Rationalism versus Sentimentalism: Why not both?

By Vu Dinh Toan (Daniel) Nguyen

Bachelor of Politics, Philosophy, and Economics, La Trobe University, 2019

Thesis submitted as total fulfillment of the requirements for
the degree of Master of Arts (by research)

College of Arts, Social Sciences and Commerce

School of Humanities and Social Sciences

Department of Politics, Media, and Philosophy

La Trobe University, Victoria, Australia

August 2021

Table of contents

Table of figures

Abstract

In this thesis, I will formulate the Affective Reflection Model (ARM), a descriptive model of moral decision-making to demonstrate how automatic affective processes (the having of emotions) and controlled deliberative processes (reflective reasoning) can function in a mutually inclusive way, by revisiting arguments made in moral philosophy, philosophy of mind and moral psychology through the lens of the ARM. Conceiving the functions of our capacities in this way, I argue, will give an adequate explanation of how a morally significant demand of our judgments – the expression of our agency – is satisfied. In demonstrating this mutually inclusive way of functioning between the two kinds of mental processes mentioned, I aim to provide an alternative way to conceive of their moral significance; which historically has been portrayed in an antagonistic sense, that is, each kind of process has some moral significant function, and, therefore, moral judgments ought to be guided by either one of them in spite of the other.

Statement of authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

Vu Dinh Toan (Daniel) Nguyen

31 August 2021

This work was supported by La Trobe Graduate Research Scholarship, and a La Trobe University Full-Fee Research Scholarship.

Acknowledgements

To complete a thesis of this level, during a global pandemic, as a person with English as his second language, is no simple task. I would not have been able to achieve this without the guidance and love from my support network. For this reason, I want to take a moment to acknowledge the people who have helped me.

I want to acknowledge the support I have from La Trobe University. The financial and bureaucratic support from the University has provided a much-needed source of security during this time.

I want to express my gratitude for the guidance and support from my supervisors, Doctor Richard Heersmink and Doctor Mary Walker. Your expertise and knowledge were a constant source of guidance, challenge, affirmation, and inspiration for me throughout this project. But more than my teachers and supervisors, you are also my confidants and friends who offered stories, experiences, and advice that helped me push through. Words cannot describe the sense of gratitude I feel. I am truly lucky to have you both as my supervisors.

I would also like to extend my gratitude to our staff at La Trobe's Department of Politics, Media, and Philosophy; especially Doctor George Vassilacopoulos, Doctor Toula Nicolacopoulos, Doctor Yuri Cath, and Doctor Nicholas Barry. Without your valuable advice, many challenges along the way would not have been possible to overcome.

And on a more personal note, I would like to thank my family and friends. Philosophy is not a well-understood field in Vietnamese culture. It took an immense amount of faith and unconditional love for my parents and brothers to support my decision to pursue this path and provide constant assurance. For this, I owe you everything.

Finally, I would like to recognise the support I have from my friends; especially Grant Richardson, James Tran, Craig Currie, Anthony Gagliano, Sam Mensforth, and Nicholas Bega. Whether it was by lending an open ear, challenging me on my ideas, giving me a fresh perspective or reassuring me, you have helped me in more ways than one in this thesis' completion.

1. Introduction

In this thesis, I will formulate the Affective Reflection Model (ARM), a descriptive model of moral decision-making. The model will demonstrate how automatic affective processes (the having of emotions) and controlled deliberative processes (reflective reasoning) can function in a mutually inclusive way, by revisiting arguments made in moral philosophy, philosophy of mind and moral psychology through the lens of the ARM. This way of functioning, I argue, will give rise to a morally significant demand of our judgments – the expression of our agency, brought to attention in modern discourse by Kennett and Fine (2009). In demonstrating this mutually inclusive way of functioning between the two kinds of mental processes mentioned, I aim to provide an alternative way to conceive of their moral significance; which historically has been portrayed in an antagonistic sense, that is, each kind of process has some moral significant function, and, therefore, moral judgments ought to be guided by either our emotion or our reasoning.

This thesis is composed of nine chapters, with **Chapter One** being this introduction and **Chapter Nine** being the conclusion. In **Chapter Two**, I will briefly present the extended history of the Rationalism versus Sentimentalism debate. Arguments on whether morality is guided by our emotions or our capacity for rational thinking dates back to Plato's idea that the ideal society is one lead by rational minds, and the Epicurean idea that pleasure, regulated by prudential reasoning is the marker of a good life (Baltzly 2019, Plato 2003). For reasons of scope of this thesis, however, it is practical to examine this tradition with David Hume as a starting point. Hume saw that passions (emotions, desires) are pivotal in our moral judgments in two regards: moral judgments imply motivation, which necessitate the having of passions; and secondly, passions are what we use to determine what is virtuous and what is vicious, something Hume takes reasoning, which pertains to matters of fact, to be ultimately incapable of.

The counterargument to Hume, by his greatest critic of the modern period, Immanuel Kant, shaped the continuation of the Rationalism versus Sentimentalism tradition in modern discourses. Kant exerted that for moral appraisals of a person's character to be valid, these appraisals must be made on actions which agents committed to with their own free will (Rohlf 2018). And a necessary condition for having and acting on a free will is that our will is not under control by some force foreign to us. For Kant, our human desires are a source of control – sometimes we succumb to our desires even if doing so warrants moral disapproval from others. So, to be a free agent, one must act according to one's deliberative choice and not one's desires. This tradition of aligning moral demands with either emotions or reasoning, and using those demands to discredit the other capacity, is echoed through modern arguments: in Smith's (1987) claim that motivations are not intrinsic in moral judgments, therefore emotions are unnecessary in moral thinking; Shafer-Landau's (2003) defence against Mackie's (1977) argument against moral realism; and Green's (2009) argument that emotions latch on to morally insignificant factors which can skew our judgments. Notions of motivations and self-control from Hume and Kant's era are used throughout these arguments.

