced practical contest implies that the agent is highly unlikely to engage in self-control. Effectively, means-end rationality adds an extra cost of control, equal to the net motivation to pursue the wayward act. The divided-mind proposal, on the other hand, sets a much lower bar for control. Because the motivational base of practical desires is irrelevant to the metacognitive question, any direct desire for control that is greater than the direct costs of control is sufficient to initiate a control action.

This difference between the two models should be testable empirically. Is it harder to engage self-control actions if the wayward desire is stronger? As a philosopher, a thought experiment is the best I can offer:

Imagine that an agent desires an alcoholic drink but she also has a metacognitive desire to control her wayward desire to drink. Fortunately, she has a control action, which is to call her mother. She knows that if she does so, her mother will talk her out of her desire to drink (no matter how strong the desire). Assume her drink-desire comes in three levels: Low, high, and extreme. Low she can handle herself, high and extreme she cannot. Between high and extreme does the strength of her desire for alcohol influence the likelihood of making the call? I would suggest it does, which is consistent with the unified model. At high she will make the call but at extreme her
desire for the drink will be so strong that the call feels like a terrible idea. Of course, this view merely reflects an intuition that means-end rationality does apply to control actions\textsuperscript{86}. The divided-mind, by removing the constraints of instrumental rationality, seems to me to go too far. It makes the initiation of a control action too easy.

7. A brief history of “willpower”

The final difference between the unified and divided models is the idea of a separate and distinct source of “willpower” that sits outside of the agent’s normal motivational system. Although this was proposed by Sripada in 2014 as a solution to the problem of synchronic self-control, it was also consistent with the then prevailing theories of effort. For example, Baumeister’s ego depletion theory supported the idea that willpower is a finite resource which is used up in acts of self-control, and that feelings of effort track the amount of the resource consumed (Masicampo and Baumeister, 2008, see Friese et al, 2019, and Hurley, 2023, for the current state of the replication crisis).

The resource depletion account of effort received sustained criticism, particularly from Kurzban (2010, 2013, 2016) who offered a rival theory, proposing that mental effort is the experiential manifestation of the calculated opportunity costs associated with a mental task (see also Székely and Michael, 2020). A 2019 paper, co-authored by Baumeister (Andre, Audiffren and Baumeister), agreed with the broad thrust of

\textsuperscript{86} One potential response that the above vignette brings to mind is that very strong desires have a concerning ability to influence reasoning. Possibly the agent’s desire for alcohol led to a process of rationalisation in which she persuaded herself that she had been working hard and deserved a drink \textit{just tonight}. In that case she doesn’t call her mother because she no longer judges it best not to drink. This conjures up the idea of a messy battleground where M2 desires influence M1 desires through self-control actions, and M1 desires influence M2 desires through motivated reasoning. I will assume that motivated reasoning is an extra argument against, rather than for, motivational partitioning.
Kurzban’s theory. The authors proposed a Mechanism of Effort which integrates perceived costs and benefits, and outputs ‘decisions regarding the intensity and the direction of the engagement in effort in ongoing or future tasks’ (p4) as well as a feeling of effort. This feeling, the authors suggested, makes the agent aware of the perceived costs associated with achieving the goal of the task. More recently Luethi et al. showed that any decline in performance due to cognitive fatigue can be reversed if the participant is given financial incentives (2016, see also Muraven and Slessareva, 2003, Stewart et al 2009). It seems that increasing motivation increases the ability to engage in cognitive control, which undermines the belief that willpower is separate and distinct pool or muscle.

Although “willpower” was part of the title of Sripada’s 2014 paper, it was not mentioned at all in his 2020 paper or in either of Altehenger’s 2021 papers. Indeed, Sripada stated that ‘effort serves as a non-conceptual valence-type representation (Carruthers, 2018), with a distinctive phenomenal character, of the disvalue of exercising cognitive control in a particular context’, and in footnote 28 described Kurzban’s theory as ‘broadly on the right track’.

In chapter 4 I presented my own Kurzbanian theory of the feeling of effort. My novel claim was that the common unit on which cost-benefit computations are performed is motivational force. Motives in favour of an action are benefits, and motives against are costs. The disvalue of cognitive control is conative. This claim is consistent with the unified hierarchical model of self-control: Feelings of effort during self-control do not indicate a depletion of willpower or the existence of a motivational partition, but that an ongoing desire to do otherwise is being inhibited.
This theory not only unifies all kinds of effort, but it also explains the different phenomenology in two cases of self-control taken from Sripada 2020:

Mo judges it is best that he not smoke, so he visits a doctor who specializes in helping smokers quit quickly and effortlessly. The doctor hands Mo a special little pill and tells him it tastes like candy. Mo has no desire to smoke at the moment, but he knows he will have one shortly. He takes the pill, and he never has a desire to smoke again.

Ro judges it is best that she not smoke. However, right now, she has a very strong desire to smoke. She immediately effortfully resists acting on this desire until it passes. As a result, she does not smoke. (p817)

Ro feels effort because her act of synchronic self-control involves temporally extended inhibition of an ongoing desire to do otherwise. In his 2014 paper, Sripada calls this ‘full-blooded willpower to reflect the fact that exerting willpower over time in this way often feels particularly difficult and effortful’ (p48). Whether Ro resists smoking in this way because her virtuous desire is stronger, or because she has a metacognitive desire for control that (when added to V) defeats the desire to smoke, is not hugely important. As long as W is being inhibited effort will be felt. Despite its name, the conative phenomenology of “full-blooded willpower” is not an argument for a separate and distinct source of “willpower” that sits outside of the agent’s normal motivational system.

In contrast, Mo’s diachronic self-control is phenomenologically silent because the pill ensures there is no desire to resist. Both experiences are explained by the framework outlined in chapter 4.
This analysis raises the question of whether the divided-mind account can stand alone, without the support of a separate and distinct source of willpower? I think the answer is that it cannot. To see why, recall from section 4 that the actions used to control desires are just ordinary actions - shouting, focusing attention etc. The motivational partition cannot stop M1 desires from competing to drive these actions. To resolve the worry of partition permeability it was proposed that M1 desires were unable to competing to *activate regulatory powers*. Step 2 of Sripada’s sequence stated that ‘A provides the motivation for exercising willpower against B, and an exercise of willpower ensues’ and in step 3, ‘the R-powers of regulatory systems that implement willpower are sufficient to defeat B’ (p57). Steps 2 and 3 were explicitly separated because activating willpower is an action. It had to be for the motivational partition idea to work. If “exercising willpower” was replaced with “executing a control action” such as shouting or focusing attention then M1 desires would naturally have a say, because these actions are also ordinary actions.

However, if willpower is no longer a separate source of regulatory power, then the intermediate activation step is removed. There is no longer an act of initiation of power which is isolated from the control action. Unfortunately for the divided-mind proposal, that leaves the motivational partition without a role. Its only task was to block M1 desires from motivating an act of initiation that no longer exists. Without a separate and distinct source of willpower, the motivational partition idea should be retired.

I therefore conclude that the mind is not divided by a motivational partition. It is unified, but hierarchical.
Conclusion

In this chapter I resolved the mismatch problem by focusing on metacognitive motives.

From the perspective of the overall thesis, I have shown that the hierarchical motivational strength idea is consistent with the possibility of intentional synchronic self-control. Metacognitive motives, despite being subject to instrumental rationality and the motivational strength idea, are able to motivate control actions.

If one zooms in and focuses purely on the practical decision, then a control action looks like an agential tipping of the scales, which asymmetrically biases the decision. But zooming out, symmetry is restored and bias removed. An agent can only engage in a metacognitive control action if she is most motivated to do so.
Chapter 6
Is Alienation a Problem?

Introduction

The focus of Part II has been on building a model of decision-making which is based on motives and motivational conflict. I have suggested that activated motives compete in a winner-takes-all contest based on motivational force, and only inhibited motives are felt. This model is simple, in that when an agent chooses, she will always choose to do what she most wants to do. It is symmetrical, in that all activated motives are treated equally and, in any motivational contest, the only determining factor is motivational force. It is hierarchical, in that motivational contests also exist at the metacognitive level. And it is non-interventionist, in that an agent cannot ‘rise above the desires to which one is subject, and to take the reins in one’s own hands’ (Wallace, 1999, p234). Intentional synchronic self-control is possible, but control actions work within and through the motivational strength idea.

In this chapter I consider the possibility that an agent may feel alienated from a force that moves her to act: She may divide the forces that move her into motives with which she identifies, and motives from which she is alienated. Thus, like a phoenix from the ashes, the worry of asymmetry returns. The symmetric motivational strength idea was shown to be consistent with conative phenomenology and intentional synchronic self-control, but is it also consistent with alien motives? This worry takes two forms, one epistemological and the other ontological.
The epistemological worry: The experience of alienation indicates that the proposed model of intentional action is flawed.

Does the possibility of alien motives indicate that any symmetrical model must be wrong? To answer this question, I need to first understand what it is that makes an agent categorise a motive as alien. By analysing potential solutions to what I will call the “subjective categorisation problem", I can deepen my psychological understanding and validate, invalidate, or perhaps enhance my model. I will find that both judgement-based and conative phenomenology-based solutions to the categorisation problem are consistent with my model. The possibility of a non-conative phenomenology of alienation, exemplified by a divergence between liking and wanting in addiction, is most interesting in that it helps address some outstanding questions. I conclude that there is nothing within the experience of alienation which undermines the motivational strength idea.

The ontological worry: The phenomenology of alienation indicates that although my model successfully explains intentional action, some of that action is not agential action.

This second worry is that I have built an accurate model, but of the wrong thing. Perhaps I have cast the net too wide, catching all motives rather than the motives of the agent\(^{87}\). I will argue against a hard version of this worry, in which agential motives and alien motives are of fundamentally different natural-psychological kinds. A softer version remains, however, in which a motive is called alien (or half alien) if it is inauthentic, or fails to satisfy a norm of agency. This would be consistent with an

\(^{87}\) Ryle famously told of the tourist, who after seeing all the colleges, libraries and research departments asked “but where is the University?” (2009, p6). Here the risk is that the tourist thinks the whole city is part of the University.
“incremental” view, in which agency captures only what is distinctive about humans. I will argue that this view is not without problems and that, for the phenomenologically informed project in which I am engaged, an “inclusive” view of intentional agency is more appropriate.

But first, to set up the problem, I introduce the unwilling addict.

The Unwilling Addict

Frankfurt famously stated that an ‘unwilling addict may meaningfully make the analytically puzzling [statement] that the force moving him to take the drug is a force other than his own’ (1988, p18). The unwilling addict feels alienated from the ‘irresistible thrust’ of the force to take drugs and denies that the desire for drugs is an expression of his agency. He feels that the force moving him to act is a force external to, and conflicting with, who he really is in his deep-self. I will assume that, at least pre-theoretically, the case of the unwilling addict supports the intuition that an agent may divide the forces that motivate her into those with which she identifies, and those from which she is alienated.

For those, such as myself, who have a motive-centred model of agency, it is tempting to focus on the fact that Frankfurt’s addicts are stipulated to have an irresistible desire to use drugs. I have defined motives as inhibitable, so if irresistible is understood as uninhibitable then perhaps the force that addicts are subject to is something more like a brute behavioural reflex than a motive. If all cases of alienation were examples of uninhibitable forces, such as in cases of alien hand syndrome (Assal et al., 2007), then the proposed model of agency based on motives
would stand. All motives would be identified-with and alienation restricted to irresistible reflexes or tics.

There are two major problems with this approach.

First, reflex responses are, from a conative perspective, phenomenologically silent, in direct contrast to cravings, which may be so strong they are unbearable. A consistent theme throughout this thesis has been the link between inhibition and phenomenology. Reflexes, tics and alien hand syndrome reinforce this idea: they are functionally isolated from the rest of an agent’s cognitive infrastructure and outside the inhibitory system, so there is no feeling of motivational pull. If alienation referred to irresistible forces, then there would be no conative phenomenology of alienation to analyse.

Second, it is an empirical fact that drug-desires work through, rather than around, the motivational system. For example, Pickard argues that addiction is not compulsive but purposive, and should be treated as such. Treatment ‘requires acknowledging that addicts are agents—agents who use drugs as means to understandable ends’ (2012, p11). Despite talk of compulsion, ‘individuals with addiction are typically highly sensitive to incentives to defer use when the situation calls for it, e.g. when a policeman “is at the elbow”’ (Sripada, 2022). I will therefore restrict my analysis to alien motives, as described in the unwilling addict vignette. I will leave it to others to elucidate the relation between hard-to-resist alien motives and the uncontrollable responses of an alien hand.

I will proceed on the basis that alienation is a philosophically interesting phenomenon that relates to an attitude that motivates action, and that alien motives interact with other motives in a motivational contest. The outstanding issue with
respect to these harmful motives, is not whether they are motivational but whether they are alien to the agent.

1. The Epistemological Worry

Although persistent conflict between motives was a central part of the model of conative phenomenology developed in chapters 3 and 4, that model did not include any conative phenomenological signpost which would allow an agent to differentiate between alien and identified-with motives. Given this, how does an agent differentiate between an identified-with and alien motive? This, I will call the subjective categorisation problem.

Jaworska has argued that there are two senses of the concept alienation: ‘The psychological, subjective sense [of] how the agent perceives aspects of his psychology and whether he regards them as his own’, and the ontological sense, ‘where the task [of the philosopher] is to pick out attitudes that properly belong to the agent from the sea of happenings in the agent’s psychic life’ (2007, 531). The qualifier subjective, attached to the categorisation problem, highlights that my focus in this section is on Jaworska’s psychological, subjective sense of alienation.

My primary aim is to confirm that the phenomenology of alienation is compatible with the findings of chapters 1 to 5. I will consider three potential solutions to the subjective categorisation problem, situating each proposal in the context of the analysis of the thesis thus far. I am open to the possibility that all three contribute in some way to feelings of, or beliefs about, alienation.
1.1 Solution 1: Judgement

The starting point for a judgement-based solution to the categorisation problem is the possibility of divergence between an agent’s predominant practical motive and her judgement about what is best: If she judges, after consideration, that she should not act in line with a motive, but does so anyway, then she might claim alienation from that motive (Watson, 1982)\(^\text{88}\). Rather than unwilling, an addict in this framework is perhaps better described as unreasonable.

Agents who act against their own best judgement are usually described as akratic, where ‘core akratic action’ is defined by Mele as: ‘free, sane, intentional action that, as the non-depressed agent consciously recognizes at the time of action, is contrary to his better judgment’ (2012, p.8). The debate over whether akratic action is possible has a long history (Plato, 1976, Davidson, 1980). I will assume that, at the very least, it is possible for an agent to believe that her own behaviour is akratic. If an agent could never form such a belief, then, obviously, beliefs about akrasia could not be the cause of feelings of alienation.

However, there is a potential worry that a simple one-to-one relationship between akrasia and alienation leads to false-positive alien-categorisation. Killmister describes what she calls the Woody Allen Puzzle: ‘those motivational attitudes that seem so unreasonable … can become an integral part of who we are’ (2015, p729). If Suzy is too scared to get on a plane, even though she judges that planes are safe, then is she alienated from the motive to avoid planes, or is it a core part of who she

\(^\text{88}\) An alternative view might be that alienation is merely a post-hoc defensive response to a failure to find a socially acceptable justification for one’s motive (for an argument along these lines with respect to reasons, see Mercier and Sperber, 2017).
is? Beliefs about akrasia, don’t seem to be sufficient to solve the subjective categorisation problem.

One possible response is that the categorisation of motives is judgement based, but that the judgement takes in other factors as well as beliefs about akrasia. As Killmister argues: ‘there is a conceptual gap between taking a motivational attitude to be reasonable, and taking it to be a reason for action’. If an unreasonable attitude is authentic, ‘in the sense that the agent takes it to express a deep truth about themselves’ (2015, p738) then it has reason-giving force, even if it is unreasonable.

Perhaps a motive is alien only if it is inauthentic? Although an attractive solution, this, of course, adds a new categorisation problem. How does an agent know whether an attitude is authentic? In the absence of a clear marker of authenticity, the most obvious proposal is that an inauthentic attitude is one that clashes with the agent’s own self-description, their narrative self. An alien motive would therefore be defined as a motive that is inconsistent with an agent’s narrative self. Although this does seem intuitive, self-descriptions are often simplifications and/or self-deceptions: ‘We can be – and often are – mistaken in our stories about who we are and why we act and feel the way we do’ (Kristjánsson, 2010, p43). ‘The true self is, shall we say, evidence-insensitive… claims made on its behalf may completely contradict all available data, as when the hopelessly miserable and knavish are nonetheless deemed good ‘deep down’. (Strohminger et al., 2017). In extreme cases, designating as “alien” narratively uncomfortable desires can be a form of repression which should be rejected in favour of acceptance that they are one’s own (Velleman, 2005).

These issues raise questions about whether coherence with the narrative-self should be used for categorisation, but for now I will accept that it is possible that
categorisation is narrative-based. I will reserve further ontological discussions, such as whether an inauthentic act can be the act of an agent, until section 2.

With respect to the phenomenological worry, if the categorisation of motives is based solely on judgements, of rationality or narrative compatibility, rather than feelings, then there is no phenomenological gap due to alienation that must be closed\(^{89}\). Judgement based solutions to the categorisation problem are perhaps the simplest to integrate within my overall thesis, as they are consistent with the motivational strength idea and are phenomenologically neutral.

1.2 Solution 2: Meta-conative phenomenology

The second potential solution to the categorisation problem is based on Frankfurt’s initial proposal (1988) that an agent self-reflectively categorises her practical motives based upon whether she wants them to be effective. She identifies with any practical motive that she wants to be effective, or is alienated from any practical motive that she wants not to be effective\(^{90}\). For example, the unwilling addict can be contrasted with a willing addict: Whilst, the latter identifies with his addictive-motive, the former does not. Unlike solution 1, this is a phenomenologically-based categorisation scheme. The agent categorises based upon (meta-)conative phenomenology.

Similar to solution 1, there is a worry that this leads to false-positive alienation categorisation, as an agent with a desire to engage in a control action would, under

-----

\(^{89}\) A second question is whether such judgements impact motivation. They could do so if the agent has an antecedent motivation (Mele, 2003), such as a higher-order desire to be rational (Velleman, 1996) or to do what one all-things-considered should. I analysed the possibilities of combined belief-desire attitudes, so-called besires (Kriegel, 2011) in chapter 3 and concluded that they were theoretically separable, and neither necessary nor sufficient for rational or moral behaviour.

\(^{90}\) The former requires a conative attitude towards all identified-with motives, whereas the latter implies that an agent identifies with any motive towards which she does not have a negative conative attitude.
this account, always be alienated from the desire she wishes to control. When Julia is considering the control action of pulling the bath plug, to reduce her desire to stay in the bath, is she necessarily alienated from her desire to stay in the bath? A simple version of the theory suggests she is, but that seems wrong. After all, she is a bath-lover! The metacognition-based categorisation scheme needs to be augmented.

Suggestions for augmentation generally take the form of somehow ensuring that the second-order desire is the right kind of desire. However, for this to be relevant to the subjective categorisation problem the agent must be able to differentiate between these two kinds of metacognitive motives. This cannot be achieved through conative phenomenology alone. It therefore seems unlikely that meta-conative phenomenology is the full answer to the categorisation problem. However, if I am wrong, and categorisation is based on the existence of metacognitive motives, then this suggestion is easily integrated with the rest of the thesis. Metacognitive motives are fundamental to the hierarchical, motivational model developed in chapter 5 on intentional synchronic self-control.

1.3 Solution 3: Non-conative phenomenology

The fact that the unwilling addict’s behaviour is driven by a purposive motive rather than an irresistible reflex response opens up the possibility that drug-desires are just ordinary desires and therefore should be identified-with. This position would be eliminativist about alienation. However, an advocate for alienation could justifiably respond that, between uninhibitable responses and ordinary motives, there remains space for alien desires. Indeed, although desires for narcotics may be theoretically
resistible, they do appear to be practically irresistible for addicts in some circumstances.

Empirical research has shown that although drug-desires work through the motivational system, narcotic use can lead to ‘sensitization [which] renders brain ‘wanting’ systems hyper-reactive to drug cues and contexts, thus conferring more intense incentive salience on those cues or contexts’ (Berridge and Robinson, 2016, p9). These motives might be called “cravings” (Holton and Berridge, 2013) and are often accompanied by a dissociation between wanting and liking. What the unwilling addict feels is conflict between the alien craving to take drugs and the identified-with desire to resist.

The empirical findings suggest that addiction impacts the system that underlies both cue-based activation of occurrent motives and the allocation of value to ends, inappropriately ‘tagging some of these options as especially salient and rewarding’ (Levy, 2015, p62). To rephrase in the language of this thesis, addiction both increases the likelihood that a motive is activated and persists, and boosts the intensity of its force. More generally, addiction distorts the agent’s evaluative landscape such that conative predictions of care value diverge from actual care value. The motivation to give in to temptation defeats the motivation to resist, in a conative cost-benefit computation.

Note that although the traditional model of addiction might imply that resistance is overwhelmed by a single strike of overwhelming force, contemporary models suggest that addiction works through multiple smaller but varied and temporally extended attacks (Sripada 2022). An analogy might be that addiction overwhelms
defences like a multi-day drone-swarm attack, instead of with a single bunker-buster bomb.

From a conative phenomenological perspective there is no difference between a craving and a particularly strong, or persistent, or frequently reoccurring desire, but the dissociation of wanting and liking highlights the fact that there is a difference in the broader non-conative phenomenology. If the question is how to differentiate between alien and identified-with motives in the absence of a conative phenomenological signpost, then an obvious answer is to make use of a non-conative phenomenology.

Berridge’s empirical findings imply that at least part of the phenomenology that accompanies unwilling addiction is non-conative: a negatively-valenced, non-conceptual, affective ‘signal’ from the psychological infrastructure. Although non-conative feelings have no direct synchronic impact on action, this solution may be able to resolve an outstanding issue from the self-control literature. When non-actional and actional forms of self-control were analysed in chapter 5, the proposals all involved an initial trigger which raised a question and/or activated a motive that ultimately resulted in a shift in the motivational balance. But it was not specified what that trigger might be. It is possible that the non-conative phenomenology of disliking a practical motive could fulfil this role. Like all aspects of the affective backdrop, such signals would immediately become part of the current state-of-affairs which activates motives. Just as feelings of uncertainty may lead to activation of epistemically virtuous motivations, feelings of alienation may lead to activation of morally virtuous
motivations. There is a feedback loop from monitoring, via affect and activation, to control.

Strikingly, this discussion has landed very close to a suggestion made by Kennett and Smith. Their suggestion was that at the very moment of an impending failure of control, an agent may have a non-actional thought, which updates the agents set of occurrent desires such that there is no longer a failure of control. In response to expected objections that their solution was ad-hoc, Kennett and Smith considered some possible causes of such a thought:

‘Feelings of shame, the thought that one is stupid, a certain inward focus of attention, any of these may have the requisite effect, and folk wisdom tells us that one or another of these, or something similar, will have the requisite effect upon most of us’ (1996, p70)

Presumably it is uncontroversial to add to this list non-conative feelings of alienation, such as disliking a currently predominant desire.

It is not only non-actional thoughts that could be triggered by the affective backdrop. The motive to undertake a synchronic “ancillary action” (Mele, 1998), such as shouting a self-command to resist temptation, could also be activated. Altehenger challenged Mele to provide ‘some sort of explanation of what, specifically, brings it about that [the agent] desires more to utter a self-command than to refrain from doing so” (2021a, p248). In chapter 5 I defended the possibility of synchronic control actions from a functional perspective. This analysis adds a phenomenological explanation for the trigger. If an agent has a negatively valenced feeling about her

---

91 A stronger version of this argument would state that there is innate or learnt value in the reduction of negatively valenced feelings of alienation. There is a more depressing, alternative perspective, however: agents may feel that it is impossible to defeat alien motives, so will not even try.
strongest motive then it is not inconceivable that this would trigger a metacognitive motive to engage in a control action.

Even if there is failure of synchronic control, negative feelings may lead an agent to take steps to influence the actions of her future self. In the search for practical steps, she can make use of the myriad of tips and tricks detailed in self-help books, relying on resolutions, community support, positive habit formation or so-called tying to the mast strategies.

If non-conative phenomenology underpins feelings of alienation, then that is consistent with the analysis of the thesis thus far, and possibly supportive.

1.4 Summary

The analysis of solutions 1-3 has revealed that the phenomenology of alienation is difficult to pin down. It is possible that an agent categorises motives by combining all sources of information: by considering not only her judgements about what is best and reason-giving, but also her conative and non-conative feelings. Whilst this methodology might be difficult to elucidate formally, an agent may apply an “I know it when I see it” approach to alien motives.