Then comes the involvement of psychologists who re-conceptualise emotions as cognitive processes – much like reflective reasoning. But unlike reflective reasoning, emotions are automatic and visceral, explaining why they almost always occur prior to our reasoning. This shift in the conceptualisation of emotions will be presented in **Chapter Three** through three studies: the first is a formalisation of emotions as cognitive processes, proposed by psychologists like Paul Ekman (1999). Then Antonio Damasio's (1994) observation of patients with damages to their Ventromedial Prefrontal Cortex (abbreviated as vmPFC) shows the dependency of cognitive processes on affective responses. These two studies established that the line between cognitive and affective capabilities in human beings is not as clear as the philosophical debate has sometimes assumed. Howard Margolis (1987) put emotions and rational thinking, in terms of automatic affective responses and conscious

reflective deliberation, in one continuous process. This constitutes two significant changes to the tradition of Rationalism versus Sentimentalism – the "rebranding" of emotions and reasoning as *automatic affective processes* and *controlled deliberative processes*; and the arguments switched their focus to the necessity (or otherwise) of the controlled, reflective process.

This development prompted modern hybrid accounts of moral decision-making. We will examine three arguments in this line of thinking in **Chapter Four** - Jonathan Haidt's (2001) Social Intuitionist Model (SIM), Jeanette Kennett and Cordelia Fine's (2009) argument for the necessity of rational reflective thinking for the expression of agency, and Karen Jones' (2018) Trajectory-Dependent Model of (Human) Rational Agency. These theses incorporate both capacities to some extent. Thus, while I take each theory to advance the debate in important ways, in these hybrid models, scholars are still largely thinking in terms of the rationalist versus sentimentalist tradition.

In studies mentioned so far, I believe Sentimentalists and Rationalists are both right and wrong. They are both right in being able to capture essential roles that our mental capacities play in our moral decision-making. They are wrong in still tending to see these processes as mutually exclusive, or either one ought to have moral primacy in our judgment; despite the progress made by cognitivist approaches to the emotions. For example, Kennett and Fine's account is too general in the role of reasoning in expressing agency, thereby failing to properly include the function of emotion; while Jones' and Haidt's accounts are too quick to dismiss the benefit of reflective reasoning. Theses from Hume and Kant's era have evolved, but the conflictive conception of Sentimentalism and Rationalism still persists. So, to advance the discussion further, I aim to demonstrate how these cognitive processes work together in moral decision-making in a way that is not mutually exclusive, and that they are capable of being mutually beneficial to one another. This is the focal point of my main argument, which leads to the building of the ARM.

To develop the ARM, the concepts of mental attitudes, i.e., desires and beliefs, will be our building blocks. These concepts will be examined at length in **Chapter Five** – including their core differences, classifications, and their own notions of strength. These concepts are pivotal to the examination of the functions of emotions and reflective reasoning in **Chapter Six**. Emotions are seen here as outputs of an automatic process, which signify changes in our attitude system (Reisenzein 2009). From this description of emotion, I will draw the basic model of motivation. Then, the reflective reasoning process and the ARM will be set up in **Chapter Seven.** Reasoning examines the rationality of our attitudes – of whether, and to what extent, our beliefs are logically consistent, our desires are simultaneously satisfactory, and our procedural beliefs can satisfy our desires (Bratman 1987, De Sousa 1987). From this, I will present how rational examination processes function within the basic model of motivation.

Then, I will take into account what I call the non-monogamous relationship between causes and effects to draw the ARM. When we make a practical judgment, we are making a judgment on the means we take to attain some ends (Paul 2015). In this process, our attitudes are examined rationally to enable agents to produce true-to-self emotional reactions, which then become inputs for rational decision-making processes. Reasoning, at the same time, enables identification of the different potential means to our ends and their likely consequences, which can prompt further emotional reactions. These processes are akin to a feedback loop, or a Circular Cumulative Causation process – mutually dependent components interact and cooperate in their functioning (Argyrous and Stilwell 2011). Exercising rational reflective thinking enhances our emotional reactions, and better emotional reactions produce better input for rational reflective thinking.

Through the ARM, I will present two inferences important to the role of emotions and reasoning in moral decision-making in **Chapter Eight**. The first notion is the distinction between justificatory and reflective reasoning, the first being reasoning for a preconceived judgment or egotistical needs, and the latter being

reasoning in service of curiosity and the pursuit of truth. This distinction is often not clear in Sentimentalist accounts, which leads them to conflate the function of these reasoning processes. In separating them, I will proceed to show how the practice of reflective reasoning can expand an agent's awareness of their actions, of the effects that their actions cause. In being aware of an effect that their action causes, and decided to carry out the act, agents show that they intended the effect mentioned (Sinhababu 2013). Intention is one's expression of one's agency in their judgment. This leads to my second point, that the expansion of their awareness, which enables more affective responses by agents, increases the sense of agency is increased in their final judgment. This, I will argue, is the role of reflective reasoning, in conjunction with emotions, in the expression of agency in one's judgment.

In presenting how the expression of agency, a morally significant demand in one's judgment, is jointly caused by both one's affective responses and one's proper practice of reflective reasoning, I hope to pivot the Rationalism versus Sentimentalism tradition from seeing the role of these mental processes in moral decision-making in an antagonistic sense to a mutually inclusive and collaborative sense.

2. Rationalism vs. Sentimentalism

In this chapter, I will present a brief history of the rationalism-sentimentalism debate. It is important to provide a context for this thesis by presenting the historical development of the reason-sentiment dichotomy in moral philosophy. This will start with Hume's argument for moral sentimentalism from motivation in 2.1, and Kant's deontological ethics in 2.2. Then, in 2.3, I will show how this dichotomy has persisted in modern moral thinking, as well as how modern thinkers have deepened this division. This will include its effect on the development of the 'trolley problem', as well as additional charges against emotions' role in moral thinking advanced by rationalists. In addressing this history, I aim to highlight key points on both sides of this debate that will be relevant to the making of my model.

2.1. The Humean critique of reasoning from motivation

I will discuss the origin of this tension between moral rationalism and moral sentimentalism in this section. Moral judgments – those that separate right from wrong, good from evil, virtues from vices – have been one of philosophy's oldest topics of interest. In the many approaches to tackling this topic, theses which used human faculties, i.e., thinking and feeling, to explain the making of these moral judgments, and by extension, what makes up moral judgments, or how we know which judgment is right and which is wrong, have generated a large body of work. We can find this type of argument in works as early as Plato's *Republic*, where the idea that reasoning can open our eyes to moral truths first took shape; in Stoic philosophy, with its focus on reason as the source of virtue; or differently, in the Epicurean idea that pleasure, regulated by prudential reasoning, is the marker of a good life (Baltzly 2019).

For this thesis, however, the starting point I choose is to examine David Hume and his most famous critic, Immanuel Kant. David Hume's critique of the role of reasoning in moral judgments is perhaps the challenge that lit the spark to ignite the

debates that followed. During Hume's time, people regarded morality as a pivotal factor in the peace of society; and thought that our moral decisions affect us unilaterally, but cannot, at the same time, support two opposing ideas. Because of this, emotions are inappropriate to guide moral principles, as people can have wildly different emotional reactions on the same subject, e.g., abortions. The answer to what can guide morality, then, falls onto our reasoning.

However, Hume disagrees with this idea. He asserts that an integral part of a moral judgment is its ability to motivate agents to act. He understands moral judgment as a type of normative judgment that differentiates good from evil, determining what is virtuous and what is vicious. I agree with Hume in this definition of moral judgments. For him, this definition implies that motivation is an integral part of a moral judgment. If I am sincere in saying "stealing is wrong", this necessarily means I am motivated to refrain myself from stealing. In asserting a moral judgment as such, I am also expressing a passion that I hate or dislike the act of stealing, or else my judgment would be insincere. Furthermore, Hume claims it is only through my passions that I can recognise moral right from wrong (Hume 2009, p. 713):

> *Take any action allowed to be vicious: Wilful murder, for instance. Examine it in all lights, and see if you can find that matter of fact, or real existence, which you call vice. In which-ever way you take it, you find only certain passions, motives, volitions and thoughts. There is no other matter of fact in the case. [...] It lies in yourself, not in the object.*

Morality, for Hume, is an essential *human* thing, and not a fact of nature which can be subjected to inquiries which concern facts of this kind. By using the judgment "wilful murder is an immoral act" as a paradigm case, which few find problematic, Hume claims that it is our passion, i.e., disgust, that tells us it is immoral. Examinations on facts about this case, e.g., whether the person died, what was the weapon used, etc., can trigger a "sentiment of disapprobation" or approval about the

fact, but these would still be our feelings, not intrinsic to the fact. (ibid) Therefore, it is my passions that show what is morally right or wrong.

But Hume does not consider every emotion to be a moral emotion. We can put actions and their outcomes in the general dichotomy of being beneficial and harmful, which excite different kinds of pleasure and pain, whether it be in the acting, receiving or witnessing of them. Morality, in particular, involves the excitation of four types of emotions – moral actions cause agents to take *pride* in their conduct and inspire *love* in others, while immoral actions cause *humility* when observed from oneself, and inspire *hatred* when observed from others (ibid, p. 717). By feeling prideful in their actions, agents know they are morally good, and by loving someone's action, others know it is morally good. Hume claimed it is only via the arousal of these emotions that we can know which action is virtuous and which is vicious.

And because motivation is such an integral part of morality, Hume thinks reasoning is irrelevant to moral judgments. According to Hume, "Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular" (ibid, p. 698). Besides the direct link between emotions and motivation presented above, Hume provides two additional arguments in his critique against the centrality of reasoning in moral judgments. The first is that in no type of reasoning, conceived separately from passion, can we find relations to motivation. Reasoning, to Hume, is "the discovery of truth or falsehood [...and] consists in an agreement or disagreement either to the real relations of ideas, or to real existence and matter of fact" (ibid, p. 4). Here, he differentiates between two types of reasoning, probable reasoning, regarding matters of fact, and demonstrative reasoning, regarding relations of ideas like causes and effects[1]. Hume's claim is that the truth, or falsehood, as determined by reasoning, has no bearing on one's motivation alone.

[1] I am content with Hume's short description of reasoning here. I will discuss reflective reasoning further, using ideas from Michael Bratman and Ronald De Sousa, in the development of my model in chapter Seven.

Whether I have the motivation to smack my brother with a bat is not solely determined by whether it is true that a bat is harder than bones or that smacking him in the face will disfigure him. Including my passion towards (or against) the smacking, or its effects, is absolutely necessary in determining my motivation.

Hume makes his second argument against the centrality of reasoning in moral judgments by pointing out the insignificance of false reasoning in moral accountability. With the example above, let us suppose that I thought the bat I have is a plastic, hollow toy bat, which is neither harder than bones, nor will it hurt others. Now suppose further that I was wrong, and my smacking indeed caused by brother pain. Hume claimed that this kind of mistake is of an innocent kind, and often taken as not related to a person's moral character. Mistakes of this kind often inspire grief and not blame towards moral characteristics (ibid, p. 583). Therefore, mistakes in reasoning do not contribute to the task of separation between good (virtuous) and bad (vicious) people, outside of excluding said mistakes from the task. Because mistakes in reasoning are morally irrelevant, reasoning itself is morally irrelevant.

2.2. The Kantian idea of free and rational agency

To Kant, on the other hand, rational thinking is required in the making of a moral judgment. He conceived that morality necessitates an agent to follow a set of rules wilfully, made by their own rational mind, and necessarily void of any material desires (Johnson and Cureton 2021).

First, to understand Kant's argument better, I believe that we must understand how he draws the connection between moral rules and persons. To Kant, moral rules, which determine what we ought to do, concern themselves with the realisation of an ideal world; being in which is what he sees as the highest good (Rohlf 2018). In making a moral judgment, we are making a normative judgment which aims to bring about the ideal world. Analogously, to be a good person is to act in order to realise one's ideal self, one's self-conception – the highest good *to a person*. Kant's ideas

fluctuate between these two senses of good-in-general and good-to-persons quite often, so it is beneficial to keep this analogy in mind as we proceed.