Although questions remain, from the perspective of the epistemological worry the analysis has been comforting. There is nothing within the phenomenology of alienation to suggest that the motivational strength idea is wrong. If the phenomenology is non-conative then it might even be supportive of a wider model which seeks to flesh out motive activation.
However, any feelings of relief should be tempered, because the ontological worry remains and is potentially significant. Perhaps I have described a model of human decision-making, but not agential decision making. The possibility remains that an intentional action is not necessarily an agential action. I will consider a hard and a soft version of this view. In the hard view of alienation, agential and alien motives belong to fundamentally different natural-psychological kinds and there is an uncrossable ontological divide between the two. I will argue against this view. In the soft view, whether a motive is agential or alien depends on whether it reflects her authentic self, somehow defined. This opens up the possibility that motives can exist in a partially alienated state. Although I am personally inclined to believe that inauthentic action is still agential action, I will argue that taking an alternative position on this question would not undermine the findings of my thesis.

Before doing so, however, it is necessary to set aside one possible explanation for the intuition that motives can be divided into different kinds: The feeling of an effort of will. If such a feeling, or indeed the existence of intentional synchronic self-control, were to indicate that the mind is divided and only one part has access to a “reservoir of willpower”, then that would be a strong argument for an ontological split. The findings of chapters 4 and 5 - that feelings of effort and intentional synchronic self-control are fully explained within the motivational strength idea and that the mind is hierarchical but undivided - are therefore relevant for the ontological question. When combined with the analysis of this section, they confirm that there is no phenomenological justification for a model that divides the mind into agent and alien\textsuperscript{92}. I acknowledge that if my findings with respect to conative phenomenology

\textsuperscript{92} In part III I show that the act of effortfully making a difficult decision is also consistent the motivational strength idea.
were shown to be wrong, then the case for the hard view of alienation would be strengthened. For now, however, I will proceed on the basis that the previous analysis was correct, or at least internally consistent.

2. The Ontological Worry

2.1 Alienation and harm

The word autonomy comes from the combination of “auto”, self, and “nomos”, norms. This suggest that an agent is an individual who interacts asymmetrically with the world according to a set of norms (Barandiaran et al, 2009).

In chapter 1 I suggested that the ultimate goal of any agent is to sustain or care for herself. Her “duty of care” is complicated by the fact that it applies to her synchronic and diachronic self, to her physical and narrative identity, and to how long she spends deliberating. This results in a crowded conative environment in which motivational conflict is frequent and significant. The management of this motivational conflict is central to what it means to be a flexible human agent and perhaps the most important driver of the phenomenology of agency.

The analysis of section 1 was consistent with this view. Whether categorised through judgements of akrasia and lack of authenticity, or a certain kind of metacognitive motive, or through dislike, an alien motive is perhaps best described as a motive that breaks agential norms. When implemented, an alien motive harms rather than promotes care for the agent (see Shoemaker, 2003; Sripada, 2015). It is important to clarify, however, that due to the complexity and multi-dimensionality of human care, as well as the winner-takes-all nature of motivational conflict, many care-based motives may be harmful on some dimension, or have harmful side-effects. Consider
the motivation to eat. This is clearly in service of self-sustainability even if it sometimes conflicts with the desire to stay thin or to keep working on a thesis. I have described human agency as sophisticated, but it is probably more accurate to say that it is complex and multi-dimensional. Given the normal rough and tumble of conative conflict, how should one think about the link between harm and alien motives?

In what follows I will initially assume an extreme definition of alienation in which an alien motive is a motive that is inconsistent with the norms of self-care on all dimensions. I am not sure that such a motive exists but assuming it does represents the best hope for the hard view of alienation. If such a harmful motive, which breaks the norms of agency, is not categorically distinct from a pure, authentic motive (which satisfies the norms of self-care on all dimensions) then alien and agential motives cannot be split into natural-psychological kinds. After conducting this analysis of motives at the extremes, I will relax the assumption.

2.2 Norms or model outputs?

Although the extreme definition of alienation is highly restrictive, it does not necessarily create a null set because conative architecture is fallible. This fallibility raises a subtle question at the core of agency: Is an agent an individual who acts in accordance with norms, or an individual who acts in accordance with an internal model which aims to satisfy norms? Put another way, does an agent's set of motives include all motives that arise from the (fallible) conative machinery that exists to implement the self-care norm, or only those motives which do satisfy the normative requirements?
This difference can be seen in the following two accounts of minimal agency:

An agent is an autonomous organization capable of adaptively regulating its coupling with the environment according to the norms established by its own viability conditions (Barandian et al., 2009, p8)

An agent is an information-processing system which is characterised by modelling activity with respect to its surrounding environment, geared towards successive interaction (Barzegar, Margoni and Oriti, 2023, p3)

Whilst the talk so far has been of care-based norms, in a naturalist model, those do not impact behaviour directly. The proximate drivers of action are motives, which are the output of a motive formation process\textsuperscript{93}. The important question for current purposes is how such a process can produce motives that are (sometimes) aligned with care-based norms. Fortunately, when considering this question, philosophers can draw on analysis conducted within epistemology on belief formation processes. These provide the basis for the first of three arguments against the hard view of alienation which, as a reminder, I expressed as follows:

Agential and alien motives belong to fundamentally different natural-psychological kinds and there is an uncrossable ontological divide between the two.

\textsuperscript{93} See Berridge and Holton (2013) for a detailed scientific description of how this process might work (and update). See Gillian and Tye for an integration of physical and social needs: ‘This collection of findings compels the hypothesis that dopaminergic signalling may be involved in the initial response to social isolation, but that downstream regions (including the BNST and CeA) might exhibit longer term remodelling/plasticity following chronic isolation. Indeed, there is considerable evidence to support a similar model for the stages of drug-evoked plasticity in the mesocorticlimbic dopamine system.’ (2019, p11)
2.3 Argument 1: The argument from attitude formation processes

Motives are not the only attitudes produced by an attitude formation process. Whilst motive formation processes aim to satisfy care-based norms, belief formation processes aim to satisfy truth-based epistemic norms. Both processes are fallible: motives can be self-harming and beliefs can be untrue. Within epistemology, the reliability of belief formation systems has been, and continues to be, widely analysed in the context of justification and warrant. For example, Alston described reliable belief formation systems as those that “yield a high proportion of truths over a wide range of situations of the sort we typically encounter” (1995, p10) and Plantinga argued that a belief has warrant ‘if and only if that belief that belief is produced in S by his [truth-aiming] epistemic faculties working properly in an appropriate environment’94 (1993, p9). As far as I am aware, no account in epistemology has questioned the fact that the output of a belief formation process is a belief of the agent. The belief may be more or less true, justified or warranted, but it is still the belief of an agent. She may have a false belief, a belief that was formed from a biased evidence set, or a belief that she wishes that she did not have, but it is still her belief. Why should motives produced by motive formation system be any different?

The fundamental requirement of a belief-formation process is that it is sensitive to evidence. A mental state that was not produced by an evidence-sensitive process is arguably not a belief, even if it satisfies some or all of the roles of a belief (see Gendler on ‘Aliefs’ (2008, but, for a contrasting view see Mandelbaum, 2013). Although controversial, one attractive implication of applying this requirement to

94 ‘Proper functioning, of course, comes in degrees’ (1993, p10)
motive formation processes would be that a motive inserted by an evil neuro-
scientist could be regarded as a different kind of motive, or even not a motive, even if
it interacts with other motives and drives behaviour in the normal fashion. In this
case, the attitude would be distinct, because its functional strength does not reflect
the agent’s conative values.

This suggest that for the hypothetical case of an implanted motive, the hard view of
alienation might be true, but that for harmful (but inhabitable) motives produced by a
motive formation system the hard view of alienation is false. Harmful motives are not
of a different natural kind to normal motives. One potential worry about this
conclusion might be that although alien motives are the output of a motive formation
process, the process is malfunctioning. This could be due to addiction, for example.
However, even if an attitude such as a craving is the output of a malfunctioning
process, that does not necessarily imply that the hard view is correct. Within
epistemology, discussion of proper functioning or reliability are relevant to questions
of warrant or justification, rather than ownership or alienation. If an agent has an
unwarranted belief, it is still her belief.

In the following section I will consider the specific case of addiction.

2.4 Argument 2: The argument from addiction

In the previous section I raised the possibility that the cravings of an unwilling addict
are the output of a malfunctioning motive formation process. In this section I consider
whether contemporary views of addiction support this view. An initial answer is that
the process of motivation formation is not necessarily malfunctioning, but the input is
extreme:
Directly or indirectly, all addictive drugs work by triggering exaggerated but transient increases in extracellular DA [dopamine] ... Such DA surges resemble, and in some instances greatly surpass, the physiological increases triggered by naturally pleasurable stimuli’ (Volkow et al, 2010, p749)

This is a normally reliable desire formation system, working in accordance with its care-based (natural) design plan, responding to highly harmful inputs. Harmful inputs can lead to harmful motives, just as false evidence can lead to false beliefs. This suggests that not only is a drug craving a normal motive in the sense that it is not irresistible, but it is also formed through a normal process. If this is true then it is difficult to see why a harmful motive should be categorised as a different kind.

However, chronic addiction does lead to neuro-adaptation and distortion of the whole conative evaluative landscape. Thus, the motivational system as a whole begins to function less well. But this is a multi-dimensional problem. It is not just that the craving is strong but also that memory may be impaired (George et al, 2008), temporal discounting curves are steepened (Ainslie, 2018), and the motivation to do otherwise is suppressed by downregulated response to other incentives (Volkow, Wise and Baler, 2017). Moreover, Sripada argues that drug addiction involves ‘large populations of distorted thoughts [which] unfold over extended periods of time’, the regulation of which is unreliable, in part because they are ‘not particularly salient’. ‘Compulsive desires win by strength, but distorted automatic thoughts win by stealth’ (2022 p 5, my italics).

Even if one resists the specifics of Sripada’s account, a complex multi-dimensional picture of addiction is emerging. It is less that one particular motive is of a different kind and should be considered outside of the unwilling addict’s agency, and more
that the unwilling addict, as a whole, is becoming less able to care for himself. Drug use has changed him. He is no longer who he used to be. The focus on an individual motive seems misdirected.

The hard view of alienation is narrow but deep, in that it isolates an individual motive and identifies it as being of a different kind. In contrast, contemporary understanding of chronic addiction suggests a broad but shallow impact in which neuro-adaptation causes a shift in the wider conative landscape. This impacts all, or many, motives, but not enough for any to jump an ontological gap and become a different kind.

2.5 Argument 3: The argument from degrees of alienation

The neuro-adaptations that occur as an agent declines into chronic addiction highlight one further argument against the hard view: Alienation comes in degrees. Not only is this a consequence of the fact that proper functioning comes in degrees, but it is also consistent with all three solutions to the subjective categorisation problem of section 1. This is most obviously true for experience-grounded subjective categorisation, as phenomenology exists on a spectrum rather than being all-or-nothing, but is also the case in judgements of rationality or coherence with narrative identity:

‘Some narrative threads become more central or deeply internalised in the overall self-narrative than others. The centrality of a feature in a self-narrative is a function of how many narrative threads include that feature and how many other narrative threads assume the truth of that first set of threads…..The self-narrative structure, therefore, operates as a kind of continuum of mineness – the central threads being at the more internal end of
the continuum and the peripheral threads being at the external end.’

(McConnell and Golova, 2003, p 72-73)

Shifting from a dichotomy to a continuum is a problem for any attempt to carve nature at the joints, as it implies that there exist phenomena that are neither one side nor the other of the ontological cut. Philosophers must resist ‘attempting to hack off parts like a clumsy butcher’ (Plato, translated by Hackford, 1952, 265e). If agents and non-agents are categorically distinct, then to which entity does a part-alien motive belong? Different individuals, or societies at different times may set a “threshold-for-alienation” beyond which a motive is called alien, but this (malleable) threshold is not a fact about the world.

Even in the case of unhealthy addiction, there are potential social and coping benefits to continued use. At the extreme these explain the case of the willing addict, but it is likely that most cases of addiction fall somewhere in between. Indeed, Holton and Berridge’s 2013 paper is called ‘Addiction Between Compulsion and Choice’.

Whilst the intuition of a human machine which is sometimes controlled by an agential motive and sometimes by an alien motive might be intuitively appealing, it is compromised, perhaps irreparably, by an acknowledgement that there is an entire spectrum of part-alien motives.

Neither philosophical analysis of attitude formation processes, nor contemporary models of addiction, nor degrees of alienation, support the idea that agential and alien motives are of an ontologically different natural kind. However, the findings thus far are consistent with the softer idea of a struggle to be authentic. I will therefore

---

95 Pickard (2021) and Sripada (2022) argue that addiction is not always irrational.
reject the hard view of alienation and turn to the arguments for and against, and the implications of, the soft view.

3. The soft view of alienation

The analysis of section 2 suggested that two things can be simultaneously true:

   a) All motives belong to the same natural-psychological kind. There is no hard ontological boundary between the two.
   b) Some motives partially or fully break the norms of agency.

This is possible because all motives are produced by conative architecture and conative architecture is fallible. Philosophers of agency are therefore left with two alternatives:

The first is to accept that harmful, inauthentic motives are a part of agency. Agency can be messy, as Walt Whitman described in ‘Song of Myself’ (1892):

   Do I contradict myself?
   Very well then I contradict myself,
   (I am large, I contain multitudes.)

Personally, I am sympathetic to this view.

Alternatively, one can argue that, although all motives are formed and interact equally within cognitive architecture, a metacognitive agent can reject some motivated action as not her action. She can do so because she has a sophisticated, but fallible, series of checks on the output of the motive formation process, and can

---

96 I will generally talk in terms of authenticity and norms of care, but readers may have different models based on rationality or other norms of agency
identify which (norm-breaking or partially norm-breaking) motives to reject, just as a worker on at the end of a production line can reject low quality products. However, to continue the metaphor, rejection is somewhat subjective and rejection without intervention has limited value. The production line worker must intervene to ensure that the low-quality product isn’t sent out to customers. Likewise, if an agent believes that her actions are irrational, inauthentic, metacognitively unwilled, or disliked, then merely labelling them as alien is not necessarily helpful. At times it may be harmful. What is helpful, is if categorisation triggers other motives towards control, whether the control is synchronic, diachronic, social or medical. Control may be extremely hard, but motives, alien or otherwise, are never uninhibitable.

In the first alternative, all intentional action is agential action, whether it is norm satisfying or not. In the second an action is agential only to the extent it is norm satisfying.

This raises the question of whether the difference between the two alternatives is terminological rather than philosophically substantive. My answer is to acknowledge that the difference may be philosophically substantive in some areas. Just as it is important for epistemologists to differentiate between beliefs that satisfy epistemic norms and those that do not, in some fields of philosophy it may be important to differentiate between motives that satisfy certain norms and those that do not. For example, there are very difficult questions about the allocation of moral responsibility that might hang on attributions of (moral) agency (see Sripada, 2015 for an overview); human flourishing is arguably dependent upon identifying and acting in accordance with an authentic self (Bauer, 2017, but see Feldman, 2014); and the aim of self-control or self-regulation, both synchronic and diachronic, is arguably to shape one’s actions, and indeed one’s attitudes, to more closely match norms of
ideal behaviour, somehow defined. The ability to engage in self-regulatory ‘mind-shaping’ (Vierkant, 2022, p102, see also McGeer and Pettit, 2002), is perhaps the greatest differentiating factor between human agency and that of animals. Finally, from a psychological health perspective, whether a patient identifies with a motive or disclaims it as alien may have a significant impact. I will leave it those with medical experience to determine whether, and in what circumstances, that impact is positive or negative.

However, there are also reasons to be wary of a focus on individual motives as opposed to overall capacities. Human agents have the capability to act morally and/or authentically and to engage in mind-shaping, but that doesn’t mean they are only an agent when they exercise that capability. Whilst metacognition, moral understanding and/or the ability to sustain a narrative self may capture what is incremental about human agency this doesn’t necessarily imply that a human agent is only that upper structure. Human agents are the full block of flats, not just the penthouse. Even in the case of the unwilling addict, he may be less responsible for his actions because chronic use has reduced his overall capacity for self-regulation, rather than because his craving was alien.

My view is that agency is broad and messy and inclusive. Inauthentic action is part of who I am, even though I strive to be authentic. Therefore, I will endorse what Schlosser, in his entry to the Stanford Encyclopedia, calls the standard conception of agency: ‘the exercise of agency consists in the performance of intentional actions’ (2019) 97.

---

97 The quote goes on to include unintentional actions to the extent that they ‘derive from the performance of intentional actions’. See chapter 2.
Even if I were to define agential action as necessarily authentic or norm-satisfying, that would not change the results of the functional and phenomenological analysis of this thesis. Whether an action is intentional would not change, and conative phenomenology would not change. As my aim is to close the gap between function and phenomenology, and explain why agency feels the way it does, it seems pragmatic to take an inclusive rather than incremental view of agency, and capture all motives rather than narrowly focusing on those that satisfy certain norms. If I had taken the latter approach then I would have had to ignore the experience of conative conflict between two motives, if both were defined as alien. Readers who favour a more restrictive view of agency due to commitments elsewhere, will hopefully still find value in the analysis, even if they sometimes disagree with the terminology.

I will therefore proceed to part III, reassured that, whilst not everyone will accept my Whitmanesque view of agency, there is nothing about alienation which undermines the core findings of the thesis thus far.
Part III

Part III consists of three chapters, focused on the decision-making process. By the end of this section, I will have a) completed my project of developing an experience-compatible model of the mind; b) shown that the phenomenology of the decision-making process does not justify libertarian beliefs; and c) solved the agency problem of compatibilism. Making a difficult decision is an action.

In part I and II, I developed a unified, hierarchical model of decision-making, based on the idea that when an agent makes a choice she does so based upon the strength of her motives.

*When an agent chooses, she will always choose to do what she is most motivated to do (The motivational strength idea)*

However, the introductory clause of this statement indicates that an agent is not compelled to decide immediately. In difficult or high stakes decisions she may withhold her choice and engage in further deliberation. For as long as deliberation proceeds, all considered action-plans remain inhibited\(^98\).

In part III I will focus on this pre-decision phase of the decision making-process. Doing so will involve a shift in the subject of my analysis, from motivational strength and conative phenomenology to the risk-of-error and feelings of confidence. I will argue that:

\(^{98}\) In chapter 2 I called this deliberation-driven inhibition.
An agent makes a choice only when she reaches a threshold\textsuperscript{99} of confidence in her decision (Confidence)

In chapter 7 I will conduct a detailed conceptual analysis of uncertainty, levels of confidence and decision-thresholds. This will be complimented in chapter 8 with an analysis of cognitive phenomenology. In chapter 9 I will incorporate the findings of Part I and II in order to explain how a difficult decision can be an effortful act. The analysis will suggest that libertarian beliefs are not supported by cognitive or conative phenomenology.

\textsuperscript{99} The threshold is not fixed. Understanding this fact is the key to understanding active decisions. See chapter 9.
Chapter 7
Levels of Confidence and Decision Thresholds

Introduction

When an agent faces a difficult decision, she is not forced to decide right away. If she is uncertain about what to do, she can engage in extended deliberation.

Decisions are made based on motivational strength, which is a non-conceptual representation of the conative value that an agent’s cognitive infrastructure assigns to a course of action. But conative values are predictions and, like any prediction, they incorporate a risk-of-error. Human decision-making architecture motivates deliberation to reduce that risk. When an agent withholds a decision, it is not because her conative values for and against an option are incommensurable, but because the difference between the pro- and con- conative values is small in the context of the possible impact of future evidence.

In section one I define what I mean by risk, evidence and extended deliberation, and introduce the idea of a probability distribution of possible futures. I propose that confidence is inversely related to risk-of-error: an agent’s confidence in her decision rises as the risk-of-error declines. In section two I analyse the concept of uncertainty, and conclude that being uncertain is not an all-or-nothing variable within the agent’s psychological architecture. The psychological variable is the agent’s level of confidence, which is continuous and “accumulates” as the risk of error declines. An
agent is uncertain about what to do until this accumulating level of confidence crosses a decision-threshold.

The concepts of uncertainty and levels of confidence are important for the thesis, because only by understanding the benefits of increases in confidence can one explain the introductory clause of the motivational strength idea and the existence of extended deliberation within the decision-making process. Without extended deliberation there would be no possibility of managerial actions such as acts of intentional synchronic self-control, or the act of making a decision.

In addition, feelings of uncertainty about what to do seem to support the libertarian intuition that agents ‘could be doing something else right here and now, that is all other conditions remaining the same’ (Searle\textsuperscript{100}, 1984, p. 95). To argue against that position, I will need to conduct fine grained functional and phenomenological analysis of the process of making a decision. Only by showing that feelings of uncertainty are feelings of low confidence, and that agents are motivated but not compelled to increase confidence, can I show that cognitive phenomenology does not support the libertarian intuition. Rather than indicating an agential power to ‘rise above the desires to which one is subject, and to take the reins in one’s own hands’ (Wallace, 1999, p234), feelings of a lack of confidence merely indicate an adaptive “hunger for information”.

1. Risk, Evidence and Extended Deliberation

1.1 Conative values, probability distributions and risk-of-error

\textsuperscript{100} Searle is not a libertarian, but he explains the libertarian intuition in order to illustrate the challenge that naturalists face. See chapter 8 for a full analysis.
Conative values are modelled estimates of the extent to which a plan-of-action satisfies care-based norms. In chapter 1, I suggested that each plan-of-action has a context-specific, norm-based, “care value”, and the aim of conative architecture is to predict that value, and motivate value accumulation. In chapter 6 I introduced the idea of model fallibility: Motives may fail to satisfy agential-norms.

If an agent makes a choice based on her current conative values, but would have chosen otherwise if her conative values had accurately represented care value, then I will call that choice an error. If the water looks inviting, I will be motivated to jump in. If it is really, really, cold, jumping in will have been an error. I am not Wim Hof! Whilst epistemic risk is the risk of believing in error (Pritchard, 2016), here I am concerned about the practical risk of intending in error.

There are two broad explanations for a divergence between conative value and care value: First, the agent’s conative infrastructure may not successfully capture agential norms. This “model failure” may be broad-based and chronic, perhaps due to illness or drug use, or narrow and temporary, if the agent is in a highly unusual environment. Second, the information set on which the conative evaluation is based may be flawed or unrepresentative of reality. Garbage in, garbage out. This second form of divergence is not due to model error but input error.

As analysed in chapter 6, there are complex questions about the extent to which agency is impacted by model failure. Similar questions could possibly arise with respect to the second source of divergence, if, for example, an agent is somehow fed biased information. However, in this chapter I want to identify what drives a normally functioning agent to stop or prolong deliberation. To do so, I will consider an idealised situation where the agent has no concerns about the functioning of her
conative architecture, and believes that relevant information is available even if she has not yet accessed it. In what follows “risk-of-error” should be understood to indicate solely the risk that arises from a partial data set. I set aside “risk-of-malfunction”.

In a situation where the only risk is from an information deficit, a plan-of-action’s care value is the asymptote to which conative value approaches as information is incorporated (figure 3). The future, perfect-information model-output value “stabilises” at the asymptote.

Figure 3. A stylised representation of the reduction in risk-of-error, and stabilisation of conative value, as unbiased information is incorporated.

Whenever an agent makes a choice, she runs a risk-of-error. Due to cognitive and temporal constraints, she will never have perfect information, but if she perceives the risk-of-error is high she can do something about it. She can deliberate further and reduce that risk.

But how does an agent’s psychological infrastructure estimate her risk-of-error? This is where the simplifying assumption becomes important. She no longer needs a separate model that somehow captures the “true” care value. The “true” value is now
equal to her stabilised future conative value. From her current data set, she can estimate the potential impact of further information, creating a probability distribution of future conative values. Her predicted risk-of-error is equal to the proportion of possible future conative evaluations which imply she should choose otherwise.

This may have been a bit quick so I will now go through an example in some detail. The framework I introduce will be central to my argument (chapters 8 and 9) that cognitive phenomenology does not support libertarian intuitions, so it is important that I ensure it is precisely specified:

Consider the case of Stephen. He has received an offer from two universities α and β, and is trying to decide which to attend. He is leaning towards university α, but he is still uncertain and hasn’t yet decided. He is going to read some more reviews and speak to his friend’s older siblings, because he really doesn’t want to make the wrong choice.

Stephen is mathematically minded so he represents the current state of his analysis in terms of the probability distribution of possible future worlds \(^{101}\) shown in figure 4. The mean of the distribution indicates the current state of his analysis. Right now, he is more motivated to go to university α. The spread of the distribution represents his estimate of the possible impacts of future evidence \(^{102}\). A wide distribution, with high variance, would indicate that he has just started his research and/or the information so far is mixed. A narrow distribution would indicate that he has reviewed a lot of information and it is all consistent. The distribution is symmetrical because he has no idea which option future evidence will favour.

---

\(^{101}\) I discuss Dennett’s possible worlds framework in chapter 8.

\(^{102}\) I assume that Stephen can model his potential future motivational balance, either because he is very self-aware, or because he has a strong antecedent motivation to attend whichever university the analysis favours.
Figure 4. A distribution of possible final conative values, assuming further deliberation. The mean reflects the current motivational balance and the variance indicates the potential impact of further information. The area to the left of the origin indicates the risk-of-error if the agent were to chooses university α now.

As can be seen, there is a non-negligible area of the distribution (6.7%) to the left of the origin. This implies that Stephen believes that it is possible, albeit unlikely, that new evidence will indicate that choosing university α is an error. This makes him ‘epistemically anxious’ (Newton, 2022), which is why he is planning to deliberate further. His current conative value differential is consistent with choosing α but it is possible that, due to an evidence-driven update, his future conative value differential will be consistent with choosing β. Although finding a piece of explosive evidence that inverts Stephen’s motivational balance would represent a highly unlikely and
somewhat shocking outcome, to some extent it is the justification for further deliberation. If such an update were impossible then further deliberation would be a waste of resources.

Stephen does not know what evidence he will find, but he knows that basing a decision on a large data set is more epistemically virtuous than on a small data set. His central expectation is that further deliberation will leave his choice unchanged but reduce his risk-of-error. Figure 5 represents an update after further deliberation in the case where the initial data set was representative of the full data set. His choice will remain the same but his confidence in the choice will increase, because there is now a lower risk-of-error. As Babic stated with respect to epistemic risk, ‘one credence function is riskier than another if it is a mean-preserving spread of it’ (Babic, 2019, p522). The orange updated distribution is less risky because it preserves the mean but reduces the spread.

---

There are significant overlaps with this analysis and the idea of “power” in statistical analysis.
Figure 5. An illustration of a reduction in error due to deliberation, in the case where the initial data set was representative of the full data set. The blue and orange lines represent the prior and posterior distributions respectively.

Stephen will only feel satisfied when the distribution suggests that the risk of error is negligible. At that point, consistent with the orange distribution in figure 5, his deliberating will end and he will make a choice. He will have one of ‘those comings-to-intend events that resolve a state of uncertainty over what to do’ (O’Shaughnessy, 1980, p297). This reveals that the question of when to decide is not directly a question about one’s current motivational balance, but about one’s estimated risk-of-error. The current conative value differential is relevant only to the extent that it impacts the risk-of-error. Figure 5 shows that an agent’s psychological state can shift from “don’t decide yet” to “decide now” if deliberation reduces the variance of the
distribution, *even if it does not impact the mean*. This is because, when making a winner-takes-all choice between two options, there is no value in further deliberation if there is no risk-of-error. If an agent is fully confident that one option is better than the other then there is no value in finessing exactly how much better. She should decide and reallocate her cognitive resources elsewhere.

Readers may agree that this is how Stephen makes his important decision, working with a checklist and a spreadsheet, but wonder about the relevance to normal, everyday decisions. Am I confusing external tools, and AlphaGo computations\textsuperscript{104}, with psychological architecture? There is something in this challenge. It is true that I felt no introspectable 6.7% risk-of-error when I was deliberating between tea and coffee this morning. But I did feel a certain level of confidence in my decision. My proposal is that the feeling of confidence, or lack thereof, is a non-conceptual representation of (an approximation of) the risk-of-error. Even if I can’t put a precise number on the risk-of-error, I can introspect the difference between high and low confidence.

Although computational neuroscientists continue to refine models of the decision-making process (Forstmann et al 2016, Rollwage et al 2020, Parés-Pujolràs et al. 2021) and many details are still being debated\textsuperscript{105}, there are a number of reasons to think that my framework is a fair approximation of how the brain works:

First, Bayesian inference relies on the idea of prior and posterior probability distributions. The impact of a piece of evidence on a hypothesis is a function of its

---

\textsuperscript{104} In a DeepMind blog describing AlphaGo’s model, the authors explained that AlphaGo favoured a scenario “where it will win by 1 and a half points with 99 percent probability over a situation where it will win by 20 points with 80 percent probability” (quote cited in Coppey, 2018). It minimised risk-of-error rather than maximising margin of victory.

\textsuperscript{105} I add my perspective to one aspect of that debate in section 2.3.
likelihood given the hypothesis. I will not defend the idea of the Bayesian brain here (von Helmholtz, 1866, Friston, 2012), but for anything like Bayesian updating of mental states to be possible, cognitive infrastructure must be able to model or approximate a probability distribution. Second, as will be seen in section 2, theories in areas as diverse as epistemology, computational neuroscience and predictive processing, suggest that psychological infrastructure instantiates at least a rough version of the model I have outlined. Third, in chapter 8 I will analyse the cognitive phenomenology of dynamically shifting levels of confidence. This provides phenomenological support for the idea that psychological infrastructure tracks some measure of the risk-of-error, and that this measure is central to the question of when to decide. And finally, it is an empirical fact that humans do prolong deliberation when they are uncertain, and decide quickly when they are confident (Dunlosky and Hetzog, 1998). The interesting question in philosophy of mind and cognitive neuroscience (Vierkant, 2022, Parés-Pujolràs et al. 2021) is not whether resolving uncertainty leads to resolving indecision, but whether it is possible to actively resolve indecision without resolving uncertainty. That question will be the focus of chapter 9.

In the introductory chapters of the thesis, I stated that I planned to develop a model of the mind that had two psychological elements at its heart: Motives and levels of confidence. Both have a central functional role in decision making, and both are felt. This was not particularly controversial at the time and it should not be controversial now. What I am trying to do here is to get a handle what is meant by confidence. My proposal is that confidence is inversely related to risk-of-error: an agent’s confidence in her decision rises as the risk-of-error declines. As Kriegel states, an agent ‘harbors two mental states: a judgement with content <p> and a confidence state
with the content '<my judgement that p is likely true>' (2015, p67). My claim is merely that p is more likely to be true when the risk-of-error is minimised.

Figure 6. A chart illustrating how an agent’s current motivational balance and level of confidence can be read off a distribution of possible future motivational balances.

Figure 6 illustrates the interrelation between motives and levels of confidence for a practical judgement. When an agent makes a choice, she does so on the basis of the motivational strength idea, but she will only choose when her level of confidence is above a certain threshold\(^{106}\). The aim of deliberation is to narrow the distribution and increase confidence until the threshold is met, just as the aim of eating is to

\(^{106}\) I analyse thresholds in detail in section 2.
decrease hunger until satiated. This is the model that I will be interrogating, supporting, and applying for the final three chapters of the thesis.

But first I will explain what I mean by evidence and extended deliberation.

1.2 Evidence and Extended Deliberation

Deliberation, if successful, improves an agent's model of the world, enhances decision-making and reduces the risk of error. It involves identifying, gathering, and incorporating relevant evidence within the decision-making process. It facilitates information gain and closes the information gap.

I have represented an agent’s current mental state with respect to a decision as a probability distribution. Anything that leads to an update in that distribution - its mean, its variance, or its shape - I will call evidence or information. This is a broad definition, including data which the agent deems irrelevant, or of which she is consciously unaware. It does not confine evidence or information to consciously recognised or endorsed reasons. Nor does it require that an agent is able to understand or justify her own decision. She does not need to be able to answer the why question. Nurses and fireground commanders (Kahneman and Klein, 2009) and radiologists (Drew et al, 2013) have been shown to make complex decisions quickly and intuitively without knowing how they do so. Some may be able to provide a post-hoc justification, but to me whether they can or not is unimportant. If it leads to an update it counts as evidence and that’s the end of it.

An agent can feel a growing sense of confidence in a decision despite not knowing what evidence (or analysis of evidence) is driving the increase. Even if she cannot answer the why question, she can still know that she is not at all uncertain about
what to do. Feelings of confidence therefore act as a bridge from the sub-personal to the personal, allowing an agent to be aware of (changes in) her current psychological state even if she cannot explain what drove those changes.

Identifying, gathering, and incorporating information in order to update one’s psychological state can involve a number of different epistemic processes and actions. Epistemic processes such as evaluation and subconscious pattern recognition are generally considered to be non-intentional, whereas epistemic actions, which can be mental (focusing attention) or physical (going to the library), are usually considered to be intentional. Physical epistemic actions such as gathering information through exploration, have been described as ‘world-to-mind directed processes aimed at bringing about mind-to-world directed states’ (Wilutzky, 2015, p6). Although an epistemic action may lead to a change in attitude, that does not imply that doxastic voluntarism is true. The action changes the mental state indirectly, by uncovering relevant evidence. This indirect action is “of the catalytic sort”, and “the rest is just waiting, seeing if anything happens” (Strawson, 2003). It is the evidence that directly changes the belief or intention, not the epistemic action.

For this chapter, I am looking for a term that captures the full, temporally extended, bundle of epistemically driven processes and actions that drive information gain and uncertainty reduction. In the absence of a better alternative, I will call this bundle “extended deliberation”.

2. Uncertainty, Levels of Confidence and Thresholds

In this section I aim to set out as precisely as possible the interaction between uncertainty, confidence, and decision thresholds, drawing on three diverse fields:
Epistemology, decision neuroscience, and predictive processing. The main finding will be that being uncertain is not an “all-or-nothing” variable within the agent’s psychological architecture. The variable is the degree of likelihood/confidence and an agent is uncertain if that variable is between thresholds.

2.1 Credal vs practical uncertainty

I will call uncertainty about the truth-status of a belief, credal uncertainty. An agent is credally uncertain if she doesn’t know whether a proposition is true. She is stuck between acceptance as true and incredulity. Practical uncertainty, on the other hand, is uncertainty about what to do, and is relevant to the formation of intentions.

As analysed in chapter 1, beliefs represent-as-true propositions about the world. If an agent is uncertain about whether to believe proposition $p$, it is because her set of evidence and other beliefs don’t decisively come down on the question of whether $p$ is true. She withholds belief in $p$ because there is a risk that $p$ is not true. Similarly, intentions represent-as-valuable a plan-of-action, aiming ultimately towards self-care. The value-representations underlying conative attitudes are predictions and all predictions are subject to risk-of-error. If an agent is uncertain about whether to choose action $A$, she withholds forming the intention to $A$ because there is a risk that $A$-ing would be an error.

Beliefs aim to satisfy a norm of truth, whilst intentions aim to satisfy a norm of care. Uncertainty in both cases create a risk of norm satisfaction failure\(^\text{107}\), a risk that can be mitigated by further deliberation and information gathering. Importantly, for both

\(^{107}\) This is in addition to the risk discussed in chapter 6 on alienation
belief formation and intention formation the resolution of uncertainty is sufficient for the acquisition of an attitude. If an agent is certain about the truth of \( p \) then she will believe that \( p \), and if an agent is certain about what to do, then she will (intend to) do it\(^{108}\). Although credal uncertainty and practical uncertainty differ in the attitudes to which they refer, they share many similarities. Each of the following five facts apply equally to credal and practical uncertainty:

First, the process of reducing uncertainty in both cases is epistemic, in the broad sense that it has a mind-to-world direction of fit. Through reducing uncertainty an agent acquires a propositional attitude (a belief or an intention).

Second, there is a symbiotic relationship between uncertainty and questioning. An agent can be uncertain about what is true, or what to do, only in the context of an open question about what is true, or what to do? Whilst uncertainty is a reason to refrain from, or delay, answering a question, without uncertainty there wouldn’t be a question. What is (subjectively) certain is accepted without question, and uncertainty does not exist within the psychology of unquestioning agents.

Third, the presence of uncertainty indicates an information gap. There is information that, if it were known by the agent, would help reduce her uncertainty. The fact that there are sometimes unknowable unknowns does not undermine the claim that information gaps exist. Rather, it reflects the fact that information gaps are not always closable.

Fourth, and notwithstanding the existence of unknowable unknowns, there is often something that an agent can do (consciously or subconsciously) to reduce

\(^{108}\) This should not be understood to prohibit akratic behaviour. It is possible for an agent to be theoretically certain about what she should do, and yet do otherwise, even if it is impossible for an agent to be practically certain about what to do and do otherwise.
uncertainty. She can identify, gather, and incorporate relevant available information to close the information gap. Indeed, as will be discussed in section 2.3, her cognitive architecture motivates her to do so. If, through deliberation, the information gap is closed and uncertainty resolved, then the question is answered. This is a function of the relationship between uncertainty and questions.

Finally, both credal and practical uncertainty are felt by human agents. I will analyse the phenomenology(s) of uncertainty in depth, in chapter 8.

Although my focus will be mainly on practical uncertainty, I will use examples and theories from the literature on credal uncertainty when licensed by the above similarities.

One example of that, which I turn to now, is in the fact that uncertainty—about what to believe (what to do) is the result of an interaction between degrees of belief (levels of confidence) and a norm-based threshold.

2.2 Uncertainty, degrees of belief and thresholds.

Within epistemology it is generally accepted that beliefs or credences can come in degrees (Cohen, 1989, Engel, 1998, Moss, 2018). For example, Muriel may have a 0.60 credence that her daughter has started smoking. This fact can be accommodated within an attitude-to-content framework either through complex attitudes to simple content or through simple attitudes to complex content. The complex attitude account proposes an array of belief-like attitudes such that, for example, Muriel “believes to the degree of 0.6” that “her daughter has started smoking”. In contrast, the complex content account proposes that Muriel “believes” that “it is likely to the degree of 0.60 that her daughter has started smoking”.

270
In her 2018 book, Moss argues that complex ‘probabilistic contents’ (p10)\(^{109}\) can be the contents of beliefs, assertions, and knowledge. In particular, beliefs with probabilistic content satisfy the four roles of belief: They play the appropriate role in explanations of action; they have the appropriate relation with the beliefs of other agents, thereby explaining agreements and disagreements; they correctly describe what happens when an agent “changes her mind”; and they appropriately ground relations of rationality and belief consistency. Although I agree with Moss that beliefs can have probabilistic contents, it is also uncontroversially true that agents face questions with simple contents. How should an agent directly respond to these questions? For example, how should Muriel respond to the questions of whether her daughter smokes, or whether she should ground her daughter for smoking? I will assume that a direct answer is one that applies to the simple content – “yes, my daughter smokes” – whereas a statistical answer shifts the content to probabilistic content – “my daughter is 60% likely to smoke”. I will also assume that it is often necessary, particularly in practical questions, to give a direct answer. The story of King Soloman and the dispute over the baby is an obvious example.

Moss’ proposal is that direct answers to questions with simple contents can be given by comparing probabilistic beliefs to contextually relevant norms. If the norm is that daughters can be grounded based on a 60% likelihood of smoking, then it is permissible for Muriel to ground her daughter. If the relevant threshold is 80%, then grounding is impermissible. A nice element of the theory is that the norm-based thresholds are action and context specific. Whilst an 80% threshold might be necessary for daughter-grounding, a 40% likelihood of rain will be sufficient to justify

\(^{109}\) Moss argues that the contents of beliefs are ‘probability spaces’ rather than propositions. However, this might just mean that ‘propositions are not the sort of objects that we thought they were’ (p15). For ease of language, I will continue to talk of propositions.
taking an umbrella to work. Although this is a compelling theoretical framework, in practice Muriel will often answer questions with simple contents by simply saying that she is uncertain. This fact must be accommodated and I believe that Moss’ account can easily do so. Explicitly making that adjustment clarifies two elements relevant to my definition of uncertainty:

First, with respect to directly answering questions such as “does your daughter smoke?” conversational norms seem to specify a tripartite system with not one but two thresholds. Above the higher threshold Muriel should answer yes and below the lower threshold she should answer no, but there is a “withholding zone” between the two where she should withhold judgement and state that she is uncertain. This seems uncontroversial and applies equally to practical questions. In the practical case, the agent should withhold action, (temporarily) inhibiting all alternatives until a decision threshold is reached. When an agent is uncertain, this is shorthand for her probabilistic content, when considered in the context of norm-based thresholds, being incompatible with a direct answer to a question with simple content. To be uncertain is to be between thresholds, in the withholding zone, unable to meet the normative requirements for providing a direct answer.

Second, Moss’ account assumes a given set of probabilistic knowledge, based on gathered evidence, which is represented within her cognitive architecture by an ‘epistemic representor’ (p181). This then interacts with the norm-based thresholds to drive, or withhold, decision-making. However, an epistemic representor is not immutable, it is evidence-responsive. If an agent has an opportunity to increase credence, then that opportunity should usually be taken in advance of making a practical decision. An agent’s utility may be better served by epistemic actions such as gathering information to update the epistemic representor, rather than practical
actions based on the current evidence set. Moss’ account explains what happens for a given evidence set, but must be augmented for agents that are able to prolong deliberation. Uncertain Muriel should gather more evidence before deciding whether to ground her daughter.

Although, I have focused on Moss’ probabilistic content model due to its detail and rigour, I am not tied to all aspects of her account. The relationship between likelihood-of and uncertainty-about can be transposed to more traditional attitude-based accounts. For example, Engel (1998, p 143-147) stated that whilst beliefs come in degrees, acceptance is an all or nothing matter. Importantly for Engel, it is acceptances that are used in reasoning. Acceptances are pragmatic and are premissable, in that they can be used as premisses in inference and action (Tuomela, 2000, p 125). As such, there needs to be a framework through which an agent extracts premissable acceptances from her epistemic landscape, which is full of degrees of belief. Presumably the framework would be very similar to that proposed by Moss: An agent accepts a proposition, and uses it for inferences and action, only if her epistemic landscape is consistent with the proposition meeting her context specific, norm-based thresholds for acceptance. If neither proposition $p$ nor $\neg p$ meets her norm-based thresholds then the agent is uncertain about whether to accept $p$. And if an answer to the question of whether $p$ is necessary for some practical reasoning then the agent should gather information, thereby updating her epistemic landscape, until she is able to accept or reject $p$.

In conclusion, uncertainty about the answer to a question is not simply a transformation of likelihood or confidence. It is the result of an interaction between likelihood/confidence and a context-specific norm-based threshold. Being uncertain is not an all-or-nothing variable within the agent’s psychological architecture. The
variable is the degree of likelihood/confidence and an agent is uncertain if that variable is between thresholds.

Somewhat surprisingly, this framework may help resolve a debate within computational neuroscience with respect to accumulation-to-bound models. Despite the broad acceptance and empirical success of these model there is not yet a consensus about what exactly is accumulating. In the next section, I introduce the models and suggest that what is accumulating is a measure of confidence.

2.3 Accumulation-to-bound models

In recent years, theories of decision-making within computational neuroscience have converged on a set of models which are broadly known as accumulation-to-bound models (Ratcliff and McKoon, 2008, Teodorescu and Usher, 2013, Forstmann et al 2016, Parés-Pujolràs et al. 2021). Roughly, these models understand decisions between options as a “race” between accumulating variables, where the process of accumulation is assumed to be stochastic, including statistical randomness as well as evidence-supported drift.
In a winner-takes-all contest, such as a decision between turning left and right, support for one option suppresses its opponent, so two enhancements to the simple race model have been proposed. I will call the models A, B and C, in line with figure 7. Model A is the simple race model. B and C accommodate lateral inhibition:

B) In drift diffusion models there is only one accumulating variable, which captures the relative preference of one option over another, but there are two thresholds, which equate to the different choices. If the upper threshold is hit then one decision outcome (right) is implemented and if the lower threshold is hit then the opposing outcome (left) is implemented.

C) In leaky competing accumulator (LCA) models, two variables race, but inhibitory competition between the alternatives is explicitly modelled. Any evidence which supports the accumulation of one variable, suppresses its opposition. This ensures an inverse correlation between the variables such
that there is no chance of both options being chosen at once. When one option (right) hits the threshold and is chosen, the other option (left) will always be at “zero”,

The three models are shown in figure 7. Whichever one is used, when an accumulating variable hits a threshold the race is over and a decision is made in favour of the outcome associated with that variable or that threshold. Dan Burnston (in submission) has argued that a decision can therefore be defined in terms of threshold-crossing:

A decision = The crossing of a threshold in a neurally implemented accumulation to bound process

The threshold in an accumulation-to-bound model is a decision-threshold.

Recently there has been some debate within the computational neuroscience literature about the precise definition of the variable that accumulates. Traditionally it has been assumed that what accumulates is “evidence” or the “true” value difference between the options\textsuperscript{110}. This position is summarized by Sripada: ‘I understand forming a decision as a process of evidence accumulation. When evidence for performing an action accumulates sufficiently so that it reaches a certain critical threshold, for example, when estimates of overall benefits exceed costs, or exceed them by a certain margin, a person-level action ensues: the person makes a decision to perform the action’ (2020, p 807). Within an agent’s cognitive architecture, the

\textsuperscript{110} ‘Though the evidence that this variable represents remains abstract in nature, there have been a variety of proposals as to what the evidence actually is: a function of the likelihood of each alternative response being correct, given the sampled information... a comparison between sampled information and a mental standard... a measure of strength of match between a memory probe and memory traces stored in long-term memory... or the difference in spike rate between pools of neurons representing the alternative options ’ (Lee, Daunizeau and Pezzula, 2023, p2).
estimate of the net benefit of a practical action is given by its conative value. At first glance, therefore, this approach seems consistent with the motivational strength idea that I have defended throughout the thesis. Perhaps micro motive-bits\(^{111}\) accumulate until they become a full-blown motive. On further reflection, however, two worries emerge, relating to (i) risk-of-error and (ii) the difference between decision timing and decision outcome:

The proposal that decisions are made when relative conative value estimates have accumulated to a threshold makes no explicit mention of risk-of-error. Sripada suggested that benefits need to exceed costs by a certain margin, but as shown in figure 8 below, a low margin in the context of low uncertainty (the green distribution) may indicate less risk of error than a high margin in the context of high uncertainty (the red distribution). If the required margin was “a conative value difference greater than 1.75”, then the agent would make a decision based on the risky red evidence set, but would continue deliberating based on the ~riskless green evidence set\(^{112}\).

\(^{111}\) Inspired by the fact that the technical name for a small iceberg is a “bergy bit” (https://oceanservice.noaa.gov/facts/iceberg.html)

\(^{112}\) ‘Under the DDM framework, higher variability in the evidence accumulation process causes the process to terminate sooner (on average; see Lee & Usher, 2021)’ (Lee, Daunizeau and Pezzula, 2023, p1362).
Figure 8. Riskiness vs conative value differences. The red distribution has a greater conative value difference than the green distribution, but a higher risk-of-error.

Ignoring the imprecision of the data set seems both irrational and maladaptive. A decision-making process that ignores the risk of error exposes an agent to harm, and this exposure is entirely unnecessary, because a) levels of variance are available to be computed from the gathered data set alongside the mean and b) prolonged deliberation can reduce that risk. This brings me to my second worry:

The crossing of the threshold is the moment that deliberation stops, so the appropriate considerations are those that bear on the costs and benefits of stopping deliberation. If there is a risk of error then that counts as a reason to continue deliberating, whereas no risk of error implies that there is no chance that an agent will change her mind, and therefore no value in further deliberation. This is the case irrespective of the magnitude of the central point estimate of benefits-casts. Further deliberation is not required to confirm that a precise value of 2.02 is greater than a precise value of 2.01. Of course, the (relative) costs and benefits of the practical
actions are indirectly relevant to the question of how long to deliberate, but only
insomuch as they bear on the costs and benefits of further deliberation\textsuperscript{113}.

One way of responding to the two worries is to propose that what accumulates in an
accumulation to bound model is not a point estimate of benefits minus costs, but
confidence. In a recent paper, Lee, Daunizeau and Pezzula suggest exactly that:
what is monitored and accumulates is the ‘momentary level of confidence across
time’:

\begin{quote}
The decision variable is quantitatively defined as a sigmoidal transformation of the absolute difference in posterior value estimates scaled by the posterior precision (i.e., the normative posterior probability that one option is better than the other, given the accumulated evidence thus far) (2023, p1363).
\end{quote}

Confidence under this definition, and consistent with the analysis of this chapter,
should be understood as an antonym of risk-of-error\textsuperscript{114}.