To Kant, free will is a necessary condition in the making of moral judgment, and holding morally accountable. Like Hume, Kant also sees that we subject not all of our conduct to moral appraisals. Hume saw mistakes in reasoning, which cause agents to commit actions that result in others' suffering unknowingly, not to be constitutive of agent's moral characteristics; simply because the agent in question did not want to cause others' suffering. But Kant also recognises that, sometimes, we do not scrutinise others' moral characters even if they want to cause suffering, e.g., when they had no other choice. Kant asserts that, for a moral appraisal of an agent's actions and character to be valid, it is implied that the agent in question, at the time of acting, was acting on their own free will (Rohlf 2018). In defending myself from an assailant, I wanted their suffering – either by a knife or a gun – instead of my own. To Kant, this does not warrant scrutiny on our moral character either, simply because our will in these cases is not free. Rather, it is in virtue of the wilfulness in committing murder, to re-use Hume's example above, that the action is conceived as immoral. So, to be subject to moral appraisal, a person must express their free will in their actions; the person must have *chosen* to act that way (ibid, p. 41).

The sense of freedom that Kant advocates for is of the negative kind, but not merely in the negative "free from" sense. In the expression of a person's free will, Kant considers the removal of, that is, being free from, any constraint to be necessary. If we are being constrained by factors outside ourselves, e.g., being forced to cause suffering with self-defence, our will is not free. But further than that, Kant thinks agents are also constrained internally – by their own desires, because we ultimately do not choose what we want. And the exercising of an agent's free will here requires that an agent can choose otherwise, which is not true if we passively follow our desires. Thus, it is the resistance of this, to choose deliberately what to act on, regardless of our wanting, that is the ultimate expression of free will. So, to Kant, agents do not express moral autonomy in being unrestrained, but in being able to

constrain themselves and their actions in principles of their own choosing (Johnson and Cureton 2021).

It is through our rational mind that we can identify and commit ourselves to these principles. We know how to act by deriving descriptive principles (maxims) from prescriptive principles (imperatives). For example, I have a means-ends belief that if I want food cheaply, I should cook to save on cost. Now suppose that I became hungry and therefore want food; I will then choose to act on the maxim "go cook something to satiate my hunger" which is rationally compatible with my imperative, in the realisation of my desire, instead of other maxims which are not compatible, like going out to buy food. In this sense, our desires set the ends of our practical thinking, by which we can apply the appropriate imperative; and, in turn, the imperative became a standard of rationality by which we choose maxims to act on.

But for Kant, these kinds of principles are unfit to describe moral principles (ibid, p. 18). Moral principles, as he perceives, cannot be conditional. In expressing "killing is wrong", we are committing to the view that under no circumstances is killing is right, not that only under some circumstances is killing wrong. This prompts him to distinguish between conditional, or material principles, which have desires as their ends, and those void of desires, which are formal principles (Rohlf 2018). To act with material principles is to be constrained by one's desire, and that, for Kant, implies a constraint that invalidates a person's free will.

Formal principles, which Kant called categorical imperatives, are the standard of rationality that we can use to respond to our desires. Desires provide reasons for agents to act; but to be truly free in their actions, one must rationally respond to their desires and not merely follow them (Johnson and Cureton 2021). It is because of self-control, i.e., the application of our reflective thinking, to either reject or allow these desires to move us, that we are free agents. To control ourselves, a standard of rationality that is void of any desire is required. If we respond to a desire using principles derived from another desire, we are simply following another reason, and not actually responding to a reason, hence our will is not free. Only formal principles,

which do not employ any desires, are suitable to be used to respond to reasons. For a will to be truly free, then, one must rationally choose how to act using formal principles. Hence, to reflect rationally on maxims by which we act is a requirement in moral thinking, as an expression of a free will.

An example of such categorical imperatives is Kant's formula of the Universal Law of Nature (Johnson and Cureton 2021). According to this, an action is morally permissible when its maxim can be universally applicable, at least to rational agents. So, to see if an action is morally permissible, agents must follow four steps, which I will present here with Kant's example[2]:

1. Formulate a maxim for your action – I ran out of money, so I go out to borrow some and promise to pay it back, with no intention to keep that promise.
2. Reform said maxim as a universal law that all rational agents must abide by – if anyone ran out of money, they could always borrow with a promise and without a commitment to keep such promise.
3. Conceive whether you can execute your original maxim in a world in which such a law is universal – It would be impossible for me to borrow money by a promise as no one would lend to me since they know my promise is empty.
4. See if you can will yourself to act in such a way, should it become a universal law – In this example, it fails on the third step so no further examination is needed.

Only through this rigorous, rational examination of a maxim, to either reject or approve it, do we exercise our freedom as a rational agent. It is only when we are free that we are morally blameworthy for our actions. Therefore, reasoning for Kant is an absolutely necessary process in moral decision-making.

[2] Here I simplify the language of Kant's example; to see Kant's original example of how these steps applied, as well as further discussions, see Rivera-Castro (2014).

Here, Kant raises an important point about moral judgments that is problematic in Hume's thesis, which is the agency that enables moral judgment. In Hume's thesis, he conceives of specific emotions (pride, love, humility, hatred) that indicate moral judgments, but he does not specify what kind of desires are responsible for these emotions. We can, in fact, feel these moral emotions from desires which are imposed on us – e.g., feeling shame about being gay because we were taught that being gay is undesirable.

However, I do not consider that Kant's approach here is entirely sensible either. In his formulation, he conceives of a condition to an ideal world, which is the freedom from control of the desires. But I cannot seem to be able to conceive of an ideal world without it being a desirable world, and in Kant's formulation of the Universal Law of Nature, desires are implicit as well. In his example, the third step fails because it was self-contradicting; but it is implied that I *want* to borrow money successfully. In changing what I want, this step can be successful. If what I want is to sew distrust in people, e.g., the maxim in the first step is "I borrow money with an empty promise to sew distrust in people", it is conceivably successful in all four steps; but we can hardly call this a morally permissible action.

2.3. Modern continuation

The ideas discussed above are found in many contemporary discussions on which mental faculty is responsible for moral judgments. One contemporary use of this dichotomy is Michael Smith's (1987) argument for a dispositional conception of desires, which says that having a desire means that the subject has the disposition to act in a certain way, rather than causing actions directly. This idea is Humean because, as Smith points out, Hume is aware of two kinds of passions: violent passions and calm passions. While the direct causal link between desires and actions can be explained in cases of violent passions, that task falls short for calm passions, which are "more known by their effects than by their immediate sensations" (Hume 2009, p.530). In this light, violent passions are desires like hunger, or pain – those

which have immediate phenomenal content. Calm passions are more akin to preferences, which have little to no phenomenal content. From this distinction, it follows that the connection between desires and motivations is a contingent connection. To Smith, having a desire does not mean that you are motivated to act in its satisfaction, but you are disposed to act in its satisfaction. Conceptualising desires as dispositions, which do not necessarily propose a direct causal link to action, can stress the essence of desire in both its violent and calm forms.

Using this, Smith inquires into the motivational character of moral judgments. Smith (1987) conceives of judgments as a different type of mental state from beliefs or desires: "Suppose a subject now accepts the judgement [...] This judgement may properly be thought of as the expression of a belief that the subject presently has" (p. 59). I think here, Smith intends to differentiate between motivation and judgment: while motivations express an agent's desires, normative judgments express both their beliefs and desires. For example, if I judge that "killing is evil", and assuming that I am sincere in this judgment, it expresses my belief that killing is morally bad, and that I also desire to avoid killing, thus I am motivated to restrain from killing. For Smith, the addition of a belief is important here to explain why our motivation tracks our judgment. When my normative judgment changes to "killing for self-defence is acceptable", my motivation to restrain from killing is reduced where I need to exert self-defence; in other words, moral motivations reliably track moral judgments.[3]

[3] These issues are also discussed at length in the literature on internalism versus externalism about moral motivation. I will leave this literature aside in this thesis.

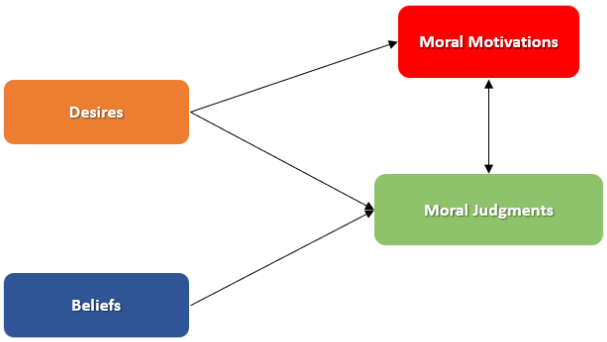

Figure 1 - Smith's separation between judgments and motivation

Smith's argument is interesting because it separates normative judgments from motivation. Smith conceives of motivation separately from justification – he regards moral judgment as requiring justification in a moral point of view, regardless of agent's motivations (p. 39). Smith's separation of judgment and motivation presents a way out of Hume's problem of motivation for moral rationalists. If they reject the Humean idea that motivation is not a part of the reasoning process, they must explain the motivational feature that moral judgments have. But, if they only claim the justificatory aspect of the judgement as a moral requirement, they can avoid having to deal with explaining the motivational aspect of moral judgment altogether. This transposes the problem of motivation to the problem of defining the relationship between moral judgments and moral motivation.

This new structure works in rationalists' favour. It is considerably easier to claim a contingent connection between moral judgment and moral motivation than to go against Hume's charge and take on explaining motivation from reasoning.[4] It is intuitive to think of a situation where normative judgments and motivation come apart, e.g., even if one judges that A is a good thing to do, one would still be

[4] There are counter arguments from internalists in this matter – such as questioning how can one honestly judge that telling lies is a wrong thing to do, without feeling some motivation to abstain, e.g., hesitation, from lying? My response to that would entail more conceptual accounts on perspective taking and identity – people are capable of saying something would be right, or wrong, in another person/collective's perspective, without actually having the desire or motivation needed for the judgment to be theirs. But for the scope of this thesis, I will leave this matter here.

dissuaded from doing A in favour of other desires. We can find one example of this in Plato's *The Republic (2003)*. Socrates asks Cephalus if it is right to return a weapon you borrowed from a friend, who now has become a madman planning a killing spree, even if he holds a moral belief that it is right to return what you have borrowed (Plato 2003, p. 7). Another extreme example is amoral people: some may be fully capable of judging what is the moral thing to do, yet be motivated to act the opposite way to that. This would mean that being able to judge what is right and wrong does not mean the agent will be motivated to act accordingly. Hence, the connection between moral judgments and moral motivations is contingent. This way, it is easier for rationalists to palm off the motivational features of emotions and treat motivation externally to normative judgments.

Another application of the dichotomy between rationalism and sentimentalism is by Russ Shafer-Landau (2003), in the course of defending moral realism. Moral realism, to put it simply, advances the point of view that there are moral truths in the world, and that moral statements can be assessed rationally. Kant's categorical imperative (as presented above) is one example of this. The real-ness of morality is demonstrated by the existence of a standard of rationality, by which judgments can be morally true or false. Shafer-Landau's defence of moral realism was directed towards Mackie's (1977) argument for moral anti-realism. Based on the observation that in making moral judgments, agents are motivated to act in accordance with them, Mackie's argument against moral realism can be put as:

P1: Moral realism is committed to the existence of moral entities that, upon being perceived, compel an agent to follow its (moral) demands.

P2: Such entities do not exist.

C: Moral realism is false.

Here, Mackie posits that every piece of knowledge needs a worldly phenomenon that has the authority to validate it. For example, for me to speak truly when saying "I know the coffee shop is two blocks away", the knowledge needs to be validated by the fact that the coffee shop is two blocks away, or my statement will be false. For moral truths to be true, Mackie asks for the same type of validation: what worldly phenomenon, person, or entity, can validate the truth of moral judgments? It is easy for religions to refer to their holy books and gods to be that entity. But in a modern, secular society, not to mention discrepancies between religions, these holy books can hardly serve as objective anchors for moral truths. Mackie's answer is simple: there is none, and therefore, no moral truth is verifiable, hence, morality is not real.