Although this shift in framing from evidence to confidence, from conative values to
risk of error, will be hugely important for my thesis, there are a number of reasons
that it is not particularly controversial:

First, for those who rightly think that decisions should be made in accordance with
the evidence, levels of confidence are entirely evidence-based. In fact, the reframing
suggests that the decision-making process incorporates the full information content

\textsuperscript{113} An agent will quickly decide in favour of an option with a much higher conative value, but only because the high differential implies that the risk-of-error is tiny.

\textsuperscript{114} To be precise, I have defined the risk of error as the area under the distribution to the left of the origin, whereas this new suggestion is (estimated difference)/(precision). For a given distribution, such as a normal distribution, one measure is merely a mathematical transformation of the other. However, if the shape of the distribution is allowed to vary, these two risk-based measures can diverge. Although I prefer my measure as it captures kurtosis, which definition is correct has no bearing on the proposal of this thesis. New research suggests that ‘optimistic’ and ‘pessimistic’ neurons map the entire complex distribution rather than just the mean plus an error term (Dabney et al, 2020) is perhaps supportive of my proposal.
of the gathered evidence, as captured by the mean plus the variance of the
distribution, rather than discarding adaptively important information about risk-of-
error, and using only the central expectation.

Second, the current suggestion is more directly linguistically intuitive, and
phenomenologically supported\textsuperscript{115}:

An agent’s likelihood of choosing increases with her confidence in her choice.

When she is confident\textsuperscript{116} in her choice she makes a decision.

When presented like this the reframing seems obviously true. Indeed, it is possible
that many supporters of the evidence-based interpretation have in mind a definition
that converges with the definition of confidence. This definition would imply that
measures of evidence are scaled by precision. If precision is excluded then, as
shown in figure 8, a decision with a high risk of error may have threshold-meeting
evidence\textsuperscript{117}.

Third, and perhaps surprisingly, support for the confidence-based interpretation
comes from the field of epistemology, as the confidence-based model is entirely
consistent with that discussed in section 2.2. Beliefs (intentions) are formed when
credence (confidence) hits a threshold, and credence (confidence) is a measure of
the likelihood that a proposition is true (a decision is best), not the amount by which it

\textsuperscript{115} It is important to note that accumulation-to-bound models are proposed to apply to sub-personal “micro-
decisions” as well as person-level value-based choice. Although it is important for me to be clear that not all
accumulation is felt, I don’t believe that this is a problem for the account.

\textsuperscript{116} The setting of the confidence threshold will be a central topic of chapter 9. I will argue that is not fixed but
can be actively adjusted by the agent. If that is true then a decision can be an act.

\textsuperscript{117} A second alternative is that the threshold is scaled by precision. Although functionally equivalent, this
solution cannot be made consistent with both the phenomenology of rising confidence detailed in chapter 8
and the ability to make a decision despite ongoing uncertainty detailed in chapter 9. In the current model the
phenomenology is a non-conceptual representation of the level of the accumulator, whereas in the second
alternative the phenomenology would have to be a representation of the relation between the accumulator
and the threshold. This would then fail to explain the phenomenology of a low-confidence decision. I therefore
reject this alternative.
is true (best). The same neural solution is implemented in both theoretical and practical decision-making. ‘Neuroanatomical continuity is taken as a guiding principle’ (Cisek, 2019, p2267).

Finally, from a historical perspective, a focus on confidence is actually a return to the original 1960s proposal that the accumulating variable, and response time, is a function of the likelihood of a response being correct (Edwards, 1965; Laming, 1968; Stone, 1960).

Although I have claimed that the reframing should not be controversial, it has significant implications. Most importantly, it implies that levels of confidence are an important element of the decision-making process. They are a key input into the metacognitive question of whether to deliberate further. Just as levels of hunger indicate the risk of starvation, levels of confidence indicate the risk-of-error. The purpose of evidence gathering is to reduce that risk, just as the purpose of foraging is to reduce starvation risk: ‘The action is performed to eliminate the conditions that motivated the action’. (Cisek, 2019, 2269). The reframing also suggests that decision timing and decision outcome are to some extent orthogonal. The interaction between levels of confidence and decision-thresholds relate to decision timing only. A third implication that I will not investigate is that ‘the ensuing decision process gracefully generalizes beyond two-alternative decisions’ (Lee, Daunizeau and Pezzula, 2023, p1364).

In summary, levels of confidence accumulate during ongoing deliberation, until they reach some (context specific, norm-given) threshold. They are the relevant input to the metacognitive question of whether to stop or prolonging deliberation. Whilst motives determine what an agent will choose to do, levels of confidence determine
when an agent will make the choice. Accumulation-to-bound models, when the accumulating variable is understood as a measure of confidence, explain the existence of the introductory clause in the motivational strength idea.

*When an agent chooses, she will always choose to do what she is most motivated to do (The motivational strength idea)*

*An agent will make a choice when she reaches a threshold of confidence in her decision (Confidence)*

Not only does this reframing seem functionally correct, it explains why an agent is motivated to allocate cognitive resources towards deliberation. An agent is willing to engage in costly information gathering actions not because she values evidence accumulation for its own sake, but because she values reducing the risk-of-error. For her, every shift higher in the accumulation of confidence is a shift lower in risk. Feelings of confidence allow her to monitor this risk, with high confidence ascribed a positive value within her conative architecture. An aversive feeling of a lack of confidence motivates the search for information, which is an adaptive response to the risks of uncertainty. Sophisticated agents monitor, and are motivated to increase, levels of confidence.

The psychological-level picture that is starting to emerge from the sub-personal analysis is one in which there is a motivational contest which answers the practical question of what to do, but then another at the metacognitive level which answers the question of how long to deliberate. Although accumulation to bound models are clearly sub-personal, they shine a light on when extended deliberation is necessary (when her confidence is below the decision-threshold) and what motivates extended deliberation (the value of reducing the error-based risk). The final piece of the
picture, the con- side of the metacognitive motivational contest, will be presented in chapter 9. A hint at the answer comes from a difference between my focus on risk-of-error and the focus within predictive processing on risk-of harm more broadly. I will analyse that difference now.

2.4 Uncertainty and risk-of-harm in the predictive processing literature

A less intuitive understanding of uncertainty can be found in the literature on predictive processing. I discuss predictive processing further in chapter 8 but roughly, the framework assumes that agents maintain homeostasis by attempting to minimise variational free energy. Variational free energy is a measure of the gap between the realized outcome and the outcome most consistent with the agent’s goal of sustaining herself. Agents minimise this sustainability gap by improving their internal model of the world (mind-to-world direction of fit) or by performing actions which bring the world closer to the model (world-to-mind).

Variational free energy is retrospective, so for an agent to use this framework to decide what to do prospectively she must have a generative model which computes expected free energy out to her planning horizon. She must estimate, and compare, the impacts of any potential future behavioural alternatives. She answers the question of what to do by (subconsciously) generating expected free energy scores for all considered alternatives, and choosing the option with the lowest score. Simply put, she minimises her subjectively estimated risk of harm. The question of what to do is automatically resolved by the expected free energy computation. If that is the case, how does it leave room for uncertainty?
In the predictive processing literature, the technical approach is usually to define uncertainty (“technical uncertainty”) as expected free energy, or expected prediction error: ‘Variational free energy approximates surprise and expected free energy approximates uncertainty’ (Friston et al, 2018, p388); ‘In some influential accounts… cognitive uncertainty [is] conceptualised as expected prediction error’ (Deterding et al, 2022, p7). Uncertainty ~ expected sustainability gap. Although this may be a powerful technical definition within the framework, it diverges from more traditional understandings of uncertainty in one major respect: It implies that pragmatic policies such as eating when hungry are uncertainty reducing because they reduce expected free energy. This follows from the fact that expected free energy is composed of (the negative of) two elements: Information gain (“epistemic value”) and pragmatic value. To minimise expected free energy is to maximise the sum of these two values. The technical conceptualisation of uncertainty as expected prediction error inverts the uncertainty-choice relationship. Rather than resolving uncertainty in order to decide what to do, an agent decides to do that which most resolves technical uncertainty.

The decomposition of expected free energy into epistemic and pragmatic elements naturally implies that uncertainty under this conceptualization can also be decomposed: Total technical uncertainty is the sum of epistemic uncertainty and pragmatic uncertainty. Pragmatic uncertainty, which is reduced by pursuing policies such as eating when hungry, might be better understood as pragmatic unsustainability or pragmatic risk-of-harm. Similarly, epistemic uncertainty could also be understood as epistemic unsustainability or epistemic risk-of-harm. Uncertainty reduction, pragmatic or epistemic is important for an agent only to the extent that it reduces the risk-of-harm.
As such, epistemic uncertainty in the predictive processing world is very close to what I have called practical uncertainty, but with a more explicit focus on harm. I have talked of the risk-of-error but within predictive processing, the important measure is risk-of-harm-from-error.

I agree with predictive processing advocates that for a model of human decision to be functionally accurate, it must incorporate both error and harm, but believe that for it to be phenomenologically accurate the two elements should be disaggregated. Levels of confidence are a measure of the risk of error, not the harmfulness of error. In chapter 9, I incorporate harm into my model of decision-making in the context of a metacognitive contest of motives. I will argue that whilst risk-of-error drives feelings of confidence, it is risk-of-harm-from-error that motivates information gathering behaviour. Through this hierarchical model, in which the motive to minimise harm from error competes with the motive to allocate cognitive resources elsewhere, I will be able to show that making a decision is an act, and explain why making a difficult decision is effortful.

This concludes my functional and conceptual analysis of uncertainty. I have argued that an agent is uncertain about what to do, when her confidence in an answer is below a norm-based, context-specific threshold. In the following chapter I will conduct a phenomenological analysis of feelings of confidence, and then in chapter 9, I will return to question of how the threshold is set.
Chapter 8
Cognitive Phenomenology: Levels of Confidence

Introduction

In this chapter I shift from the primarily functional analysis of chapter 8 to phenomenological analysis.

In section 1, I chart the shifting phenomenology that accompanies a successful process of deliberation, starting with uncertainty (or unconfidence), through the pleasurable feeling of rising confidence, and ending with the pleasurable feeling of trust and confidence.

In section 2, I consider the implications for two forms of libertarianism, anti-natural and psychological. I conclude that cognitive phenomenology does not support either libertarian belief.

In section 3, I briefly set out the adaptive benefits of cognitive phenomenology, responding to a challenge from Searle.

Before proceeding I should make clear that although I use the phrase cognitive phenomenology and derive support from Kriegel (2015), my focus is almost entirely on feelings of confidence and I remain neutral on questions of irreducibility. If forced to choose, I would say that the cognitive phenomenology of confidence is reducible to (metacognitive) algedonic phenomenology. This is consistent with my view that feelings of uncertainty are feelings of “hunger for information”.

286
Uncertainty or Unconfidence?

In chapter 7 I aimed to be precise about the interaction between uncertainty, confidence and risk of harm. I argued that uncertainty is not an all-or-nothing variable within the psychological architecture. The variable is the level of confidence and an agent is uncertain if that variable is between thresholds. What does that imply about feelings of uncertainty? I think that although an agent can be uncertain about what to do, what she feels is a level of confidence or risk-of-error. In this chapter I will therefore sometimes use the word unconfident, rather than uncertainty, to indicate a low level of confidence.

Personal level feelings of confidence

Although probability distributions and accumulation-to-bound models are sub-personal, feelings of confidence are experienced at the personal level. As an illustration, imagine that you are on a game show. There are two questions, and a prize if you get both answers right. You are allowed to consult a friend on only one question:

<table>
<thead>
<tr>
<th>Q1. What is the capital city of France?</th>
<th>Q2. What is the capital city of Switzerland?</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Bordeaux</td>
<td>2. Bern</td>
</tr>
<tr>
<td>3. Reims</td>
<td>3. Zurich</td>
</tr>
</tbody>
</table>

As question one is asked, you have a pleasurable feeling of knowing. You may even smile to yourself: ‘Mind at ease puts a smile on the face’ (Winkielman & Cacioppo,
2001). As the second is asked, however, you feel less confident. You search your memory for historical conversations or facts that might support one answer over another. You are familiar with Geneva and know that the UN headquarters are located there, but also know that Zurich is larger and is the financial centre. But there is something niggling you about Bern. Could it really be Bern? Or is Bern in Germany? If forced to guess you would choose Bern, but, given the prize at stake, it would be uncomfortable. The host asks you if you would like to consult a friend and, with relief, you request help with question 2.

This reveals a number of facts about confidence: First, levels of confidence in scenarios such as this are transparent to an agent. You know for which question you can state the answer with high confidence, and for which question you should call a friend\textsuperscript{118}. Second, there is often a strong link between confidence and willingness to act. In Q1, you are confident that the capital of France is Paris, so are keen to share that answer straight away. In Q2, your confidence is low, so you want to withhold your answer and engage in the epistemic action of phoning a friend. Third, there is something pleasurable about confidence and aversive about uncertainty. And finally, there is an inverse link between feelings of confidence and feelings of alternative possibilities. There is a sense of openness and ambiguity (James, 2014) about what answer you will give to Q2 – whatever answer your friend tells you to give - that doesn’t exist for Q1. Your psychological state with respect to Q1 is consistent with only one answer.

The language of the preceding paragraph suggests that phenomenology can help answer psychological level questions, but what exactly is the psychological level and

\textsuperscript{118} This does not mean that feelings of confidence are always reliable. An agent could be confident that Zurich is the capital of Switzerland if, for example, she knows it is not Geneva and has never heard of Bern.
how does it relate to experience? List describes the psychological level as ‘the level at which our best theories of psychology describe and explain human cognition and behaviour’, and gives the examples of ‘intention and action, belief and desire’.

Although List does not mention feelings of confidence and states that ‘the psychological level must not be confused with a person’s subjective first-personal perspective’, I will assume that feelings of confidence act as a window into the psychological level. If a person feels fully confident in a decision, then she is aware that her psychological state is consistent with only that decision. Alternatively, if she is uncertain during deliberation, then this makes her feel that her psychological state is consistent with more than one final decision. This will become highly relevant in section 2.

In what follows I analyse feelings of confidence throughout the decision-making process. I start with an initial feeling of low confidence, and chart its transformation (in the case of successful deliberation) into a final pleasurable feeling of trust and confidence in a decision. Although I won’t discuss accumulation to bound models in this chapter, the shifting experience of levels of confidence throughout a process of deliberation provide phenomenological support for the reframing of the accumulator in terms of confidence, as opposed to evidence or conative differentials. If what accumulates is a level of confidence (which it is) and the accumulating level of confidence is felt at the personal level (which it sometimes is119), then experiencing that accumulation gives us an insight into our own psychological state and allows us to answer psychological level questions.

---

119 I reiterate that accumulation-to-bound models are proposed to apply to sub-personal “micro-decisions” as well as person-level value-based choice. Not all accumulating processes are felt.
1. Cognitive phenomenology of levels of confidence

1.1 Feelings of low confidence

For sophisticated agents, levels of confidence are a key input into the decision-making process. They are monitored and are central to the evaluation of whether to stop or prolong deliberation.

In human agents at least, results of this metacognitive monitoring are revealed phenomenologically to the agent through what Dokic calls noetic feelings, where feelings are noetic if they ‘concern epistemic states, events, or skills’, (2012, p303, see also Proust 2013). Feelings of low confidence are generally aversive and ‘the curious individual is motivated to obtain the missing information to reduce or eliminate the feeling of deprivation’ (Loewenstein, 1994, 87). Therefore, the confidence-based negative feedback system is a metacognitive version of an organic circuit such as the hunger feedback loop. There, aversive feelings of hunger activate the desire to eat, whereas here the aversive feelings of low confidence activate the desire to gather information. In both cases ‘the action is performed to eliminate the conditions that motivated the action’. (Cisek, 2019, 2269). This efficient repurposing of a simple evolutionary solution to a high-order problem, Cisek calls ‘resynthesizing … through phylogenetic refinement’ (2019).

Although the primary purpose of metacognitive monitoring of levels of confidence is to inform the question of whether to stop or prolong deliberation, the fact that uncertainty is felt gives an agent a window into her own psychological state. It reveals to her that the reason her decision-making process is not yet concluded is

---

120 In both cases the aversive feeling creates a desire to respond rather than an uninhibitable reflex response. This gives the agent flexibility to respond to a greater need first, if they are more motivated to do so.
because there remains more than one possible future decision consistent with her current psychological level of understanding. The feeling of uncertainty during ongoing extended deliberation supports an agent’s belief that although she will eventually resolve the question of what to do, and choose one alternative or another, until that happens more than one final decision is consistent with her current psychological state. Taking inspiration from James, I will call this a belief in psychological openness (and ambiguity), defined as follows:

During ongoing deliberation an agent’s psychological state is consistent with more than one final decision.

Philosophers have linked the experience of alternative possibilities to the experience of freedom: ‘The experience of freedom, that is to say, the experience of alternative possibilities, is built into the very structure of conscious voluntary intentional human behaviour’ (Searle, 1984, p94). Here is a potential logical argument that leads from the premise of psychological openness to the experience of alternative possibilities:

1. During ongoing deliberation an agent’s psychological state is consistent with more than one final decision (Psychological openness)
2. An agent will do (or at least try to do) whatever she eventually decides to do (Intentional action)
3. Therefore, during ongoing deliberation an agent’s psychological state is consistent with alternative possible future actions (Psychologically possible alternatives)
4. Feelings of low confidence during ongoing deliberation make an agent aware of her own psychological openness (Experience of psychological openness)
5. Therefore, when experiencing low confidence during ongoing deliberation an agent is aware that her psychological state is consistent with alternative possible future actions (Experience of alternative possibilities)

In section 2, I will argue that neither the existence, nor the experience, of psychologically possible future alternatives in the early stages of deliberation justify libertarian beliefs. This is because psychological openness is contingent upon further deliberation. As soon as deliberation stops, psychological openness is false, even if the agent continues to feel low confidence in any decision.

The phenomenology of uncertainty in the early stage of deliberation has been covered widely in the libertarian vs naturalist debate but other confidence related experiences have received much less attention, despite being widely studied elsewhere. In what follows I will bring the findings from other areas of philosophy to bear on the debates within the philosophy of agency.

1.2 The phenomenology of an increase in confidence

A human agent can experience her level of confidence. When combined with her ability to remember previous experiences, this allows her to observe the progress of her confidence over time. However, the phenomenology of an increase in confidence that I analyse in this section is more than just a series that changes over time. It is the phenomenology of the rate of change (or first derivative) of confidence. A sharp rise in confidence is felt as pleasurable, whereas a steep fall is aversive.

As the experience of increasing confidence has not been previously analysed in the context of the phenomenology of agency, the following will draw on work in other philosophical fields, with focus on the aha! moment, games, flow, and most
relevantly, predictive processing. This analysis will not only serve to confirm that the feeling of increasing confidence exists, but also explain why.

The clearest example of a sudden and pleasurable increase in confidence is in an aha! or Eureka! moment, as commonly associated with Archimedes in his bath. Topolinski and Reber (2010) give four defining characteristics for an aha! experience: Suddenness; Ease; Positive Affect; and Truth and Confidence. Importantly positive affect and truth and confidence are felt intuitively, before the solution has been formerly assessed. I will return to feelings of truth and confidence in section 1.3.

The suddenness element of an aha! moment raises two possibilities for the cause of the positive affect: Either it is driven by the high rate of change or it is driven by the positive end state. Topolinski and Reber seem to imply that it is the change that is important: ‘The sudden change in processing fluency increases positive affect’ (p., my italics, p404). Arguably, the pleasurable experience arises from (fast) progress towards an end, rather than from reaching the end itself. Perhaps surprisingly, this shift in focus from, as it were, the destination to the journey, is corroborated by recent philosophical work on games.

Nguyen (2020), in his analysis of aesthetic striving games, addresses the question of why humans spend so much time and energy pursuing such arbitrary goals:

In the rest of life, we are used to justifying our goals by looking at the value of the goals themselves or by looking forward, to what follows from those goals.

\[^{121}\] Feelings of uncertainty or confidence in the psychology literature are usually linked to what is called processing fluency, which indicates coherence, contextual accessibility of applicable knowledge, and the frequency of concept collocations in the corpus of natural language (see Schwartz and Jemstedt, 2021)
But with the goals of games, we often need to look *backward*…. In ordinary practical life, we usually take the means for the sake of the ends. But in games, we can take up an end for the sake of the means. Playing games can be a *motivational inversion* of ordinary life (p1).

A game designer invents a goal, and creates obstacles to achieving that goal, building an environment through which a player can strive and make progress. The problem solving and obstacle overcoming, if calibrated at the right rate, can create an aesthetic experience. Nguyen calls this a motivational inversion because the goal and obstacles of the game are merely means to the end of aesthetic striving. He talks of the search for that ‘most delicious of harmonies – harmony of capacity’ (p111) where there is a fit between one’s maximum skill level and the demands of the task. This can be targeted in games by modulating the design, opposition, or difficulty level. The positive phenomenology that Nguyen details is an example of the well-documented experience of flow (Csikszentmihalyi,1988), for which there are two preconditions: First, the agent must have a sense that the challenges are at an appropriate level for her abilities. Second there must be immediate feedback about the progress that is being made. Nakamura and Csikszentmihalyi described flow as ‘intrinsically motivated or autotelic activity’ (2014, p 239). It is rewarding in and of itself independent of the end goal.

Nguyen’s motivational inversion and Csikszentmihalyi’s intrinsic motivation follow from the fact that a positive rate of progress generates positive affect. Agents want to experience more of the same, either in the moment via a continuation of the activity, or later by seeking out environments in which such an experience can be replicated. ‘Experiencing flow encourages a person to persist at and return to an activity because of the experiential rewards it promises’ (p249). An understanding of this fact
has been used (for good and for ill) by technology companies who encourage participation by “gamifying” various tasks.

Philosophers in the predictive processing field, have taken the positive phenomenology of uncertainty reduction\(^\text{122}\) seriously because it provides a way out of what has been called the *dark room problem* (Kiverstein et al., 2019, Van de Cruys, 2017). In summary, this problem arises from the core assumption of predictive processing, that agents maintain homeostasis\(^\text{123}\) by attempting to minimise prediction error, or what they call surprisal. Survival depends upon avoiding surprising encounters and physiological states that are uncharacteristic of a given phenotype (Friston 2009). To achieve this, agents generate models of expected prediction errors and are motivated to reduce those errors, either by improving their internal model of the world, or by performing actions which bring the world closer to the model\(^\text{124}\). Agents thus aim to reduce epistemic risk and pragmatic risk.

Although an attractive and far-reaching theory, the worry is that, absent qualification, predictive processing might imply that agents should seek a dark room and stay there, because that is where expected prediction error would be minimised. This would be inconsistent with the empirical facts: Agents engage in novelty seeking behaviour such as starting a crossword, exploring a new city, or listening to a good joke. The simple uncertainty-is-bad model needs to be augmented.

An obvious first response to the dark room worry is that part of what agents “expect” is their own sustained survival. Being in a dark room would quickly diverge from such

---

\(^{122}\) Uncertainty in predictive processing is understood as risk-of-harm broadly. Risk-of-harm-from-error is a subset. See chapter 7.

\(^{123}\) The definition of homeostasis can be broad. For example, it can include what Gillian and Tye call ‘social homeostasis’ (2019, p5)

\(^{124}\) ‘The models needed for this may be rudimentary and fixed or complex and flexible, depending on complexity and volatility of the organism’s Umwelt’ (Van de Cruys, 2017, p4)
an expectation, so even organisms subject to the simple predictive processing model would exit the dark room in search of food or warmth. However, this negative reply to the objection falls short of explaining the fact that even warm and satiated agents seem to positively seek out novel environments in which prediction errors will increase.

A positive solution to the dark room problem becomes possible once one realises that, in the model as described, the time horizon of the expectation has not been specified. Whilst an agent who is trying to minimise expectations of short-horizon prediction errors, may rush to a dark room, her cousin who is aiming to minimise expectations of long-horizon prediction errors might head out into the world to increase her understanding. A long-term agent is willing to accept short-term pain for long-term gain.

This solution authorises not unconstrained but bounded exploratory behaviour. The dark room is replaced by a “safe room” in which an agent lets out some rope, but not enough to hang herself. An agent is rational in increasing the distance to a goal (prediction error), if that moves her to a position of strength, from which she can approach the goal with larger certainty (Friston et al. 2015). Thus, it is rational for an agent to seek out prediction errors, but only to the extent that they are reducible. She should choose novelty only when novelty plus its expected resolution leads to a decrease in long-term expected prediction errors.