Shafer-Landau (2003) defends moral realism by arguing that Mackie's P1 is wrong; that moral realists do not have to commit to such entities to explain the motivational aspect of moral judgment. Instead, he points to the agent's motivational profile, or their motives, that are responsible for their motivations, moral or not. He raises the fact that if a person does not have any motive, being presented with a fact will not move them at all. For example, knowing cheating on a test is an unethical thing to do would not compel me to not cheat without me having a higher desire to do the moral thing over passing the exam. On a more negative note, if a person has an immoral motive, being presented with a moral fact will not dissuade them. Say, if A really wants to kill B, and assuming A's hostility towards B is not impulsive because B raped and killed A's sister, who A loves very much, and got away with it, then presenting the fact that killing is morally wrong would not dissuade A from killing B, if A has the chance to[5]. Hence, similarly to Smith's idea above, moral motivation can be conceived independently of moral facts, and moral facts can be determined without referring to a motivational function. We can judge an action as a right thing

[5] One would be right to point out that this is a very Humean argument. But Shafer-Landau uses it to advance moral realism (and by extension, moral rationalism) by coupling it with Smith's separation between judgment and motivation above.

to do without requiring a motivational aspect to be imbedded in said judgment. A judgement about what is right is only motivating if the agent themselves wants to do the right thing – some people know that they are causing suffering, that what they do is "the wrong thing", and they are motivated to do so because they want to cause suffering. Therefore, moral judgments, which work on moral facts, have a contingent connection with moral motivation.

Another application of this dichotomy between rationalism and sentimentalism is Joshua Greene et. al.'s (2009) argument against sentimentalism, using the popular thought experiment created by Phillipa Foot (1967), dubbed "the trolley problem". Traditionally, this thought experiment is closely associated with theses in utilitarianism, which examine the idea that morally good actions aim for the highest amount of happiness. But here, we are looking at the role of our emotions and rational thinking in defining what "the highest amount of happiness" means. Specifically in this section, we are looking at Greene's use of this thought experiment to reject sentimentalism.

What initiated the thought experiment was the problems surrounding abortion. Foot wanted to show that the reason we think abortion is problematic is because of the intention behind the act, not the act itself. From this, she provided many other thought experiments surrounding the theme of 'permissible sacrifice' to show the differences in intentions, and how morally blameworthy we find them to be. A coupled thought experiment was deployed for this purpose, which were later labelled the 'Court Case' and the 'Trolley Problem' :

Suppose that a judge or magistrate is faced with rioters demanding that a culprit be found for a certain crime and threatening otherwise to take their own bloody revenge on a particular section of the community. The real culprit being unknown, the judge sees himself as able to prevent the bloodshed only by framing some innocent person and having him executed. [...] To make the parallel as close as possible it may rather be supposed that he is the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one

track and one man on the other; anyone on the track he enters is bound to be killed. In the case of the riots the mob has five hostages, so that in both the exchange is supposed to be one man's life for the lives of five. The question is why we should say, without hesitation, that the driver should steer for the less occupied track, while most of us would be appalled at the idea that the innocent man could be framed.

Foot wants to show that intentions matter in moral judgments. *Vis* killing, there is a difference between *direct intentions,* i.e., the Court Case, and *oblique intentions*, i.e., the Trolley Problem[6]. Foot noted that there are different types of effects which one action, or a course of actions, can produce and they can be classified and differentiated according to this distinction in intention. This differentiation she calls the Doctrine of Double Effect (following Aquinas, and others). She applies this doctrine to draw a parallel between different levels of moral accountability and intentions in her attempt to solve moral dilemmas, such as the permissibility of abortion.

The trolley problem caught the attention of many scholars, but it has not kept Foot's original intention. Based on Foot's mind experiment, Rationalists and Sentimentalists have developed many other variations to further their arguments. Sentimentalists come up with variants that show how emotional connections to the victim can affect one's decisions, while rationalists claim that these effects are harmful to moral judgments. The two most recognised ones are the 'Footbridge' and 'the Doctor' variant. In the Footbridge variant, an agent must choose whether he or she wants to push a big person off a footbridge to stop the trolley. In the Doctor variant, they decide whether they would dissect a healthy patient to harvest his organs to save 5 dying patients, or to let the healthy patient go and allow the 5 to will. These variants were designed to assess what morally irrelevant factors can influence the agent's actions, not to assess what are people's inner intentions as Foot's original experiments did.

---

[6] This is an interesting distinction, which will be relevant in chapter Eight.

Using Foot's example, Greene et al. (2009) deploy a counter argument to Hume's "Argument from motivation", using the Trolley problem and the Footbridge variant. Following Kauppinen (2018, p. 61), one way to characterise Greene's view is:

P1–Suppose that the deontological belief 'killing is wrong' is proximately caused by disgust (an emotional response) towards killing.

P2-The disgust response is sensitive to factors such as the use of personal force or up-close-and-personal causing of harm to others (it is less disgusting if you did not have to directly do it up close).

P3-Whether killing is executed with personal force, and up close, or not, is morally irrelevant.

P4 (P2+P3)–The deontological belief 'killing is wrong' is formed in response to morally irrelevant factors.

C (P1+P4)–The deontological belief 'killing is wrong' is epistemically unwarranted.

This argument charges sentiments in a similar manner to Hume's charges against rationality. Recall that Hume proposed that because rationality lacks a motivational component that disqualifies it as morality's core. Here, Greene and Singer posit that our sentiments can latch on to morally irrelevant considerations. The reason is that those affective responses (disgust at causing harm to others) are sensitive to irrelevant factors (the use of personal force, up-close-and-personal). This is further shown by modern moral psychological experiments, where it was found that irrelevant factors like whether there is a hand-sanitiser bottle present on a table can affect an agent's moral judgment (Helzer and Pizarro 2011). Hence, forming moral beliefs based on sentiments is unreliable.