For an agent to be appropriately motivated, this theoretical explanation of what an agent should do must be implemented in her cognitive architecture, and this is where the positive phenomenology becomes relevant. Because problem-solving and prediction-error reduction is accompanied by positive affect, this becomes something
that agents seek to replicate. Agents are thus motivated towards epistemic affordances. They seek out environments in which they anticipate that, by allowing some rise in uncertainty in the short term, they will be rewarded with the positive phenomenology of uncertainty reduction. Think of the growing sense of expectation when listening to a joke and anticipating the joy of a great punchline. Equally, agents should and will try to avoid environments where uncertainty is autocorrelated, because this is an environment associated with aversive phenomenology.

The positive phenomenology of uncertainty reduction turns agents into what Andy Clark has called “slope chasers”125. And individual instances of slope chasing can, over time, be scaffolded into huge gains in understanding, skill acquisition and sophistication:

Once an agent can track and learn to expect certain rates of change in prediction errors, it arguably will show a distinct propensity to explore and learn. This continuous, active search for reducible prediction errors … may in turn have enabled the development of rich social relations and culture. (Van de Cruys, 2017, p11)

With respect to uncertainty, all’s well that ends well.

I have argued that the phenomenology of rising confidence is important for sophisticated agency, but why is it relevant for the discussions of this chapter? In short, because it is an experienced closing-down of alternative possibilities. In the early stages of extended deliberation, when uncertainty about what to do is high, an agent is aware that her own psychological state is consistent with more than one eventual decision. As she samples the environment, and reduces the risk-of-error

---

125 In conversation with Kiverstein et al, as reported in their paper (2019, p2865).
she feels her confidence rising. The phenomenological experience that accompanies successful deliberation reveals two things to the agent. First, she is narrowing her possible worlds towards the point when all worlds are consistent with only one decision. Second, she is accomplishing this narrowing of worlds through a natural process of evidence sampling and deliberation. Her increasing confidence is not due to luck, but follows from behaviour that is targeted towards information gain. She is taking advantage of epistemic affordances because they will reduce her risk of error. An agent can only be a slope chaser if she knows how to find a slope.

The positive affect of increasing confidence, which is powerful enough to be create Nguyen’s motivational inversion, adds a nuance to seemings-based justification for beliefs in alternative possibilities. Whilst an agent at the beginning of extended deliberation may experience openness and ambiguity, she should become aware, through the pleasurable experience of uncertainty reduction, that the process of extended deliberation closes down alternatives. Whether it narrows the possibilities to a singularity or not depends on how certain the agent feels at the moment of decision. I turn to that now.

1.3 The phenomenology of confidence in a decision

In an ideal decision-making process, uncertainty about what to do is resolved by the process of extended deliberation. And if uncertainty is resolved then, necessarily, indecision is resolved^{126}.

^{126} The question can be reopened, but only when uncertainty returns. See chapter 2 on intention stability.
The phenomenology of increasing confidence is pleasurable, and motivates slope chasing behaviour. However, it is possible that if uncertainty begins at a high level, it can fall without being fully resolved. The phenomenology of uncertainty reduction is not enough to guarantee uncertainty resolution. For that we must turn to another of Topolinski and Reber’s defining characteristics of an aha! experience, the feeling of truth and confidence. Poincaré gives the following example with regards mathematical transformations:

One day, as I was walking on the cliff, the idea came to me, again with the same characteristics of conciseness, suddenness, and immediate certainty, that arithmetical transformations of indefinite ternary quadratic forms are identical with those of non-Euclidian geometry (1914, p53/4).

This feeling is of an end state of confidence, rather than a rate of increasing confidence. It does not relate to the process of solving a problem but to the moment in which the solution becomes apparent. Indeed, in the example given by Poincaré, it happened suddenly when he was, consciously at least, taking a break and thinking about something else.

The fact that a flash of insight can come out of the blue doesn't mean that it is unrelated to cognitive processing. Rather, as Holton analyses in his account of choice, some elements of a process of extended deliberation can be unconscious:

Unfortunately, the feeling of confidence does not guarantee that a conclusion is epistemically or normatively right. The phenomenology indicates coherence with other beliefs rather than necessarily something true about mind-independent reality.

Picard (2013, p. 2496) reports on two patients experiencing a feeling of intense bliss during epileptic seizures originating in the insula. One patient describes the experience: "...all the ordinary facts about the environment seem suddenly to become infused with certainty and a sense of inevitability... The sense that I had when I was experiencing some of these seizures was not unlike a continuous series of profound “a-ha!” moments”

This is consistent with my definition of extended deliberation in chapter 7.
‘Very often [our decision-making process] will respond to features that we have registered but of which we are unaware’ (2006, p9). More specifically, agents have a cognitive capacity for recognising patterns. Individual regularities are learned through experience in the environment, and those regularities are encoded in the form of associations. When cued by incoming data, the encoded associations generate expectations based upon prior experience. These expectations can either be used prospectively to make predictions, or in the moment to spot expectation-incongruity.

An emotional response to a situation can influence action. Sometimes the agent will realise later why she chose to act the way she did, but at other times she may never know. The decision was driven by incorporating evidence about relevant aspects of the world and spotting patterns that matched experience, even if the mechanisms that performed this analysis were not conscious.

Kriegel, in his analysis of cognitive phenomenology, argues that there is something it is like to form a judgement. He lists sixteen of what he calls ‘phenomenological platitudes’ (2015, p65) which, taken together create a phenomenological Ramsey statement. Relevant to this analysis, the cognitive phenomenology of making a judgement that \( p \) involves a feeling: of being committed to the truth of \( p \); of being rationally compelled; of reasons-responsiveness; that \( p \) is sufficiently supported by evidence; and of involuntariness\(^{130}\). With respect to the final platitude of involuntariness, Kriegel states: ‘If I consider that the evidence supports \( \neg p \), I cannot make the judgement that \( p \)’(p68). Through cognitive phenomenology an agent experiences the fact that her set of possible worlds has narrowed and is now consistent with only one decision.

\(^{130}\) Kriegel gives 16 platitudes. Those referenced are 1, 10, 12, 13, 14 and 16.
Perhaps unsurprisingly, analysis of the feeling of confidence is a reminder of the discussion of the phenomenology of intellectual seemings in introductory chapter (i). Plantinga’s ‘phenomenology that goes with seeing that such a proposition is true’ (1993, p105-106), and Brogaard’s ‘understanding that is so intense that we have a hard time envisaging that others may not feel the same way’ (2014, p388-389) are descriptions of the cognitive phenomenology of full confidence. Somewhat ironically, intellectual seemings are reliant on the fact that their cognitive phenomenology indicates that there are no alternative possibilities. An agent feels rationally compelled to accept the truth of the proposition.

Whilst the phenomenology of low confidence justifies a belief in psychological openness at the early stages of the extended deliberative process, feelings of rising confidence and of truth and confidence, justify a belief in psychological determinism at the end of successful deliberation. If uncertainty about what to do is resolved, then so is indecision, and the agent’s cognitive phenomenology does not indicate to her that she could decide otherwise.

I develop this argument further in section 2.2 by introducing the synchronicity requirement: For phenomenology to justify a seemings-based claim, the phenomenology and the object of the claim must be synchronous.

2. The phenomenology of psychological openness and the libertarian intuition

Human agents make decisions and act in the world. When they face difficult decisions, they engage in extended deliberation and consider various options before finally coming down on one alternative over another. During deliberation it feels to an agent as though her decision is open. She has what James has called open and
ambiguous possibilities (1956). Even after deciding and acting, she may feel that she could have done otherwise.

I have argued that the risk-of-error underwrites the possibility of psychological openness, defined as follows:

During ongoing deliberation an agent’s psychological state is consistent with more than one final decision.

Not only is psychological openness, under that definition, true from a functional perspective, but its truth is revealed by cognitive phenomenology. The purpose of a feeling of low confidence may be to highlight the risk of error and encourage further deliberation, but a natural consequence of such a feeling is a person level awareness of the fact that such deliberation may serve to change the decision outcome.

Pre-theoretically this phenomenology might seem to support the (anti-naturalist) libertarian intuition that:

Agents have alternative possibilities. They can make a decision, ‘the making of which cannot be fully explained by the laws of nature and a prior state of the world.’ (Fischer\textsuperscript{131}, 1998, p165).

I now present three arguments that it does not. The first shows that psychological underdetermination is compatible with naturalism. Arguments 2 and 3 show that psychological openness does not support beliefs in psychological libertarianism, the claim that decision outcomes cannot always be explained by psychological factors

\textsuperscript{131} Fischer was not endorsing this view, but setting it out clearly.
(List, 2019). If cognitive phenomenology doesn’t support psychological libertarianism, then it cannot support anti-naturalist libertarianism.

2.1 Argument 1: Psychological openness reveals a limitation, not a power.

Here I use Dennett’s framework of nearby possible worlds, to show that psychological openness is caused by psychological underdetermination, which is consistent with naturalism.

In order to understand common usage of phrases such as alternative possibilities, or could have done otherwise, Dennett suggested drawing on the idea of nearby possible worlds (2003, Taylor and Dennett, 2005). For example, when considering the question of whether Austin (1961) could have made a putt that he, in fact, failed to make, Dennett argues that one should reframe the question as follows: Is there a possible world, within a defined set of near possible worlds, in which Austin sinking the put is consistent with the physical laws?

Although there is only one actual current world, an agent’s understanding and descriptions of that world will always be at a relatively coarse level of granularity. Granularity of description is modulated in accordance with the context, but an agent’s description will never be fine-grained enough to describe only one current world. Cognitive and temporal constraints mean that any agent’s understanding of the world can only ever be partial. It will always be consistent with multiple nearby current possible worlds. I will call this fact psychological underdetermination:

An agent’s understanding is never fine grained enough to describe just one unique world. An agent’s subjective set of possible worlds at any moment in
time, her set of psychologically possible worlds, is the set which contains all worlds consistent with her then understanding.

According to naturalism, each world has a single future, which is an evolution of the current world in accordance with natural laws. Therefore, if an agent’s subjective set of possible worlds includes multiple current worlds, then even under the assumption of naturalism, this subjective set is also consistent with multiple possible future worlds.

This framework offers an interpretation of the experience of alternative possibilities, that competes with the libertarian intuition. Rather than the experience justifying a belief in a super-natural agential power, inconsistent with naturalism, it merely reflects the agential cognitive limitations which cause psychological underdetermination.

Retrospectively, psychological underdetermination underwrites beliefs such as Austin’s that he could have made a put, even after it turned out that he did not. This is because, when Austin reviews the set of possible worlds consistent with his current understanding of the moment in which he attempted the put, many (or even most) of these worlds involve making the put. Of course, if he later found that he was carrying a shoulder injury he might update his view. His updated understanding would include reasons for believing that the put was much less likely than he had originally thought.\footnote{Although Dennett (2003) developed his nearby possible worlds account in opposition to the first half of a footnote in Austin’s chapter, the second half of the footnote reveals that Austin’s position was nuanced: “But if I tried my hardest, say, and missed, surely there must have been something that caused me to fail, that made me unable to succeed? So that I could not have holed it. Well, a modern belief in science, in there being an explanation of everything, may make us assent to this argument. But such a belief is not in line with the traditional beliefs enshrined in the word can: according to them, a human}
The common meaning of “could have” is differentiated from “did” precisely because it incorporates the fact of psychological underdetermination.

Prospectively, the fact of psychological underdetermination explains why it would seem to an agent that more than one future is open even if naturalism were true. If more than one current world is consistent with her psychological state then, naturally, more than one future world is also consistent with her psychological state.

Perhaps the most commonly experienced form of psychological underdetermination is when an agent is in the early stages of extended deliberation. Here it is not just that more than one future world is consistent with her current psychological state, but more than one future decision is consistent with that state. Her current set of nearby possible worlds is consistent with a set of future worlds which contain more than one decision. Dennett’s nearby possible worlds are the worlds captured in what I have previously called the agent’s probability distribution. If she is uncertain about whether to choose A or B then by definition her set of possible future worlds includes at least one world in which she chooses option A and at least one in which she chooses option B.

Psychological openness is compatible with naturalism and explained by psychological underdetermination.

Although I have shown that naturalism is compatible with psychologically openness, there is something worrying about the argument from psychological underdetermination. The worry is that if more than one future decision is consistent with her psychological state and yet she is the one that decides, then perhaps that ability or power or capacity is inherently liable not to produce success, on occasion, and that for no reason (or are bad luck and bad form sometimes reasons?). (1961, p166)
supports a belief that she can ‘rise above the desires to which one is subject, and to take the reins in one’s own hands’ (Wallace, 1999, p234). I will call such a belief psychological libertarianism, defined as follows:

An agent sometimes has alternative possibilities in the sense that:

a) Her decision outcome cannot be fully explained in psychological terms. Psychological factors are not a sufficient cause of decision outcome.

b) Her decision is not fully constrained or determined by psychological factors.

c) It is not true that when she makes a decision, only one decision outcome is possible for her.

Psychological libertarianism is a theory at the psychological level of description, and is compatible with naturalism, but there are three reasons why it should be taken seriously. First, some believe that for questions of free will, it is the psychological level that is important: ‘Free will is very much a higher-level phenomenon’ (List, 2019, p7). Personally, I am unconvinced of the veridicality of this claim but as I provide no argument against, I will assume that it may be true. Second, for anti-naturalist libertarianism to be justified by cognitive phenomenology, it is necessary (but not sufficient) that psychological libertarianism is justified by cognitive phenomenology. Third, psychological libertarianism is inconsistent with the motivational strength idea. If psychological libertarianism is true then that would defeat the central claim of this thesis.

If the phenomenology(s) does not support psychological libertarianism then it does not support anti-naturalist libertarianism. If I can show that libertarian intuitions, even at the psychological level, are not supported by the phenomenology(s) then, when
combined with argument 1 and empirical support for naturalism, I will have created a persuasive defeater for both forms of libertarianism.

Before presenting my argument against psychological libertarianism, I will first introduce what I will call the synchronicity requirement of seemings-based justification. Pre-theoretically, psychological underdetermination seems supportive of psychological libertarianism, but the synchronicity requirement undermines that support.

2.2 The synchronicity requirement

In introductory chapter (i), I analysed seemings-based justification. Here I argue against using experienced-based seemings from an early stage in the decision-making process to justify beliefs about a later event - the event of intention formation.

First, a (shortened) reminder of Brogaard’s position of sensible dogmatism:

If it seems to S as if \( p \) and the seeming is grounded in … S’s … experience, then, in the absence of defeaters, S thereby has at least some degree of justification for believing that \( p \) (2013, p279)

Let \( p(1) = \) it is snowing. S cannot justifiably base a claim that “it seems to me as if it is snowing” purely on an experience of it snowing in the (somewhat distant) past. That experience is relevant to the proposition “it was snowing” rather than the proposition “it is snowing”. To ground a belief that it is snowing on historical experience, S must also have justification for believing a second propagation proposition along the lines of “nothing has happened in the intervening period to change the truth-status of proposition \( p \)”. If it is snowing at time \( t_1 \) and nothing has
happened to change the weather between \( t_1 \) and \( t_2 \) then (maybe) \( S \) is justified in believing that it is snowing at time \( t_2 \). For some examples of \( p \) such a propagation proposition might be justified, but consider, in the context of ongoing deliberation, the proposition \( p(2) = \text{my psychological state is consistent with more than one decision} \). Given that the aim of deliberation is to narrow the probability distribution and reduce the risk-of-error, the propagation proposition cannot be assumed to be true for \( p(2) \). In a successful deliberation which ends when an agent feels trust and confidence in his decision, it is always false.

The psychological world is not a closed system. Through the course of a life, a myriad of external factors (parents, teachers, published research, internet searches etc) shift an individual’s psychological state, and over the course of extended deliberation new evidence and information does the same. Psychological states are evidence-responsive. A prior psychological state cannot tell you whether a future decision will be open.

Psychological openness, as defined above, is true but irrelevant for questions of libertarianism. For an agent to use experience-based seemings to justify beliefs in libertarianism those experiences must be synchronous with the claim. They must be felt at the moment of intention formation.

This implies that no experiences during the early stages of deliberation can ever justify libertarian claims, anti-natural or psychological. But it remains possible that an experience at the moment of intention formation could do so. In arguments 2 and 3, I analyse this possibility and conclude that it does not.
2.3 Argument 2: The cognitive phenomenology at the end of “successful” deliberation does not support beliefs in psychological libertarianism.

Talk of the psychological impossibility of reducing the set of nearby possible worlds until it includes only one unique world may seem depressing. But it should not. Agents have the power to narrow the set of worlds and in the context of decision-making this can be very valuable.

To answer the question of which university to attend, Stephen doesn’t need to know the extent to which time runs slower at the top of Everest, or whether his Achilles tendon will handle tomorrow’s long run. He just(!) needs to confirm, that in no nearby possible world consistent within his current psychological understanding is university β more attractive than university α. To use another Dennettian phrase, all unknowns unrelated to university choosing are differences that don’t make a difference (1991).

And if Stephen is uncertain, it is possible for him to do something about it. He can engage in extended deliberation - the bundle of epistemically-focused processes and actions that drive information gain and uncertainty reduction - to narrow the set of possible worlds. He can zoom in on information relevant to the matter-at-hand, modulating the granularity of description and closing in on a set of worlds that includes only one decision. Whilst reducing his set of possible worlds to a single world is impossible, narrowing the size of the set until all worlds involve one and the same decision is extremely common. Indeed, the aim of extended deliberation is to shift an agent’s psychological state from one consistent with psychological openness, to one consistent with psychological choice determinism. Of course, sometimes uncertainty remains, but let me set that possibility aside for a paragraph or two and focus on a “successful” deliberation, which resolves indecision by
resolving uncertainty. Decisions that follow successful deliberation are passive ‘comings-to-intend events that resolve a state of uncertainty over what to do’ (O’Shaughnessy, 1980, 297). During ongoing deliberation more than one final decision is consistent with the total current psychological state of the agent but after successful deliberation only one decision is consistent with her total psychological state. The phenomenology(s) of rising levels of confidence during, and of trust and confidence as a consequence of, successful deliberation support this view of a deliberation-driven transition.

The feeling of low confidence in the early stages of deliberation does not support the libertarian position, because it can be, and often is, followed by a feeling of trust and confidence at the moment of intention formation. Indeed, if the process of extended deliberation is a natural process, then the phenomenology of increasing confidence counts as seemings-based evidence for naturalism and against the libertarian viewpoint.

2.4 Argument 3: The cognitive phenomenology at the end of “unsuccessful” deliberation does not support beliefs in psychological libertarianism

Given the synchronicity requirement and the lack of alternative possibilities at the conclusion of successful deliberation, the last hope for psychological libertarians is the case of unsuccessful deliberation, which results in a decision being made despite an ongoing risk-of-error. Here is a potential claim:
The cognitive\textsuperscript{133} phenomenology of making a decision despite an ongoing risk-of-error justifies a belief in psychological libertarianism.

I will argue that, although an ongoing risk-of-error is experienced as low confidence, this claim is false. Let’s return to Stephen, who represented his psychological state as a probability distribution of future possible worlds. The tails of the distribution represent unlikely future possible worlds that are less and less “nearby”.

![Figure 9. Current and possible future motivational balances, and levels of confidence.](image)

Stephen thinks that choosing the right university is an important decision and so he plans to deliberate further and reduce the risk of error. Although in 93.3\% of the

\textsuperscript{133} I consider the conative phenomenology of making a decision despite ongoing uncertainty in chapter 9
futures that he believes are possible, he will eventually choose university α, he wants to get that percentage higher. But suddenly Stephen’s parents run in. They have re-read the offer conditions and the decision deadline is today! His deliberation time has run out and he must make a call right now. Does his experienced risk of error indicate that he has alternative possibilities? No, it does not.

To illustrate why, imagine that I asked you to choose between two bags of money. You ask which bag has a greater value and I answer that although I can’t tell you with certainty, I can give you the following information: It is 92% likely that the right bag is worth between zero and £4,000 more than the left bag, and 6% likely that the left bag is worth between zero and £1,000 more. In fact, here are the probabilities for each £1,000 range:

<table>
<thead>
<tr>
<th>Bag with highest value</th>
<th>Difference between values</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>1,000-2,000</td>
<td>1%</td>
</tr>
<tr>
<td>Left</td>
<td>0-1000</td>
<td>6%</td>
</tr>
<tr>
<td>Right</td>
<td>0-1000</td>
<td>24%</td>
</tr>
<tr>
<td>Right</td>
<td>1,000-2,000</td>
<td>38%</td>
</tr>
<tr>
<td>Right</td>
<td>2,000-3,000</td>
<td>24%</td>
</tr>
<tr>
<td>Right</td>
<td>3,000-4,000</td>
<td>6%</td>
</tr>
<tr>
<td>Right</td>
<td>4,000-5,000</td>
<td>1%</td>
</tr>
</tbody>
</table>

Figure 10. Which bag would you pick?

Which should you choose in a single round game? The right bag, obviously. You might be worried about the 7% risk of error but that does not justify picking the left alternative. After all, picking the left option has a 93% risk of error. But there is a possible objection: perhaps you don’t like money, or want to be contrary.

134 In a multi-round game, it might make sense to choose non-optimally in order to gain further information about the distribution. This would be an explore rather than exploit strategy, pursuing epistemic value rather
Now imagine that the values are not just monetary but conative. They capture all your motives (including any motives to engage in exploratory behaviour), so that there is no motive external to the conative value computation. Does the risk of (conative) error indicate that you can or should choose the left option? No. You would never pick a 93% risk of (conative) error over a 7% risk, especially as the potential harm associated with erroneously choosing left is much greater than the potential harm associated with erroneously choosing right\textsuperscript{135}.

And then, finally, remember that the range in figure 9 indicates not current motivational balances, but possible future motivational balances contingent upon further deliberation. Your current motivational balance is there in the centre of the distribution. You are currently more motivated, by a significant margin, to choose the “right” option. If you have to choose now, then you will choose to do what you are currently most motivated to do and there is no uncertainty about what that is.

Stephen’s probability distribution indicates that it is unlikely, but possible that he will eventually be more motivated to go to university $\beta$, but that possibility is contingent upon two facts. First, it requires that Stephen continues deliberating. Without this his motivational balance will remain unchanged in favouring university $\alpha$. Second, it requires that, as a result of his further deliberation, Stephen uncovers a shocking, explosive piece of evidence that overturns his analysis and inverts his motivational balance. Without such explosive evidence, his motivational balance will remain unchanged and his confidence in university $\alpha$ will slowly increase.

\textsuperscript{135} As seen in chapter 4, a foregone conative value is a conative cost (ie a harm)
Some readers might remain resolute, despite the arguments of part I and II, that an agent can choose to override his strongest motive. That is a position which might be justified by the feeling of effort involved in making a difficult decision. For that reason, I will analyse the conative phenomenology of an effortful decision in the following chapter. But I hope those readers will agree that I have shown that there is nothing in the cognitive phenomenology of (un)confidence, or the functional implications of a risk-of-error, that justifies such a view. Whilst levels and feelings of confidence are important for questions about whether an agent should engage in further deliberation, and about when an agent should make a practical decision, outside of that they are not relevant for the question of what practical decision an agent will make. The feeling of a lack of confidence does not support beliefs in psychological libertarianism even if it continues to exist at the time of the practical decision.

I have argued that neither the cognitive phenomenology of confidence nor of unconfidence justify beliefs in psychological libertarianism. Before I close this section, I will analyse the implications of decisions in the absence of conscious judgement. I will conclude that they are not a problem for the account.

2.5 Decisions in the absence of conscious judgement

In chapter 7 I argued that the evidence and processes that drive a decision can be sub-personal. Could the possibility of an agent who is unaware of the underlying reasons for her choice impact the analysis of this chapter?

Throughout part III, I have talked of the feeling of low confidence as being similar to the feeling of hunger. This suggests that I can answer the question with a whimsical, but relevant metaphor:
Imagine that you are an octopus and the colour of your skin is determined by what you have eaten so far during the day. You have no idea how it works or which foods drive the different colours. What you are certain of, however, is that ongoing eating implies the possibility of skin-colour change. And hunger is correlated with eating, so you associate hunger with the possibility of skin-colour change. But the hunger-to-skin-colour relationship is contingent upon a particular response to hunger: Eating. If you commit to eat no more, then it is not true that your skin-colour might change. This is the case no matter how opaque the underlying mechanism, nor how hungry you are.

For a more formal response to the question, I rely heavily on Richard Holton’s detailed analysis of ‘choice in the absence of conscious judgement’ (2006, p8). Here is an example Holton cites of a decision made by a fire lieutenant, as originally documented by Gary Klein:

The lieutenant leads his hose crew into the building, to the back, to spray water on the fire, but the fire just roars back at them. “Odd,” he thinks. The water should have more of an impact. They try dousing it again, and get the same results. They retreat a few steps to regroup. Then the lieutenant starts to feel as if something is not right. He doesn’t have any clues; he just doesn’t feel right about being in that house, so he orders his men out of the building. (1998, p. 32.)