In all three arguments considered in this sub-section, notions of motivation and self-control are used throughout. Sentimentalists' arguments often claim that,

because motivation is essentially of emotions, morality, as a normative enterprise, is also a matter of the emotions. On the other hand, Rationalists often stress the importance of using reason to control our motivations and actions. This signifies the effect of early Sentimentalist and Rationalist thinking from Hume and Kant's era on contemporary thinking. This presents the context of our discussion in this thesis – of whether emotions, which relate to one's motivation, and reasoning, which Rationalists taken to be essential to self-control, are truly incompatible. In the next chapter, I will present modern studies in psychology which reconceive emotions as fast, automatic, cognitive processes. These studies changed the conceptual compatibility between emotions and reasoning, which shaped modern models of moral decision-making, especially those that will be presented in the following chapters.

3. Emotions as cognitive processes

In the previous chapter, I have established the dichotomy between rationalism and sentimentalism as an important debate in the study of ethics since its beginning to its modern-day continuations. The dichotomy took form with the debate between Kant's deontology and Hume's sentimentalism. It then develops with the involvement of more modern moral philosophers and philosophers of mind. Based on these involvements, modern rationalists and sentimentalists have produced many additional arguments on why either human cognitive processes, like rational thinking, or emotion are practical judgments' *raison d'être*.

But as with everything else in philosophy, paradigms are always open to challenges. 20th century scholars doubt this distinction's validity in moral judgments, especially with experimental psychology emerging as a prominent scientific field. This scepticism paved way for new studies that recognise traces of both what we traditionally classify as emotional and rational in people's moral judgments. In this chapter, I will show this next step in the study of how ethical judgments are formed, the re-conceptualisation of emotion as a cognitive process.

This chapter will show this shift in the conceptualisation of emotions through three studies. The first is a formalisation of emotions as cognitive processes, proposed by psychologists like Paul Ekman (1999)[7]. Then I examine Antonio Damasio's (1994) observation on patients with damages to their Ventromedial Prefrontal Cortex (abbreviated as vmPFC), which shows the dependency of cognitive processes on affective responses. These two studies established that the line between cognitive and affective capabilities in human beings is not as clear as the philosophical debate has sometimes assumed. I will explore these studies in 3.1. Following this, in 3.2, I turn to Howard Margolis' (1987) recognition of two distinct

[7] Philosophers such as Ronald De Sousa (1987) and Justin Oakley (1992) also conceive of emotions as cognitive processes, which open themselves to rational assessment. As de Sousa's work is relevant to my model, I will discuss it further in chapter Six.

cognitive processes in an agent's mind: an automatic cognitive process, intuition, and a conscious cognitive process, rationalisation. Margolis' thesis puts emotions and rational thinking in terms of automatic affective responses and conscious reflective deliberation, which he then put into one continuous process. This prompted a movement in moral philosophy to look into how intuition and reasoning shape moral thinking together, rather than dwelling in the rational-emotional dichotomy any longer.

3.1. Emotions do have cognitive content

Major developments in psychology have helped to explain how people primarily make their judgments intuitively, that is, using both their cognitive and affective capabilities. To show this, I will use two studies.

Psychologists in the 1980s, like Ekman (1999), noticed the logic behind emotions: they are akin to an appraisal program. Emotions, as noted by psychologists, are a process whereby an agent recognises input patterns and prepares his or her brain to respond accordingly; mainly to appraise whether something is beneficial or harmful to the agent. This formulation does not take away the fact that emotions are reactive in nature. But it challenges the view that emotions are purely visceral or non-cognitive. The mind is still at work, even when people are expressing an emotive reaction: they observe, they assess, and they react accordingly. This shows that, even though emotions are instantaneous, they are still cognitive processes. So then, arguments that contrast emotions and rational thinking because emotions are not cognitive processes come under scepticism.

This scepticism deepens in 1994, when Damasio's (1994) observations on patients with damage to their vmPFC (the region of brain that sits just behind and above a person's nose bridge) were published. Essentially, when a person's vmPFC is sufficiently damaged, their emotional capability vanishes. Damasio's finding is that these patients lost the capacity to exert an emotional reaction to any situation; this can range from a simple preference, like whether they prefer strawberry or chocolate

ice-cream, to having no reaction to abhorrent images, like torture videos. Because of this, they lost their capability to make choices altogether. This finding is in direct opposition to what Rationalists believe – that, when emotions are excluded from the decision-making process, agents would make the optimal choice without being distracted by emotional responses to irrelevant factors. What Damasio's finding does show is the fact that, at least in making choices, reasoning requires affective responses as a source of information – an input to the reasoning process. With emotions being established as a cognitive process, and that rational thinking relies on affective responses, the distinction between emotion–cognition is no longer a meaningful one.

But this raises a question: is there a difference, as a cognitive process, between affective responses and rationalisation? Margolis suggests the answer is yes.

3.2. Automatic versus reflective processes

Margolis (1987) subscribes to the idea that all judgments are cognitive in nature, but he recognises two distinctive cognitive processes in play when people make decisions. The first is an *automatic* process, which is rapid and effortless, like simple pattern-matching. The second is a *controlled* process that tracks the logic of any given situation and gives a narrative, either through induction or deduction. He justifies this idea by pointing to an experiment called the Wason four-card task. In this experiment, subjects were shown four double-sided cards on a table:

Figure 2 - Wason's Four Cards task

Subjects were told that there is one rule to these cards: if there is a vowel on one side, then there is an even number on the other side (Johnson-Laird and Wason 1977). Then, subjects were told to verify this rule by flipping the least number of cards. Most people would flip the E card first. But for the second card to flip, most people would flip the "4" card, while the right answer would be the "7" card, for finding a vowel on the "7" card would violate the rule, but finding a B on the "4" card would not. The rule constrains vowels, not even numbers.

The findings of this experiment led to the conclusion that judgment and justification are separate processes (Margolis 1987). Wason found that, in cases when people went for the popular answer (E and 4) and were given the right answer (E and 7), they tend to resist the right answer to defend theirs. But when given an answer up front without telling them why, whether it be the right or the wrong (but popular) answer, people seem to be equally capable of justifying the given answer. This led Wason to believe that in everyday-life decision-making, automatic affective judgments and reflective justifications are separate processes in a person's mind.