Soon after the firemen leave the building, the floor where they had been standing collapsed. Holton picks up the story:

It turned out that the source of the fire was in a basement. The lieutenant had picked up on various indicators of this—especially the great heat of the fire
and the lack of noise relative to this heat—and this produced an emotional response to the situation that influenced his action. But he didn’t realize that he had picked up on these factors. (2006, p8).

The first question that this example raises is whether the lieutenant’s psychological state was consistent with more than one decision. The final sentence of Klein’s text indicates that it was not. The lieutenant didn’t feel right, so he ordered the men from the building. He may not have known why he didn’t feel right, but this was an experienced fire lieutenant who trusted his instincts. Not feeling right was sufficient for making the decision to get out. The lieutenant is neither irrational or unusual in doing so, as Holton states in his detailed analysis, ‘we can be good at doing something without making any [conscious] judgments. The psychology literature is full of examples’ (p7). The lieutenant was most motivated to get his team out, so that is what he did. The lack of a conscious judgement is not, on its own, relevant to questions of libertarianism\textsuperscript{136}.

But what if we ratchet down the lieutenant’s strength of feeling that something is not right? Will that mean he could do otherwise? That depends on whether the declining feeling is conative or cognitive. Assume, first, that the lieutenant’s motive to leave remains higher than his motive to stay, but his feeling of confidence in the decision to leave falls. His increased uncertainty indicates that further deliberation would be valuable in resolving the question of whether leaving is necessary. But he is in a burning building, he needs to make a call right now. Does his decline in confidence from 95\% to 80\%, for example, mean he may decide differently? No. If he has to decide now then he will make the best decision he can based on the (conscious and

\textsuperscript{136} See Holton for the full argument. I diverge from his account in chapter 9, where I explain the feeling of effort in a difficult decision.
unconscious) evidence he has. The fact that further deliberation would open up a hypothetical set of worlds, a small minority of which involve a change of mind, is irrelevant to his current real-world decision.

If, on the other hand, it is the conative phenomenology of motivational pull towards leaving that declines, then of course the lieutenant might change his mind. If he is more motivated to stay then he will stay, consistent with the motivational strength idea.

There is one situation in which an agent’s psychological state is legitimately consistent with two possible decisions: When the functional motivational contest results in a tie. Here the agent will feel torn between two equally attractive possibilities, like Buridan’s ass between two carrots. Can she decide then? The story of Buridan’s ass raises the possibility that she cannot. If there is literally nothing she can do to resolve indecision, then her current psychological state is not consistent with two decisions but with indecision. She would be compelled to deliberate further or to give up and leave the question unanswered. In real life however, she can harness external or internal randomness. She can flip a coin, physically or metaphorically, to break the tie. But this does not justify beliefs in psychological libertarianism for two reasons. First, harnessing randomness is a technique for narrowing possible worlds. Whilst the decision may be psychologically underdetermined while the (metaphorical) coin is in the air, it is psychologically determined when it lands. Second, and perhaps most importantly, random picking based on blind chance is not indicative of an interventionist agential ability to bias a decision, but a sign of cognitive limitations. It is not the source of agential power, but the end of the agential road (Holton, 2006, Vierkant, 2022, List, 2019, Kane, 1999).
However, there is a nuance. It is plausible that there is a zone, before the agent reaches indifference, in which she loses the ability introspectively to determine which of her own motives is strongest. This is the zone where *it feels like picking, but it is not picking*. In chapter 1 I called this “seeming incommensurability”. Even if cognitive phenomenology of low confidence does not support libertarian intuitions, could seeming incommensurability do the job?

It is true that an agent who must decide now, but has neither consciously available reasons for a decision nor confidence in her decision, nor an introspectable motivational difference, is in an unpleasant position of opacity with regards her own psychological state. It would be fair for her to throw up her hands (or eight legs) and say that she has no real insight into what is going on within her own decision-making process. But such opacity with regards what is going on at the psychological level is not necessarily a justification for bold metaphysical claims. As List said, ‘the “psychological level” must not be confused with a person’s subjective, first personal perspective. What people subjectively experience need not coincide with how our best theories of psychology describe human cognition and behaviour’ (2019, p90).

What could those best theories of psychology be? An initial response is that scientists and psychologists have sophisticated naturalist accounts, and that Holton’s detailed analysis explicitly accommodates seeming incommensurability in a naturalist framework. But to see why those have not yet been persuasive, it is perhaps worth returning to a discussion from introductory chapter (i): I suggested that whenever anyone argues against a seemings-based claim p, they should expect to be asked a simple question: “If p is false why does p seem to be true?” If unanswered, this question creates a worry that it is the defeater that is flawed. My
account, which might be considered a thesis-level defence of the commensurability of motives, did suggest an answer to that question in chapter 3:

Whilst an attitude’s phenomenological intensity is conceptually separable from details of action readiness, from a practical perspective it is difficult to isolate. If the functional strengths of two motives are similar then, when the plan-contents are quite different, the agent may not be able to introspectively rank their intensities. The window into the functional strength of motives is translucent rather than transparent, so motives may seem to be incommensurable due to other aspects of conative phenomenology, even if from a functional perspective they are not. Seeming incommensurability is predicted and explained within my account. Of course, my definition of a persuasive defeater doesn’t guarantee that libertarians will be persuaded!

To return to the central focus of the chapter, a feeling of unconfidence does not imply that the agent’s psychological state is consistent with making more than one decision right now. It only implies that she is “hungry” for more information. Whilst the level of confidence bears on the metacognitive question of whether to stop or prolong deliberation, it is the agents motivational balance which bears on the practical question of what to do. This is the argument of chapter 7, but in reverse. Whilst levels of confidence (relative to the threshold) determine when an agent will make the choice, motives determine what an agent will choose to do. By carefully differentiating between the metacognitive and practical questions, one can ensure that the proposed inputs to the evaluations are relevant. The feeling of a lack of
confidence does not support beliefs in psychological libertarianism even if it continues to exist at the time of the practical decision.

In the final section of this chapter, I will review the adaptive benefits of cognitive phenomenology, responding to a quote from Searle.

3. Evolutionary reasons for the experience of alternative possibilities

In the conclusion of his perceptive analysis of libertarian beliefs, Searle self-deprecatingly states that:

For reasons I don't really understand, evolution has given us a form of experience of voluntary action where the experience of freedom, that is to say, the experience of the sense of alternative possibilities, is built into the very structure of conscious, voluntary, intentional human behaviour. For that reason, I believe, neither this discussion nor any other will ever convince us that our behaviour is unfree (1994, p94).

The analysis of this chapter can shed light on Searle’s question of evolutionary reasons.

At the core of any even slightly sophisticated agent is a monitor-control loop in service of the integrity of the organism. The agent monitors the (internal and external) environment and responds to increase the likelihood of survival. The monitor-control loop informs norm-driven striving.

In this chapter, I have analysed the cognitive phenomenology, which raises to conscious awareness a variable, confidence, that is being monitored. Low levels of confidence are aversive, whilst (sharp) increases in confidence are experienced as
pleasurable. These activate motives which aim to increase the former or the possibility of the latter. This link between valence of affect and the activation of motives is not unique to human metacognition. Quite the opposite. It is an evolutionarily ancient solution which has been repurposed for use in service of the reduction in uncertainty-based harm\textsuperscript{137}.

There are hard questions about why agents have any conscious experience, and the extent to which we can know about the consciousness of other creatures, but I will set those questions aside. They apply just as much to feelings of hunger or pain as they do to feelings of confidence.

If levels of confidence are experienced, then the phenomenology(s) of low confidence, of increases in confidence, and of trust and confidence, will naturally follow. The first of these experiences underwrites what Searle describes as the experience of alternative possibilities. There is no puzzle as to why we have this experience. The puzzles only arise when this experience is used to justify physical-level, asynchronous or non-contingent claims.

Indeed, a more fine-grained analysis of the complex and shifting phenomenology reveals that uncertainty about what to do is resolved by the natural process of narrowing down alternatives through the interrogation of facts, just as hunger is resolved through eating. Humans satiate their hunger for certainty by foraging for

\textsuperscript{137} This is supported by neurobiological findings which ‘lead to the compelling conclusion that the same cognitive and neural processes underlie much of human behaviour involving cognitive search – in both external [foraging] and internal [deliberation] environments’ (Hills et al. 2015, 47). Even more specifically, recent experiments in neuroscience indicate that both foraging (Hayden, Pearson, and Platt 2011; Li et al. 2012) and cognitive control (Shenhav, Botvinick, and Cohen 2013) are associated with activity in the anterior cingulate cortex.
information, and that should be no harder to understand than the fact that they satiate more prosaic hunger by foraging for food.

The interesting question is not why an agent has such a feeling of hunger for information, but why she sometimes makes a decision without satiating that hunger. That is the question I will address in the following chapter and pleasingly the answer has a lot in common with the answer to why an agent would not eat when hungry. The metaphor can be sustained.

Conclusion

I have shown that feelings of low confidence do not justify libertarian beliefs. However, I am cognisant that I have not yet considered the conative phenomenology of effort associated with making a difficult decision. In the following, and final, chapter I will argue that making a difficult decision feels effortful because it involves resisting the desire to deliberate further. If a decision can be made despite an ongoing desire to deliberate, then this implies that the threshold of confidence at which a decision is made is not fixed: confidence in a decision is a want-to-have, rather than a must-have. Whether the current level of confidence in a decision is sufficient given the context, is a metacognitive question which compares the costs and benefits of prolonged deliberation. It is resolved in accordance with the motivational strength idea. Making a decision is a motivated, intentional action. It is the act of concluding deliberation, despite the ongoing desire to deliberate.

I consider the implications of this proposal for libertarianism and find (perhaps surprisingly) that it strengthens my argument further. This is because the act of concluding deliberation causes a shifting lower of the decision threshold, stopping
the decision-making process “early”. To use a racing analogy, a passive decision is a
race until one of the competitors reaches a static finish line, whereas an active
decision is a shift in the finish line until it meets the first competitor. Such an action
does not and cannot change who is currently in the lead.
Chapter 9
The Act of Making a Difficult Decision

1. Introduction

1.1. Difficult Choices

Some choices are easy, but others are hard. They may be hard to make, because they involve uncertainty about what to do, or hard to stick with, because of an ongoing desire to do otherwise. Phenomenologically, both kinds of hard choice involve the feeling of effort, which Massin (2017, p235) categorises as 'cognitive effort' and 'effort of will' respectively. Although there is plausibly a correlation between choices which require cognitive effort and those that require effort of will, the two should be considered separately as they sit on opposite sides of the moment of intention formation. Cognitive effort is felt in the process of forming intentions, whilst effort of will is felt by an agent trying to act in accordance with a formed intention. In this chapter I limit my analysis to the experience of cognitive effort that accompanies the making of difficult choices.

Difficult choices are hard but making them is central to the experience of being an agent. In debates regarding free will, both compatibilists and libertarians agree that a paradigmatic experience of freedom is the felt effort associated with making a difficult choice. As William James describes, ‘those soul-trying moments’ are ‘what

138 I set aside Massin’s third category of muscular effort.
give the palpitating reality to our moral life and make it tingle’ (2014, p183). Despite this agreement, there is dispute over the process by which a difficult choice is made, whether making a difficult choice is an action, and why it feels effortful. My aim in this chapter is to provide an evolutionarily plausible, actional account of deciding which explains the phenomenology.

In section 2 I begin by reviewing two arguments for a standard non-actional model of decision making. I then argue that the phenomenology of effort indicates that something is missing from this standard account. What is missing is acknowledgement that further deliberation is a motivated rather than compelled response. Making a difficult choice involves concluding deliberation, despite ongoing risk-of-error. From this foundation I outline a positive proposal, which explains why making a difficult choice is an act, and why it feels effortful.

The act of making a difficult choice is triggered by a decision to intentionally conclude deliberation, despite being motivated to increase confidence further. This decision is the output of a metacognitive cost-benefit computation, which evaluates whether to stop or prolong the cognitive process.

More formally, when making a decision about what to do, an agent performs two evaluations:

Evaluation E (cognitive): An evaluation of whether to perform action A or ¬A

Evaluation M (metacognitive): An evaluation of whether to stop or prolong E-ing

Here, A is an intentional act of control and ¬A is anything other than A. ¬A includes doing nothing, an action B, multiple possible actions B or C or D etc, or multiple combined actions B+C+D etc.
Metacognition is the series of processes that monitor, evaluate and control cognitive activity. Evaluation M informs the metacognitive act of stopping (or prolonging) cognition, which is the metacognitive version of the evolutionarily antecedent ability to stop (or prolong) physical activity. I will suggest that evaluation M compares the value of increasing confidence with the opportunity costs of prolonging deliberation. Whatever its precise form, evaluation M has the functional impact of stopping or prolonging the evaluative process E.

Importantly, and perhaps controversially, the pros and cons of A-ing are irrelevant to evaluation M, other than indirectly via confidence. In slogan form one might say that evaluation M compares decision confidence with decision urgency, and influences decision timing. In situations of urgency, such as fight-or-flight, or low stake decisions, evaluation M will stop evaluation E very quickly. In scenarios where time is no constraint but making the correct choice is of high importance, evaluation E will be prolonged, possibly for many days.

In section 3 I will explore the links between this hierarchical proposal and the analysis of previous chapters.

In section 4 I show that the insights from models of physical stopping behaviour in foraging, as studied within the field of ethology, are directly applicable to the metacognitive stopping behaviour necessary for the making of a difficult choice. I then review evidence that metacognitive monitoring in primates leads to physical stopping behaviour; and finally turn to human metacognitive control of stopping and prolonging study time. As well as supporting my proposed revised model of decision making, this explains its development through a process of what Cisek calls ‘phylogenetic refinement’ (2019, p2265).
In section 5, I review contemporary theories of the feeling of effort. Consistent with these theories is the claim that evaluation M is not only the source of the binary output of whether to stop or prolong cognition, but is also the source of the analogue feeling of effort that tracks the costs of that decision. Making a difficult choice feels effortful because the mind-directed intentional act of stopping the cognitive process has risk-related costs, and those costs are revealed to the agent through affect.

Finally, in section 6, I explore the implications of the proposed account for libertarian beliefs.

The act of making a difficult choice feels effortful. Through an understanding of the inhabitable metacognitive motivation to increase confidence, one can finally explain and understand this paradigmatic experience of agency.

1.2. Clarifications and Assumptions

Cognitive effort includes both the temporally extended effort of ongoing deliberation, and the more momentary effort felt when stopping deliberation and making a difficult choice. As Kriegel states, the process of deliberation ‘has a tangible duration to it but the act of deciding, of making up one’s mind, is….. instantaneous’ (2015, p85). The account I develop focuses narrowly on what Mele calls ‘momentary mental actions of intention formation’ (2017, p85), when an agent stops deliberating and decides what to do. This is the moment to which I refer when I use the phrase making a (difficult) choice. However, as will be seen, there is a pleasing symmetry between explanations of effortful prolonging and effortful stopping.
I will assume that extended deliberation captures the full decision-making process, within which are evaluations, mental acts of stopping or prolonging cognition and any further acts of what Hieronymi calls ‘managerial control’ (2009, p138).

Evaluations may be fast and unconscious or may explicitly follow learned propositional rules. Either way, I will assume that an evaluation is involuntary in the sense that the output is a natural function of the inputs and the agent’s cognitive computational process. Despite this, evaluations by different individuals may lead to different outputs, reflecting their differing experiences and values. Some may seem sub-optimal to an external observer.

Finally, I use choice and decision interchangeably, confidence should be understood as inversely proportional to risk-of-error, and any mention of uncertainty should be understood as shorthand for practical uncertainty about what to do. I will assume that it is possible for an agent to be practically uncertain about what to do even if she knows what she should do, and that, necessarily, any agent with an unsettled question of what to do, is uncertain about what to do.

Uncertainty about what to do results from an interaction between the agent’s level of confidence in a decision and a threshold. The key insight of this chapter is that active decisions are only possible because decision-thresholds are flexible. When confidence is low, agents are motivated to prolonging deliberation, but motives are never irresistible.

2. Making a Difficult Choice

2.1. Confidence and Decisions
In the philosophical literature there has been significant debate over whether deciding is, or can sometimes be, an intentional act. Proponents of a non-actional view of deciding argue that decisions are passive events of intention acquisition, and are the automatic culmination of a process of evaluation:

The movement of the natural causality of reason ... to its conclusion in choice or decision is lived (by some) as action when it is really just reflex; distinctively rational reflex, to be sure, but not in any case a matter of action. (Strawson 2003, p244)

As Mele explains, ‘according to this view, to decide to A is simply to acquire an intention to A on the basis of practical reflection, and acquiring an intention—in this way or any other—is never an action’ (2017, p82). In order to approach the question of actional vs non-actional decisions, I first analyse the interaction between decisions and confidence.

The question of what to do arises when an agent is faced with alternative courses of action and she is uncertain about what to do. She must decide between the alternatives. If there is no alternative or no uncertainty then no decision is required. Low confidence implies a risk-of-error. To reduce the potential harm associated with error, a decision-making process is engaged. In an ideal world this process would incorporate all relevant evidence and activate all of the agent’s relevant beliefs and desires, enabling her to (consciously or unconsciously) evaluate the pros and cons of an action A vs its alternatives until she is confident about what to do. Consistent with this, O’Shaughnessy describes decidings as ‘those comings-to-intend events that resolve a state of uncertainty over what to do’ (1980, p297). I will call this the standard model of decision making, as illustrated in figure 11. In this model,
uncertainty is *resolved* when the level of confidence exceeds a threshold such that deliberation comes to a natural end. The end is *natural* because the agent is no longer motivated to deliberate. I will therefore call the threshold reached in a standard (or ideal) decision the *natural threshold*. At the natural threshold an agent feels full confidence in her decision.

This is a classic negative feedback system: ‘The action [deliberation] is performed to eliminate the conditions that motivated the action’ (Cisek 2019, p2269). Decisions flow naturally and passively once uncertainty is resolved. Although the transition from low to high confidence can be sharp, as in an aha! moment, more commonly it is smooth. As Holton says, ‘forming an intention can sometimes seem more like a rolling ball finding its equilibrium settling point, and less like the tripping of a switch’ (2006, p11).
The argument that decisions are non-actional is strengthened by the link between actions and intentions: If a decision is an act, then, as Davidson has argued, it must be intentional under some description (1980). Equally, as per Bratman (1987), the content of any intention is an action-plan. The difficult question for proponents of actional decisions is what specific action-plan is contained within the intention to decide? The obvious proposal in the case where an agent decides to A, that the content of the prior intention is the plan to decide to A, seems problematic. An intention to decide to A is at minimum inefficient and, as Kavka’s (1983) famous toxin puzzle illustrates, in normal situations is just an intention to A. Perhaps the reader can imagine some complicated scenarios where an agent is obstructed from forming an intention to A directly, but I will follow Mele in assuming that ‘in normal cases, there is no purpose in non-actionally acquiring an intention to decide to A that is not more efficiently served by a non-actional acquisition of an intention to A’ (2017, 92).

This implies that for an actional decision to exist in normal situations, the content of the prior intention must be something other than <decide to A>. Let us call it the intention to <X>. If making a difficult choice is an act, then it is the act of X-ing. Therefore, in order to answer the question of what triggers the act of making a difficult choice, and to explain its phenomenology, I must first elucidate what X is. It will then be possible to determine the process through which the intention to X is formed, and why X-ing is effortful. I will return to these questions in section 2.3, after a discussion of the importance of phenomenology.

139 The assumption that the prior intention is non-actionally acquired is necessary to avoid (potentially infinite) regress.
2.2. The Phenomenology of Effortful Choices

Despite the concerns outlined in section 2.1, many philosophers (Mele 2017; Shepherd 2015; Pink 1996; Holton 2006) believe that the making of a difficult choice is an intentional mental act. This position is consistent with folk intuitions, which are themselves grounded in phenomenology. Difficult choices are effortful and, as Bayne and Levy state, ‘the experience of effort involves an experience of the self as a source of force’ (2006, 63). The cognitive effort that accompanies the making of a difficult choice makes the agent feel as though the choice is an intentional act. If making a difficult choice is effortful, and effort is a marker of intentional action, then making a difficult choice must be an action.

Holton goes further, arguing that ‘choice is the primary ingredient in the experience of free will’ (2006, p2). This raises a challenge for Holton as he is a compatibilist, engaging with the broader literature regarding the gap between the standard mechanistic or computational accounts of how agency works, and everyday experience of how agency feels. This phenomenological gap explains the intuition that there is something missing from the standard account. As Velleman (1992) explains, a passive weighing of beliefs and desires doesn’t capture what it feels like to be an agent. Difficult choices feel like actions taken by agents, so compatibilists need to be able to place the agent within their model rather than replacing her with a ‘passive vessel’ (Holton 2006, p5).

Vierkant has called the challenge of accounting for the experience of effortful choice within a compatibilist framework ‘the agency problem of compatibilism’ (2022, p116).
The worry is that if the compatibilist account fails to solve this problem, then that will drive some to libertarianism.

Holton’s solution is to state that difficult choices are made ‘in the absence of judgement’ (2006, p7). Whereas in the non-actional account decisions are assumed to flow naturally from judgements, in difficult decisions no decision flows naturally because the agent is unable to form a judgement. Importantly for Holton, a choice in the absence of judgement is not explained by the standard model. Released from these tight constraints, it can be an action\(^{140}\).

Having suggested a solution in which a decision can be an action, Holton’s main concern is to deflate worries that decision making in the absence of judgement is just random picking. He relies on a wealth of empirical data, to argue that when making a choice in the absence of consciously available reasons, the individual relies on her subconscious, which can be quite skilful. Although relevant for the question of why the agent chose what she did, Holton’s analysis of the subconscious doesn’t seem to explain the phenomenology of making an effortful choice. Specifically, why would the shift to a mechanism where choices are influenced by unconscious states make a decision more active and more effortful? The choice would be harder to justify, and may feel more confusing, but, if anything, feelings of effort arising from unconscious processes might be expected to be less forceful than their conscious equivalents. Vierkant worries that Holton’s solution might imply choices ‘have no [conative] experiential echo at all’ (2018, p4). At the very least, further work is needed to elucidate this issue.

\(^{140}\)The implication seems to be that decisions made in the absence of judgement are of a different (effortful) kind. I will argue against that view in section 4.
The central questions highlighted in 2.1 remain. With regards to the intention to X, what it is its precise content and how is it formed? And when the intention to X is executed, why is there a feeling of effort?

Inspired by Holton’s critique of the standard model, I would like to propose an alternative solution based, not on a lack of conscious judgement, but on a lack of confidence. If there is no uncertainty about what to do there is no decision to be made. If a decision is made after confidence has been increased through a successful process of deliberation, that is a standard decision. However, if a decision is made whilst confidence remains low, that is a difficult choice. Difficult choices are not explained by the standard model, so making a difficult choice can be an action.

In the next section, I outline a positive proposal for what motivates difficult decision making and why it feels effortful. Although I agree with Holton’s insight that the standard model does not account for difficult choices, I suggest that the outstanding questions of section 2.1 can be answered through analysis of metacognitive motives, rather than of the subconscious expertise on which Holton focuses. I develop this proposal further, and support it with empirical data, in section 4, where I suggest a revised model of decision making, and section 5, where I incorporate contemporary theories of mental effort.

2.3. Low Confidence and Difficult Choices

In the account set out thus far, confidence has played two roles. First, full vs low confidence is the differentiating factor between standard decisions and difficult choices. Second, low confidence was described as a reason to engage in deliberation. In what follows, I expand on this latter point.
I take it as given that a cognitive infrastructure exists that encodes an agent’s disparate values (Padoa-Schioppa 2011; Smith et al 2010; Levy and Glimcher 2012; Westbrook and Braver 2016). These may be innate or learned, intrinsic or extrinsic, subconscious or explicitly endorsed. A valued end is an end consistent with those values. An agent’s valued ends can all be considered ends to which she is implicitly or explicitly committed. To be committed implies being motivated, in appropriate circumstances, to act in ways that achieve states of affairs consistent with those ends.

As Cisek says:

> At its heart [a behavioural control system] is an evaluation of the animal’s current state in relation to a range of desirable states. Deviations of the… state outside the desirable range constitute the motivation for actions that improve the state. (2019, 2268-9)

To reduce the potential harm associated with error, humans monitor confidence through noetic feelings (Dokic 2012), and ascribe value to information and to increases in confidence. Ascribing a negative value to risk-of-error and a positive value to information explains the existence of curiosity as a motivating force. ‘The curious individual is motivated to obtain the missing information to reduce or eliminate the feeling of deprivation’ (Loewenstein, 1994, p87). Miščević goes so far as to state that ‘curiosity is the central intrinsically motivating drive for achieving knowledge and understanding’ (2020, p85). Put simply, the value of reducing the risk-of-error motivates deliberation. This is an implicit assumption of the standard model.
However, there is also value in appropriately allocating finite cognitive and temporal resources. ‘Sometimes continued deliberation is costly’ (Shepherd 2015, p346). Spending too long focusing-on and deliberating-over a single issue has opportunity costs. These relate to the specifics of the situation as well as the general fact that focused attention lowers the ability for quick and flexible reconfiguration to a new task when the environment changes (Kurzban et al. 2013; Musslick and Cohen 2021). The greater the opportunity costs the greater the urgency to curtail deliberation. ‘It would thus be useful if an agent possessed the ability to terminate deliberation ‘at will’’ (Shepherd 2015, p346). Conflict between the value of increasing confidence and the value of deciding quickly creates an explore-exploit tension (Hills et al. 2015) or speed-accuracy trade-off (Bogacz et al. 2010) that an agent needs to manage. The agent evaluates these costs and benefits to determine whether to stop or prolong deliberation. This metacognitive evaluation, evaluation M, can culminate in the decision to stop deliberation and make a choice, even if confidence remains low.

Importantly, the costs and benefits of evaluation M do not include any (conative) evidence for A-ing or ¬A-ing. As Kane states, choices ‘are made intentionally or on purpose not by virtue of a specific prior intention to make the particular choice made but by virtue of the general intention to resolve indecision in one way or another’ (1999, p139). Making a difficult choice resolves indecision without resolving risk-of-error. It is the act of concluding deliberation.

Note that although the explanandum of this chapter is the process by which difficult choices are successfully made, it is possible that the decision-making process ends with a decision between two acts, A and B, left unmade. This can be separated into two scenarios. In the first scenario, the agent actively and effortfully stops the
process for good and gives up on making a decision between A and B. This is captured above by the specification that \( \neg A \) includes doing nothing. A decision between doing A or B or nothing, was concluded in favour of doing nothing. In the second scenario the agent is unable to reach a decision between A and B for now, but (consciously or unconsciously) intends to return to the issue, in the hope that more evidence or inspiration will lead to a satisfactory resolution later. This is not concluding but pausing the decision-making process, what might be called interrupted prolonging. It could possibly be effortless if, for example, the need to resolve indecision is not pressing.

Going forward I will use the stronger word conclude, rather than stop, to indicate that deliberation has been resolved (even if the decision is to do nothing) rather than paused.

Having clarified this nuance, I suggest that the content of the intention which is required for a difficult choice to be an action, is *conclude deliberation now, despite ongoing risk of error*.

Henceforth I will use a shortened version:

\[ X = \text{conclude deliberation now} \]

In the following section I will explore the links between this proposal and the analysis of previous chapters

3. Active decisions in context of previous findings

3.1 The motivational strength idea
Deliberation can conclude naturally and passively when confidence is high, but if there is ongoing risk-of-error an agent is motivated to deliberate further. Actions are the result of motivational contests, and the act of concluding deliberation is no different. For an agent to actively make a decision, her metacognitive motive to conclude deliberation must defeat, and laterally inhibit, her motive to deliberate further. Intentionally concluding deliberation is only possible when an agent is more motivated to conclude deliberation than she is to increase confidence through further deliberation.

3.2 Inhibition

O'Shaugnessy described decisions as ‘coming-to-intend events’ (1980, p297) and I argued in chapter 2 that an intention is an uninhibited motive. All competing practical motives are inhibited while deliberation is ongoing, but once the question is settled inhibition is released from the strongest motive. Concluding deliberation is therefore a coming-to-intend event because deliberation-driven inhibition (of the chosen motive) is released. This is true whether the deliberation is concluded naturally and passively, or actively.

3.3 Accumulation to bound models

In chapter 7 I argued that: at the sub-personal level a decision is ‘the crossing of a threshold in a neurally implemented accumulation to bound process’ (Burnston, in submission); the accumulator measures confidence; and the threshold is norm-driven and context dependent. Decision timing is therefore a function of the interaction between the accumulator and the level of the threshold. A crossing event
can be driven by a rising accumulator, but it can also be driven by a falling threshold. In passive decisions, deliberation concludes because confidence has accumulated and crossed the natural threshold. In active decisions, on the other hand, the motive to allocate cognitive resources elsewhere causes a shifting lower of the decision threshold. The threshold moves lower until it “crosses” the current level of the accumulator “from above”. When an agent forms the intention to actively conclude deliberation, this is implemented in neural architecture by a shift lower in the decision threshold, which diverges from the natural threshold\(^\text{141}\) and causes an “early” decision at a lower level of confidence. To use a racing analogy, a passive decision is a race until one of the competitors reaches a static finish line, whereas an active decision is a shift in the finish line until it meets the first competitor.

I illustrate this point with the two charts in figure 12, which are consistent with the drift diffusion model. In the example on the left, full confidence is required, and deliberation is ongoing. It remains possible, but unlikely, that some new, explosive evidence is uncovered which causes the accumulator to shift trajectory and eventually hit threshold B. In the example on the right, however, the thresholds were narrowed until they hit the accumulator. At that point the decision was made, inhibition was released from action-plan A, and implementation began\(^\text{142}\).

\(^{141}\) The importance of the fact that the natural threshold is above the decision threshold and therefore implies an ongoing, but defeated, motivation to deliberate further, will become apparent when I turn to the feeling of effort.

\(^{142}\) Readers might wonder whether I endorse a model in which the metacognitive decision is also made by an accumulation to bound model. My answer is yes, although the question of whether the current level of confidence is appropriate will often be an easy question. Even if the question is hard and the process of accumulation slow, this may be an example of an accumulation to bound process that is not consciously accessible.
3.4 Psychological libertarianism

Concluding deliberation despite ongoing risk of error is possible if an agent’s willingness to accept a higher risk-of-error is implemented via a lower confidence-based decision-threshold. This implies that when an agent makes a difficult choice, she is not asymmetrically biasing the decision towards one outcome or another, nor is she increasing confidence. She is merely stopping the deliberative process, and forcing herself to decide now despite low confidence. When the decision threshold is lowered one option will win, not because it is boosted by agential powers, but because it was slightly ahead, and its small lead is now deemed to be enough to be declared the winner. In the example above, no amount of threshold shifting will cause option B to win the race now. This provides further support for the analysis of chapters 7 and 8. Whilst changes in risk-tolerance can determine when an agent will make the choice, only motives can determine what an agent will choose to do.

The directionless agential power to affect decision timing rather than decision outcome may feel a little underwhelming to some readers, but it does have the benefit that it mitigates any Melean concerns of inefficient and purposeless
intentions. In the proposed model the prior intention is not an intention to decide to A, but merely an intention to conclude deliberation. The latter is trivially unproblematic.

I will return to this discussion in section 6 but will now sketch a preliminary answer to the question of why resolving indecision is effortful. Section 5 will provide a more detailed argument based on contemporary theories of cognitive effort. If I can provide a satisfactory answer, then the phenomenological gap with respect to the act of making a difficult choice is closed.

3.5 Feelings of effort

Making a difficult choice involves the mind-directed intentional act of concluding the deliberative process despite ongoing risk of error. The agent forms an intention to conclude deliberation, not because she no longer values increasing confidence, but because the benefit of reducing the risk of error is outweighed by the cost of continued deliberation. By valuing high confidence, she is motivated to prolong deliberation, but this is outweighed by the motivation to stop. Therefore, whilst executing her intention to conclude deliberation, the agent must inhibit her ongoing motivation to prolong deliberation. What makes difficult choices hard ‘are the attempts to inhibit’ (Vierkant 2022, p144) the desire to consider the question further. As anyone who has felt temptation knows, inhibiting or resisting an ongoing motive is effortful. This applies to resisting a ‘lust of the mind’, to use Hobbes’ (1997, p44) evocative description of curiosity, just as it does to resisting physical desires.

This suggests a new definition of X, where the subordinate clause is updated:

\[ X = \text{Conclude deliberation now, despite ongoing motivation to deliberate.} \]
Although this is consistent with the previous proposal, it more clearly reveals why X-ing is effortful. It is also more precise. If the motivation to deliberate and the risk-of-error ever diverge, then effort is felt in the case of motivation without risk-of-error, and not in the case of risk-of-error without motivation. I will continue to use the shortened version, with this updated subordinate clause assumed but unstated.

But there is a possible concern here. In section 1.1 I stated that cognitive effort and effort of will should be considered separately as they sit on opposite sides of the moment of intention formation. Am I now undermining this distinction? The answer is that cognitive effort and effort of will sit either side of the intention to A, but there is now an additional and earlier intention to X. Cognitive effort is prior to the lower-level intention to A but subsequent to the higher-level intention to conclude deliberation. With respect to the intention to A the effort of making a difficult decision is cognitive effort, but with respect to the prior, metacognitive, intention to X, it looks (and feels) like effort of will.

Similarly, as the quote from Vierkant indicates, merely deciding to conclude deliberation is not enough. For any kind of stopping, once an agent has decided to stop, that is not the end of the matter. She must then try to stop, and trying to do something can be hard. An intention, which is a commitment to a plan of action, is a decision joined to a trying, and ‘trying always mobilizes effort’ (Kriegel 2015, p90). In the case of making a difficult choice, effort must be mobilized to inhibit the ongoing motive to increase confidence.

As part of his wider argument for conative phenomenology as an irreducible phenomenological primitive, Kriegel concludes that ‘conative phenomenology is in the first instance a phenomenology of deciding-cum-trying’ (ibid, p94). There is much
of interest in Kriegel’s detailed analysis but the relevant point here is that the intention to conclude deliberation involves deciding-cum-trying, with its associated phenomenology, just as does the more run of the mill intention to A.

This framework also helps to set aside, for the purposes of this chapter, a concern about whether it is always possible to conclude deliberation and resolve indecision. I can remain neutral with regards this question because it is well understood that effortful trying does not guarantee success. The important question for this chapter is not whether deciding-cum-trying to conclude deliberation is always successful, but what is it that motivates this kind of deciding-cum-trying, and why is it effortful?

The arguments of this section suggest that from both a computational and phenomenological perspective, the metacognitive decision to conclude deliberation, shares many similarities with an ordinary decision. Both follow from an evaluation of relevant pros and cons, and can be effortful if the cons exert an ongoing motivational force which must be inhibited. This suggests that making a difficult choice is an action, but also that one might be able to gain insights into metacognitive stopping from studying other stopping behaviour in nature.

In the following section I aim to show that the insights from models of physical stopping behaviour in foraging, as studied within the field of ethology, are directly applicable to the metacognitive stopping behaviour necessary for the making of a difficult choice. As well as supporting my central claim that evaluation M culminates in the act of making a difficult choice, this will also explain its phylogenetic
development. As a result of this analysis, I will be able to propose a revised model of decision making.

4. A Revised Model of Decision Making

Ethologist, Tinbergen (1963) argued that in studying behaviour, one must address questions in four different areas: mechanism, adaptive significance, phylogenetic development and ontogenetic development. Taken together, section 3 will allow us to approach understanding of those four areas as they relate to a specific kind of human behaviour, the making of a difficult choice.

4.1. Stopping and Prolonging the Physical Activity of Foraging

Animals forage to satisfy their energy needs. When their current nutrient state deviates from a desired state, foraging action is initiated. A simplified negative feedback model would suggest that, in the absence of any constraints, animals should continue to forage until their current state of satiation is equal to their desired state. This is represented in the flow diagram in Figure 13.

143 ‘Neuroanatomical continuity is taken as a guiding principle’ (Cisek, 2019, p2267).
Figure 13. Flow chart showing foraging behaviour in the absence of constraints. Rust coloured boxes indicate monitoring and yellow boxes indicate acts of control. Blue boxes indicate evaluations and grey boxes indicate inputs that are temporally or physically ‘external’.

As can be seen, monitoring of the current internal state (rust coloured box) is distinct from control over the state, achieved through stopping or prolonging foraging behaviour (yellow boxes). The evaluation (blue box) is a simple comparison between the observed and desired state. The feedback loop creates an organic circuit (Dewey, 1896) where control follows monitoring, in the sense that feelings of hunger are the relevant variable input into goal driven behaviour, but also monitoring follows control, in the sense that feelings of hunger indicate the state subsequent to foraging actions.

However, for most animals this model is too simple. By taking into consideration constraints and risks, the picture becomes more complicated and the stopping and prolonging behaviour more sophisticated.

Laland (2017) describes the behaviour of two types of stickleback fish. Nine-spined sticklebacks have small spines and thin lateral plates making them vulnerable to predators. Three-spined sticklebacks, on the other hand, have three large dorsal spines as well as tough lateral plates, which makes them much safer from predators
than their cousins. Due to this vulnerability difference, three-spined sticklebacks spend longer exploring the environment and sampling food patches through trial and error, whereas nine-spined sticklebacks are much more often found hiding from potential predators. The modulation of time spent on energy gathering activity, such as foraging, with respect to vulnerability to harm is an obviously adaptive trait.

Even in environments without imminent risk, foraging behaviour is regulated, as first detailed in Charnov’s Marginal Value Theorem (1976). This theorem, which has been shown to be consistent with the behaviour of ‘worms, bees, wasps, spiders, fish, birds, seals, monkeys and human subsistence foragers’ (Gazzaniga et al., 2014, p525), states that an organism will change foraging location if the rate of change of energy in the current location is lower than the expected rate of return of energy from a competing foraging location. Here stopping foraging in a certain location is not based on predation risk, but on the more sophisticated optimization of efficient foraging across multiple sites by incorporating opportunity cost computations. Being inefficient is still harmful in the long run, if not necessarily in the short run.

Animals are able to naturally combine these marginal value evaluations with assessments of vulnerability. For example, ‘the bluegill is able to assess changes in both foraging profitability and predation risk’ (Werner et al. 1983, p1545).

Figure 14 captures the added complexity. The evaluation now captures the full cost-benefit analysis of prolonging foraging, including the current state of satiation, and the constraints not captured by the simple model: foraging profitability, predation risk and opportunity cost.
Figure 14. Flow chart showing foraging behaviour in a constrained environment where profitability must be incorporated. Colouring as in figure 13.

It is uncontroversial to state that foraging involves both the physical act of foraging and the evaluation, innate or learned, conscious or unconscious, of when to stop foraging. I will assume that if E was the physical act of foraging, then readers would accept that for many animals the evaluation of whether to stop or prolong E-ing exists. I am simply arguing that for humans at least, the stopping and prolonging evaluation applies not just to physical acts but also to mental acts such as deliberation. This, I call metacognitive stopping and prolonging.

4.2. Procedural Metacognition in Animals

Metacognition is higher-order cognition, which includes the series of processes that monitor and control cognitive activity. Whilst metacognitive monitoring refers to the subjective assessment of one’s own cognitive processes and knowledge,
metacognitive control refers to the processes that regulate cognitive processes and behaviour.

Metacognition can be divided into analytic and procedural (Dokic 2012; Proust 2013). The former is reliant on meta-representation and involves explicit beliefs about first-order mental states and processes. Procedural metacognition, however, does not require meta-representation. It can be revealed through noetic feelings attached to cognitive processes, such as a feeling of knowing when considering a question. Unlike analytic metacognition, evidence suggests that procedural metacognitive monitoring exists in some non-human animals such as macaques.

Zakrzewski et al. (2014, following Smith 2003) conducted an experiment where participants had to indicate whether a screen presented to them was densely or sparsely populated. Human and macaque participants accumulated food tokens for correct answers but lost all their tokens on an incorrect answer. Alternatively, before making the dense/sparse decision, participants could press a third button that allowed them to cash-out their chips. Inbuilt time delays on cash-out ensured that the optimal behaviour was to cash-out only if the decision was unclear. Strikingly, macaques behaved even more optimally than humans\textsuperscript{144}, becoming more likely to cash-out as the stakes grew higher and more accumulated tokens were at risk. The two macaques averaged 82% accuracy when one coin was at risk, but this increased to 98% when 7 coins were at risk. They cashed-out when stakes were high unless the task was very easy. The authors concluded that macaques ‘monitor their uncertainty about the present trial and determine whether their probability of answering correctly justifies answering to build the reward cache higher without

\textsuperscript{144} Suboptimal human behaviour has been explained by overconfidence and/or perceiving that uncertainty responses are signs of weakness.
cashing-out’ (2014, p14). Cashing-out is physical stopping behaviour, driven by the interaction between the metacognitive awareness of the risk of harm, as measured by potential token loss.

This experiment suggests that sophisticated animals such as macaques monitor decision confidence and incorporate this into evaluations that drive physical actions, such as cashing out. This ‘feeling of not knowing’ has been interpreted as ‘the first tender shoots of metacognition’ (Vierkant 2022, p150). I will now turn to humans who are able to perform mental as well as physical actions. This allows them to not only monitor and evaluate, but also to intentionally control the cognitive process. Macaques may be able to engage in procedural metacognitive monitoring but humans are able to exert full blown analytic metacognitive control.

4.3. Human Metacognition and Study Time

There is a significant literature on metacognition, both procedural and analytic, in humans. Relevant to the current proposal, which focuses on decision timing rather than decision outcome, is research into impacts of metacognition on allocations of study time.

Since the late 1990s, models of the connection between metacognition and study time in adults\textsuperscript{145} have become increasingly sophisticated. Initial discrepancy reduction models (Dunlosky and Hetzog, 1998) suggested that adults continue to study until a perceived level of learning meets a desired level of learning. This mimics the simplified model of foraging (figure 13) and the standard model of

\textsuperscript{145} In children, the ability to monitor and report judgements-of-learning develops in advance of the ability to use that information to optimally control studying (Metcalf and Finn, 2013). Ontogeny parallels phylogeny.
decision making (figure 11). More recently, however, it has been shown that study time is modulated by opportunity costs (Metcalfe and Kornell, 2005). First, if faced with multiple problems, individuals will allocate time not to the hardest items, but to the easiest as-yet-unlearned items - picking the low hanging fruit. Second, adults use ‘their judgements of the rate of learning’ (ibid, p465) to decide when to stop. When making decisions, timing is ‘modulated by a context-dependent urgency signal’ where the relevant context is the rate of evidence accumulation (Parés-Pujolràs et al. 2021, p8). This additional rate-of-change complexity mirrors that shown for foraging in figure 14, and the analysis of the rate of change of uncertainty in chapter 8.

Indeed, Metcalfe and Jacobs suggest that study-time theory ‘could benefit from study of the more complex, but also more realistic, persistence equations used in optimal foraging theory’ (2010, p218).

Following Metcalfe and Jacobs’ recommendation, figure 15 is a suggested flow diagram for the cognitive process of making a decision, inspired by the earlier discussion of foraging. Foraging is replaced by what I have called evaluation E, while the metacognitive evaluation that incorporates urgency is the previously defined evaluation M.

I will call this the revised model of decision making, to contrast it with the standard model.
Figure 15. Flow chart showing the cognitive process involved in making a decision. Rust coloured boxes indicate metacognitive monitoring and yellow boxes indicate metacognitive control. Blue boxes indicate evaluations and grey boxes indicate inputs that are temporally or physically ‘external’. Examples of cognitive strategies include embodied, extended, and social cognition as well as implementation of the norms of reasoning.

Taken together, the above analysis suggests that stopping and prolonging evaluations are ubiquitous in nature. This is because in a constrained environment it is adaptively beneficial to allocate the time spent on any task appropriately, bearing in mind risk of harm, the available resources and opportunity costs. As Hayden et al. state, ‘deciding when to leave a depleting resource to exploit another is a fundamental problem for all decision makers’ (2011, p933). Given this, it should not be surprising that the same principles are applied to deciding when to stop deliberation\footnote{This is supported by neurobiological findings which ‘lead to the compelling conclusion that the same cognitive and neural processes underlie much of human behaviour involving cognitive search – in both}. 

\footnote{146}
In line with the previous discussion regarding foraging, the cost-benefit analysis of whether to stop or prolong deliberation is informed by metacognitive monitoring of the rate of change of confidence as well as the level of confidence. How long a decision-making process would take in a world without constraints can diverge from how long that process should take when constraints are incorporated. There can be very hard decisions that must be taken quickly or easier decisions that should be made slowly. The guiding principle however, consistent with evolutionary fitness, is the appropriate allocation of a scarce resource.

In the case of simple animals, foraging time is subject to complex stopping and prolonging behaviour, but there is no metacognition. More sophisticated animals such as macaques seem not only to be aware of their own level of uncertainty but able to incorporate this in decision making. However, this metacognitive awareness is procedural rather than analytic. Finally, humans are able to conceptualise a feeling of confidence as such and represent it to themselves or others. This is fully fledged meta-representation. Importantly for this chapter, humans are also able to engage in explicit metacognitive control via the intentional mental acts of stopping or prolonging cognition.

Having proposed a revised a model of human decision, and explored its ‘phylogenetic refinement’ (Cisek 2019, p2265), I will now analyse how this revised model relates to contemporary theories of effort.

external [foraging] and internal [deliberation] environments’ (Hills et al., 2015, 47). Even more specifically, recent experiments in neuroscience indicate that both foraging (Hayden et al, 2011, Li et al 2012) and cognitive control (Shenhav et al, 2013) are associated with activity in the anterior cingulate cortex.
5. Cognitive Effort and The Symmetry Argument

Making a difficult decision can feel effortful, but what makes it so has hitherto remained unclear. There has been much recent debate in psychology and philosophy of mind on the topic of effort, and I presented my unified model in chapter 4. Whilst broader discussions consider effortful actions in general, here I will focus purely on the cognitive effort associated with mental actions.

In recent years, theories of mental effort have shifted from a focus on resource depletion towards a focus on cognitive resource allocation.

Resource depletion models understood will-power as a finite resource that is consumed by intentional mental acts, and mental effort as a measurement of the amount of the resource consumed. Originally an ego depletion model was hypothesised which suggested that mental effort tracked calorific depletion in the brain (Masicampo and Baumeister, 2008). However, this received sustained criticism, particularly from Kurzban (2010, 2013, 2016) who offered a rival theory, proposing that mental effort is the experiential manifestation of the calculated opportunity costs associated with a mental task (see also Székely and Michael, 2020; Sripada, 2020). I will call this the opportunity-cost theory.

A recent paper (Andre, Audiffren and Baumeister, 2019) suggests a similar model which I will call the integrated theory. The authors propose a neural network called the Mechanism of Effort which integrates constraints, perceived costs and benefits, and information relating to the current state of the organism. This mechanism then outputs ‘decisions regarding the intensity and the direction of the engagement in effort in ongoing or future tasks’ (p4), and a feeling of effort, that corresponds to the awareness of the perceived costs associated with achieving the goal of the task.
The discussion of the cost-benefit computation, evaluation M, in section 3 focussed on the mind-directed action that follows, but the perceived cost of foregoing alternative actions is an integral part of the computation. Although evaluation M leads to a binary decision, to stop or prolong deliberation, the costs associated with that decision can be anywhere on a continuum from very low to almost equal to the benefit. The opportunity-cost or integrated theories, claim that (for prolonging of cognition, at least) rather than being discarded, this more fine-grained information is revealed to the agent in the form of an analogue experience of effort. The same cost-benefit computation, or part thereof, that controls action is parsimoniously employed to modulate the intensity of the phenomenological experience:

According to the opportunity cost view, when systems that can be used for multiple purposes are engaged in a task, the potential benefit of ending the present task in order to perform some other task is computed. This computation is the opportunity cost of persisting in whatever it is that one is doing (Kurzban 2016). The sensation of mental effort is the output of mechanisms designed to measure the opportunity cost of engaging in the current mental task (Kurzban et al. 2013). This implies that, for effortful prolonging of cognition in particular, the computation that informs evaluation M leads to two connected outcomes; (i) the decision to prolong deliberation and; (ii) the feeling of cognitive effort, which is a function of the opportunity costs associated with prolonging. In the language of chapter 4, the feeling of cognitive effort represents-as-costly prolonging deliberation.

If contemporary theories of effort support the existence of evaluation M in the context of effortfully prolonging cognition, they open up the possibility of symmetrically applying the same model to explain the central question of this chapter: What it is that makes concluding deliberation effortful?
Metacognitive stopping is an act of cognitive control that follows from evaluation \( M \), just as is the prolonging of deliberation in the face of distractions. Cost-benefit computations are neutral with regards the direction of control. As Kurzban says, ‘according to the opportunity cost view …… When this cost is sufficiently great, outweighing the computation of the (potentially long-term) benefits of persisting the task is abandoned.’ (2016, p68-69). A benefit of stopping deliberation is an opportunity cost of prolonging deliberation, and vice-versa. The same mechanism that outputs a feeling of effort if one action is chosen in the face of opportunity costs, will symmetrically output a feeling of effort if the alternative action is chosen in the face of opportunity costs.

Accounts such as Kurzban’s focus on the effort of prolonging deliberation. However, the existence of effortful choice suggests a symmetry between metacognitive stopping and prolonging, consistent with that seen in physical stopping and prolonging and in cost-benefit computations more generally. By assuming symmetry in evaluation \( M \) one can explain the felt effort associated with making a difficult choice. Without this symmetry, difficult decisions would be unexplained, effortless or impossible.

Assuming symmetry one can state that:

When making a difficult choice, the computation that informs evaluation \( M \) leads to two connected outcomes; (i) the decision to stop deliberating and make a choice, and; (ii) the feeling that stopping deliberation is effortful, which is a function of the opportunity cost associated with stopping. The feeling of cognitive effort represents-as-costly concluding deliberation.
How can something as seemingly cold as the opportunity cost of an evaluation could be linked to feelings of effort? My suggestion is that, for the idea of a cost-benefit computation to be more than a metaphor, an agent’s cognitive infrastructure must include an action-focused common unit on which the computation can be performed. And for a decision to be an agent’s decision, that common-currency must be a contextually relevant manifestation of her own values. When the question of what to do arises, relevant values are activated. Values in support of an action will be benefits, those against the action or in support of an incompatible action will be costs\(^{147}\).

An agent is motivated to achieve a state of affairs consistent with her values. In the case of effortfully concluding deliberation, the opportunity costs in evaluation M reveal that the agent is motivated to reduce the risk of error. Concluding deliberation is effortful when there is conative value in reducing the risk of error and that value is foregone if deliberation is halted. Contrarily, if there is a risk of error but the agent’s cognitive infrastructure assigns no conative value to reducing that risk, then she will not be motivated to deliberate, and concluding deliberation will be effortless. What is felt when making a difficult decision is the conative phenomenology of a trying (to conclude deliberation) being resisted (by the ongoing motivation to deliberate).

Whilst a lack of confidence indicates the existence of a risk of error, whether the agent pursues actions to reduce this risk, is dependent upon whether she is most motivated to do so. Epistemic actions like any other actions are governed by the motivational strength idea. This aligns with the intuition that an agent is more motivated to reduce the risk of error in high stakes decisions than in low stakes decisions.

---

\(^{147}\) See Berkman et al’s ‘Self-Control as Value-Based Choice’ (2017) for a similar view.
decisions. Stephen will and should spend more cognitive resources on choosing between universities than choosing between flavours of ice-cream. It is also consistent with the technical harm-from-error based model, and the direct competition between reducing epistemic risk and pragmatic risk, proposed in the field of predictive processing. Ultimately, an agent aims to maximise care for herself by reducing the perceived risk of harm, be it pragmatic or epistemic.

Having said that, the activation of conative values is complicated, fallible, and at times unpredictable. An observer, or even the agent herself, may struggle to see the link between her motivation to deliberate further and ultimate self-care. Just as with ordinary practical motives, she may feel that her desire to deliberate further is inappropriate, as in incessant checking (Friedman, 2019), or she may be uncertain as to whether it really has value. Nguyen’s motivational inversion and Csikszentmihalyi’s autotelic activity are examples of conative value being assigned to exploratory behaviour and increasing confidence, in a way that may sometimes lead to long term benefits, but at other times will lead to time and financial resources being “wasted” on computer games.

The proposal of this chapter is neutral with respect to the objective value of further deliberation, but insistent that the question of whether to conclude or prolong deliberation is resolved in a motivational contest just like more practical questions. This is why making a difficult choice is an action. The standard account of decision making, taken at face value, would imply that the need to increase confidence until it reaches a natural threshold is a sub-personal uninhibitable response, which is functionally isolated from the rest of an agent’s cognitive architecture. I suggest that this view is both functionally and phenomenologically unsupported.
The opportunity-cost theory explains not only why an action is effortful, but also the strength of the feeling. The amount of effort felt in concluding deliberation is a function of the conative value of further deliberation\textsuperscript{148}. The hardest choices are ones such as Sarte’s case of the young man torn between joining the Resistance and caring for his mother\textsuperscript{149}. Here the stakes were so high and confidence in the decision so low that concluding deliberation was seemingly impossible. As the subjective value of further deliberation declines, stopping deliberation becomes less effortful until, at the extreme, a so-called passive decision feels effortless, because it has no opportunity costs. The lack of a feeling of effort in making an easy decision represents-as-costless concluding deliberation. I therefore assume that human cognitive architecture implements the revised model instead of, rather than alongside, the standard model.

In section 4, I proposed a revised model of human decision making that incorporated metacognitive stopping and prolonging. In this section I have provided further support for the model by showing it makes sense of the relation between metacognitive control and the phenomenology of effort. Contemporary theories of effort suggest that the experience of mental effort and the decision to prolonging deliberation are be explained by the same metacognitive motivational contest. Applying this finding symmetrically to metacognitive stopping explains the phenomenology of an effortful choice.

\textsuperscript{148} I recommend further research into the link between the rate of change of uncertainty and feelings of effort. My expectation is that a low rate of change increases the likelihood of stopping deliberation, but does not decrease the value of uncertainty reduction, so feelings of effort remain. However, a high rate of change would reduce the opportunity costs of prolonging. This might explain the state of flow where prolonged cognition is effortless or even pleasant.

\textsuperscript{149} Sartre also seems to endorse a value-based approach: ‘The value of his feeling for his mother was determined precisely by the fact that he was standing by her’ (1946, p7).
The analysis suggests that the proposed model is not only supported by functional and phenomenological analysis, but is also evolutionarily plausible. I will therefore conclude that it is at least roughly correct, and turn to the implications for agency.

6. Implications for Agency: An argument against libertarianism

If the account set out in this chapter is correct, then the phenomenological gap with respect to the act of making a difficult choice has been closed.

Making a difficult choice is an action because it is driven by the intention to conclude deliberation, despite ongoing motivation to deliberate. This intention is the culmination of a process that evaluates the costs and benefits of further deliberation. Making a difficult choice feels effortful because it involves opportunity costs, equal to the conative value of further deliberation, and those opportunity costs are felt. The feeling of effort represents-as-costly the act of concluding deliberation now.

This conclusion indicates that when an agent effortfully makes a difficult choice she is neither asymmetrically biasing the decision towards one outcome or another nor increasing confidence. Therefore, the feeling of effort in the act of making the decision does not undermine the motivational strength idea, nor does it justify beliefs in psychological or anti-naturalist libertarianism. Rather than indicating an ability to decide otherwise it reveals an ability to decide now. When the decision threshold is shifted lower one option will win, not because it is boosted by agential powers, but because it was slightly ahead, and its small lead is now enough to be declared the winner. The experience of effort may be an experience of the self as a source of force, but it is the force to resist the hunger for information rather than an ability to
‘rise above the desires to which one is subject, and to take the reins in one’s own hands’ (Wallace, 1999, p234).

A libertarian might respond that, although this description is functionally coherent and compatible with the conative phenomenology, it does not address the cognitive phenomenology. After all, the agent still feels unconfident.

Here, it is worth returning to the difference between the standard and the revised model. In the standard model, an agent who feels low confidence is necessarily within the withholding zone between thresholds. She cannot make a decision until confidence is increased. She is compelled to satiate her hunger for information. For an agent subject to such constraints, low confidence implies a future which includes more than one possible decision, because it is a future that necessarily includes further deliberation. In the revised model, however, the link between low confidence and the requirement to deliberate is broken. When an agent actively makes a decision, she severs the link between confidence and psychological openness, collapsing the probability distribution by removing the opportunity for further evidence gathering or evaluation. The ability to make a difficult decision is therefore both an opening and a closing of opportunities. It opens the opportunity to turn attention elsewhere, but shuts down alternative possible resolutions to the current deliberation. But that is ok. The decision about whether to close deliberation is made based on (metacognitive) conative values, which aim to align behaviour with care-based norms. An agent will only conclude deliberation if she is most motivated to do so. In the standard model a feeling of lack-of-confidence implies that the agent’s psychological state is consistent with more than one future decision, but in the revised model it merely implies an unsatiated hunger for more information, a hunger that can be resisted just as a dieter can resist eating.
Conclusion

Chapter 10

The Surprising Importance of Inhibition

“Inhibition is therefore not an occasional accident; it is an essential and unremitting element of our cerebral life.” (James, 1890/1983, p1185-6)

In this chapter, I review the findings of the thesis through the lens of inhibition. I conclude that the inhibitability of motives is central to what it means, and how it feels, to be a human agent.

Human agents flexibly navigate a complex landscape of risks and opportunities, guided by subjective, context-dependent, representations of conative value. They are motivated towards ends of value, and when making a choice, will always choose to do what they are must motivated to do. However, all motives are inhibitable. No matter how strong an activated desire, it is always possible that the desired action-plan will be rejected if the agent is more motivated to do otherwise, or delayed if deliberation is incomplete.

There are two major differences between doxastic and conative attitudes. The first is their direction of fit. Conative attitudes move an agent to act on the world, whereas doxastic attitudes are moved by observations of the world. The second difference is that, whilst doxastic attitudes are subject to the coherence requirement, conative attitudes can persist in the face of conflict. When two directly conflicting motives persist within a single agent, they ready her to implement two inconsistent actions.
But implementing two inconsistent actions simultaneously is impossible, so conative architecture must facilitate control of the transmission from readiness to action.

When an agent is torn between two motives, she decides between them through a motivational contest. To manage all competing motives while the contest is ongoing, and any persistent unchosen motives once a decision has been made, conative architecture must inhibit unchosen motives. And for this to be possible, motives must be inhibitable. If motives were reflex-responses that compelled action then neither the withholding of a decision whilst deliberation is ongoing, nor the making of a decision in accordance with the motivational strength idea, would be possible.

To be clear, the fact that motives are inhibitable does not support beliefs in a Cartesian inhibitor who can intervene in the conative landscape. All motives are inhibited until the question of what to do is settled. And once the question is settled, inhibition is only removed from the strongest motive\(^{150}\).

Once a decision has been made, inhibition is lifted and an intention is formed. An intention is therefore an uninhibited motive. But do intentions also have some additional features, such as intrinsic stability? Many philosophers follow Bratman in believing that they do. However, the analysis of chapter 2 suggested that intention stability is a function of extrinsic factors rather than anything intrinsic-to-the-attitude. I therefore concluded that an intention is merely an uninhibited motive.

As Bratman has famously detailed, intentions are different from other conative attitudes in that they are subject to ‘consistency constraints’ (1999, p31), just like

\(^{150}\) Following Mele, I assume the definition of settled is ‘thin’ (mele, 1992, p168), in the sense that it need not follow from deliberation or any other extended mental action of settling-upon. In that case there is no uncertainty, no conflict and the only motive is, by definition, the strongest motive.
doxastic attitudes. But this should not be surprising. The winner-takes-all model, implemented via inhibitory gates, is an incoherence filter. Whilst conflict can persist amongst motives, it cannot persist amongst the subset of motives that are uninhibited. Bratman’s demand for coherence is satisfied by the inhibition of unchosen motives.

In chapters 3 and 4, I analysed conative phenomenology and found that this too is dependent on inhibition:

With respect to the conative phenomenology of motivational pull, I found that the correspondence relationship between the phenomenological intensity and functional strength of motives is a conditional relationship, and that the necessary condition is inhibition: The phenomenological intensity of a motive is a representation of its functional strength if and only if it is inhibited.

Feelings of effort are central to the phenomenology of agency (Bayne and Levy 2006) or of control (Pacherie 2008; Lukitsch 2020). My analysis of the conative phenomenology of effort mirrors that of motivational pull. While the positive pull of a motive indicates that an action-plan is attractive, the aversive feeling of effort reveals that an action is controversial or unsustainable or costly. And what is the measure of costliness? It is the strength of the (inhibited) motive to do otherwise. If the strength of a motive towards an action represents its expected value, then the strength of a motive against the action represents its (opportunity) cost. The feeling of effort only arises when an agent is performing some action whilst simultaneously being motivated to do otherwise. And it is the strength of the latter, unchosen and inhibited motive, that drives the phenomenology of effort.
Whilst functional analysis celebrates the victory of the strongest motive, phenomenological analysis reveals the strength of its weaker, inhibited cousin. This inversion encourages a new perspective on the phenomenology of agency. It is not the experience of implemented intentions, but of frustrated desires and of (as yet) unfulfilled plans. Humans are flexible agents, motivated but not compelled by inhibitable conative forces. The conative phenomenology(s), of motivational pull and effort, evidence this lack of compulsion.

In addition, the fact that motives can be inhibited allows an agent to delay practical action and engage in extended deliberation until she has reached an appropriate level of confidence in her decision. For human agents, who have metacognitive abilities, extended deliberation can include not just non-intentional processes such as evaluation and pattern recognition, but also intentional managerial actions. These actions come in three kinds: Evidence gathering actions, (self-)control actions and the act of stopping or prolonging deliberation.

Chapter 9 was an account of the (effortful) act of making a difficult decision. The inhibitability of motives was doubly important for this account. First, managerial actions are only possible because practical motives can be inhibited during an extended period of deliberation, and second, making a decision is only an act because further deliberation is a motivated rather than a compelled response to uncertainty. The motivation to reduce uncertainty can be defeated (and therefore inhibited) by a stronger motive to conclude deliberation now. Decisions can be made despite ongoing uncertainty only because the motive to reduce uncertainty is inhibitable.
Finally, in his analysis of intentional agency, Bayne describes intentional actions as ‘actions that are carried out by agents themselves, and not by some sub-personal or homuncular component of the agent’ (2013, p162). For Bayne the differentiating factor between person-level actions and sub-personal responses is the level of integration within the broader cognitive economy. An action that is ascribed to the person herself must be motivated by an attitude that is cognitively integrated ‘in a flexible and appropriate manner’ (ibid, 164). This difference between person-level actions and sub-personal responses can be cached out in terms of inhibitability. A knee-jerk response, once triggered, is uninhibitable, so can never be flexibly integrated. The agent, as a person, can have no influence over the reflex. Motives, on the other hand are always inhibitable. This inhibitability is what underpins the flexible integration of motives within the cognitive economy, and ensures that motivated actions are intentional actions.

Human agency, from both a functional and a phenomenological perspective, is dependent upon inhibitable motives.
Chapter 11
Experience-Compatibilism

A simple model of decision-making

In this thesis my aim was to develop a functionally accurate and experience-compatible model of a decision-making agent. To do so, I focused my analysis on two psychological phenomena at the core of decision making - motives and levels of confidence - both of which integrate the functional and the phenomenological.

I suggested the following simple model:

*When an agent makes a choice, she always chooses to do what she is most motivated to do*\(^{151}\) (*The motivational strength idea*).

*An agent makes a choice if and only if she is sufficiently confident in her decision. Until then, she deliberates about what to do* (*Confidence*).

A prerequisite of the simple model was the inhibitability of motives. All considered motives are inhibited during deliberation and only the chosen action-plan is released from inhibition once the question is settled. An intention is merely an uninhibited motive.

Three potential problems

\(^{151}\) More formally: When making a choice between A-ing and ¬(A-ing), the agent will choose to A if, at the moment the choice is made, the strength of her motivation towards A-ing is greater than the strength of her motivation towards ¬(A-ing).
Motives and levels of confidence are felt, the former through the conative phenomenology(s) of motivational pull and effort, and the latter through the cognitive phenomenology(s) of levels of confidence and the rate of change of confidence. This raised the enticing possibility that, for those who reject an error theory of phenomenology, introspection and fine-grained phenomenological analysis could confirm or refute the simple model.

At first glance, refutation seemed the most likely outcome.

Three potential problems for the simple model immediately arose: First, phenomenologically intense motives can be resisted by opposing motives that don’t feel as strong. This is the problem of relative strength. Second, an agent seems to be able to actively shift her own motivational balance. This is the problem of intentional synchronic self-control. Third, many philosophers believe that decisions can be made actively by an agent, even in situations of low confidence. This is the problem of active decisions. All three potential problems raise the possibility that something other than motivational strength can (sometimes) bias a decision.

Although it is possible to deny the existence of intentional synchronic self-control and active decisions, this seems inconsistent with the phenomenology of effort. If the feeling of effort is a marker of intentional action, then agents can intentionally shift their motivational balance and actively make decisions. And if feelings of uncertainty about what to do imply psychological openness to alternative decisions, then perhaps the ability to actively conclude deliberation implies an agential power to overrule one’s strongest motivation and decide otherwise.

Thus, the conative and cognitive phenomenology of the decision-making process seem, pre-theoretically at least, to undermine the simple functional model and
support libertarianism. To ascertain whether this intuition is correct, I engaged in a fine-grained analysis of conative and cognitive phenomenology. I concluded that it is not.

**Conative phenomenology**

I first analysed the conative phenomenology of motivational pull. I found that the phenomenological intensity of a motive is a non-conceptual representation of its functional strength if and only if it is inhibited. The problem of relative strength vanished, replaced by three compatible statements:

1. When making a choice, an agent always choses to do what she is most motivated to do (The motivational strength idea)
2. The phenomenological intensity of a motive is a non-conceptual representation of its functional strength if, and only if, it is inhibited (Conditional Correspondence)
3. The phenomenological intensity of an unchosen motive can be greater than that of a chosen motive. (Experience)

Second, I analysed the conative phenomenology of effort. This revealed that the feeling of effort is the feeling of a trying being resisted. It is resisted by the motive to do otherwise, to stop trying. This is consistent with the motivational strength idea and confirms that the feeling of effort is a marker of intentional action. More specifically, it implies that there are two motives for each feeling of effort, one towards and one against the chosen action. For each feeling of effort, philosophers of mind should aim to identify those motives.
The simple model is made possible by inhibition but, perhaps surprisingly, inhibition is also key to understanding conative phenomenology. Feelings of motivational pull only refer to inhibited motives and feelings of effort only refer to tryings which are laterally inhibiting the motivation to do otherwise. The intensity of both feelings reflects the motivational strength of the inhibited motive.

The simple, but hierarchical, model

In addition to the phenomenological analysis, I engaged in a predominantly functional analysis of intentional synchronic self-control, and defended the idea of a hierarchical but undivided mind. My argument was based upon the idea that metacognitive attitudes exist and can motivate “managerial” control actions during ongoing practical deliberation. These metacognitive motives, which are attitudes towards attitudes, compete in motivational contests and are bound by the motivational strength idea.

The simple model can be updated to make this hierarchical element explicit:

- **When an agent makes a choice, she always chooses to do what she is most motivated to do (The motivational strength idea).**

- **An agent makes a choice if and only if she is sufficiently confident in her decision. Until then, she deliberates about what to do (Confidence).**

- **Whilst deliberating an agent can engage in managerial actions, if and only if she is most motivated to do so (Managerial control).**

In situations of intentional synchronic self-control, the agent faces two questions simultaneously: the practical question of whether to engage in a virtuous or wayward
action, and the metacognitive question of whether or not to exert control. The metacognitive motive towards exerting control is (usually) only relevant to the latter question. Therefore, if an agent’s metacognitive motive towards control is strong enough, she can choose to implement a control action to shift her own motivational balance, even if she is simultaneously most motivated towards a “wayward” answer to the practical question. This is true even under the assumption of full instrumental rationality. Intentional synchronic self-control is not a problem for the hierarchical simple model.

Finally, I turned to the cognitive and conative phenomenology of making a difficult decision.

**Low confidence decisions**

After analysing uncertainty and levels of confidence from a functional and phenomenological perspective, I concluded that although an agent can be uncertain, what she *feels* is confidence or a lack thereof. Uncertainty is not an all-or-nothing variable within cognitive architecture, but is the result of an interaction between the agent’s level of confidence in a decision, which increases as her risk of error decreases, and a decision-threshold. An agent is uncertain about what to do if her level of confidence is below the threshold.

If an agent responds to low confidence by engaging in deliberation, then she will gather new evidence and, because she is evidence-responsive, this may shift her eventual decision. A deliberating agent is psychologically open to alternative decisions, but that openness is contingent upon further deliberation. When she concludes deliberation, the agent severs the link between low confidence and
psychological openness, collapsing the probability distribution by removing the opportunity for further evidence accumulation and integration. Cognitive phenomenology does not justify the belief in alternative possibilities at the moment of choice.

Perhaps surprisingly, this argument was strengthened by my analysis of conative phenomenology. Feelings of effort imply a trying and a motive against that trying. In the case of a difficult decision the agent is trying to conclude deliberation now, despite ongoing motivation to deliberate. The fact that she is still motivated to deliberate indicates that she has not reached full confidence in her decision. This, in turn, implies that the decision-threshold is not fixed. The motive to conclude deliberation now, if successful, shifts the decision-threshold down, until it crosses the current level of confidence. At that point the current level of confidence becomes sufficient, given the context. This shift, to a context-dependent, rather than fixed, definition of “sufficient” has the intuitive implication that deliberation is a flexible, motivated agential action rather than an irresistible reflex response. Unconfident agents may be hungry for more information but that hunger can be resisted, just as a dieter can resist eating.

The fact that the decision-threshold is lower does not give the agent an ability to choose an option other than the one that is most conatively attractive. To see why, imagine two sprinters running a 100m race. A referee who moves the finish line to 60m may indirectly impact who wins the race (if one competitor is a specialist in the final 40m, for example) but she cannot change who is in the lead at the 60m point. Wherever the finish line ends up, the winner is the athlete who touches the finish line first. Similarly, whilst the interaction between levels of confidence and the threshold determines when an agent will make the choice, motives determine what an agent
will choose to do. The feeling of effort indicates an agential power to resist the hunger for information, which affects decision timing, rather than an agential power to bias the decision outcome.

Despite initial appearances, not one of the three potential problems represented an actual problem for the simple model, which I present in its final form:

*When an agent makes a choice, she always chooses to do what she is most motivated to do (The motivational strength idea).*

*An agent makes a choice if and only if she is sufficiently confident, given the context, in her decision. Until then, she deliberates about what to do (Confidence).*

*Whilst deliberating an agent can engage in managerial actions, if and only if she is most motivated to do so (Managerial control).*

**Metaphysical implications**

I have proposed and defended a simple, but hierarchical, theory of agency, and have done so by focusing almost entirely on inhibitable motives and levels of confidence. The final model is experience-compatible and evolutionarily plausible. The presented account confirms that interactions of inhibitable motives are governed by the motivational strength idea, and solves a number of outstanding puzzles within philosophy of mind. It does this without recourse to distinct intending attitudes, motivational partitions, reservoirs of willpower, Cartesian homunculi, or error theory. There is just a metacognitive agent, who does what she most wants to do.
Through functional analysis, I have shown that intentions are merely uninhibited motives; intentional synchronic self-control is possible; and making a difficult decision is an action. Through phenomenological analysis I have been able to reveal the relationship between the phenomenological intensity and functional strength of motives; provide a novel unified theory of the feeling of effort; and elucidate the complex and shifting cognitive phenomenology(s) experienced throughout extended deliberation.

Although I believe that these findings are valuable in their own right, they are most powerful when considered together, as they leave no phenomenological gap which would support beliefs in super-natural powers. The final model is experience-compatible. Whilst naturalism alone was an unpersuasive defeater to libertarian-seemings, naturalism combined with a fine-grained account of conative and cognitive phenomenology is a persuasive defeater. An understanding of conative phenomenology supports the motivational strength idea, and an understanding of cognitive phenomenology reveals that feelings of low confidence do not indicate the physical or psychological possibility of doing otherwise. Even an unconfident agent who effortfully resolves indecision, will always choose to do what she is most motivated to do.

By closing the phenomenological gap, I have removed seemings-based justification for libertarian views, psychological or anti-naturalist. The analysis may have been functional and phenomenological, but the implications are metaphysical.

---

152 I call a defeater unpersuasive if it leaves the question “If p is false, then why does p seem to be true?” unanswered. Whether libertarians are actually persuaded is another matter.
References


Burnston, D. (In submission). Nagelian Reduction of the Concept of 'Decision'.


Mill, John Stuart (Robson, J.M. ed.), 1872 (1979), *An Examination of Sir William Hamilton's Philosophy*, University of Toronto Press, Toronto


Sripada, C. (manuscript). Human Agent Architecture Has a Valuationist Structure


Helmholtz, H. V. (1948). Concerning the perceptions in general, 1867.