Building on Wason's finding, Margolis puts forward a formulation of how people make their judgments, including moral judgments. He establishes a linear cognitive ladder – an evolutionary account of how cognitive capabilities developed in humans, and how it has shaped our cognitive function (Margolis 1987). There are seven steps in total (minus the eliciting situation), which are shown in this figure:

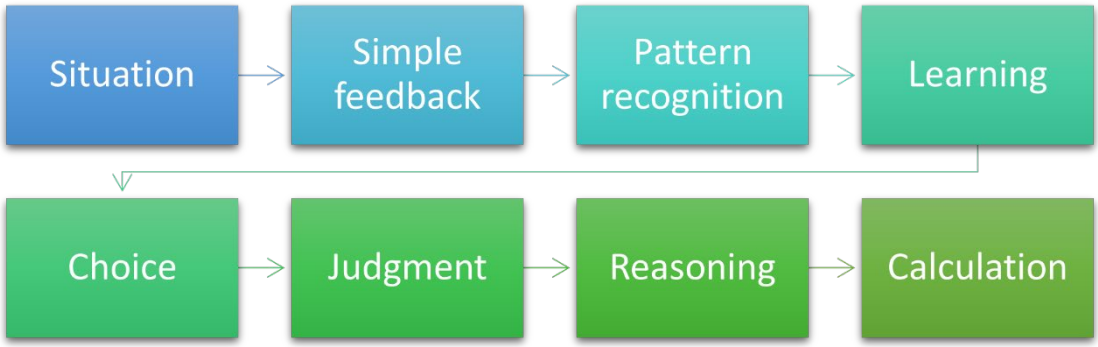

Figure 3 - Margolis' cognitive ladder

Margolis divides this ladder into two processes, split by when we have a judgment in mind. Here, he posits that emotions can fit just the first step, simple feedback, as seen when I feel pain in my hand and retract it from whatever was causing it. Or it can encompass all first four steps, such as upon seeing a person falling, reading the panic in their eyes and facial expression, having learnt that I am within reach to catch them and choosing to help. All these steps happen instantaneously as I perceive the situation – an automatic process. Margolis calls this automatic, instantaneous part 'intuition', and it happens before a judgment is formed in our mind. He claims that the primary way people make judgments is in an automatic, intuitive way that depends largely on their perception. In the second interpretation of this automatic affective process, it contains four steps, one of which is pattern recognition, which means it is a process with cognitive content; but it is not necessarily conscious. He pointed out that, in Wason's four-card task, when people choose to flip the '4' card after the E card, what they seem to do is:

> *They seem to be doing simple-minded pattern matching: There was a* ***vowel and an even number*** *in the question, so let's turn over the* ***vowel and the even number***. (Haidt 2012, emphasis added)

Reasoning, which Margolis defines as judgment *plus language*, can only come in later to interpret those perceptions and reactions, to oneself and to others. When asked to give a reason for their choices, people seem to come up with one, rather than having one before making their choices; as is shown when they resist the correct answer when given, and justify any answer given to them up front. People are far more likely to use their reasoning skills to justify their choices, not to make their choices. Margolis then concludes that rationalisations are ex post processes that follow, and therefore are constrained heavily by, initial intuitive judgments.

Since these studies, automatic affective processes and conscious reflective processes are no longer seen as incompatible. This sets up a new dynamic between the mind's capabilities, which puts the use of automatic processes and controlled processes together, in building a model that represents how people make general judgments, with moral judgments as one type of judgement. This will be the core of Haidt's model, as well as most decision-making models in moral psychology since the 1990s.

4. Hybrid Theories

This chapter will focus on how the development of psychology and neuroscience, which redefined the emotion - reasoning dichotomy by re-conceiving emotions as cognitive processes, produced hybrid theses which regard moral judgements to be formed using elements of both automatic and reflective processes: Haidt's (2001) Social Intuitionist Model, Kennett and Fine's (2009) argument for the conceptual necessity of reflective thinking in moral judgment, and Jones's (2018) Monitor Model of the role of conscious reflective deliberation and judgment in moral decision-making. I choose Haidt's model, explained in 4.1, because it is one of the most engaging studies in this new age of moral theories. It is primarily an intuitionist model, but it also focuses on the social aspect of moral judgment to explain the justificatory role of moral deliberation. Then in 4.2, I examine Kennett and Fine's work, which will present a counterargument that rational thinking, which results in our reflective endorsement of some desires but not others, is a necessary process to express agency in our actions. Finally, in 4.3, I will then discuss Jones' argument that, while actions that are reflectively endorsed can express agency, emotional actions themselves can also express agency because of past reflective processes. An important reason to include Jones' model is because it draws attention to how affective automatic processes and controlled reflective processes can intertwine with one another. Ideas in these theses bring forth the inter-connectedness between both processes, which will be essential to the development of my model in the next chapter.

4.1. Jonathan Haidt's Social Intuitionist Model

To understand Haidt's model, first we must understand that primarily, Haidt is a behavioural psychologist. For the most part, his work is empirically based. But *vis* moral philosophy, he is a sentimentalist. Haidt (2012) questions the rationalist's view that moral rules are the fruit of a person's growth in rational thinking when they

grow older. An observation he makes on cross-cultural behaviour prompts this scepticism, regarding what people deem as 'morality' in different social groups. If we were to stay consistent with the rationalist view, he questions, how are we supposed to explain the divergence in what people consider being moral rules in different cultures? For example (ibid, p.14):

> *The Hua of New Guinea [...] developed elaborate networks of food taboos that govern what men and women may eat. In order for their boys to become men, they have to avoid foods that in any way resemble vaginas, including anything that is red, wet, slimy, comes from a hole, or has hair.*

Here, the Hua consider following these practices to keep a boy pure, so when the time comes, he can become a man. In contrast to modern Western societies, the Hua, and similar cultures, have these rules about purity that they believe to be moral rules, not just social conventions. If morality is nothing but rationalisation of rules concerning harms and pleasures, how did we arrive at a moral judgment on a person's purity? Haidt concludes that there must be more to morality than just purely rationalisation.

With advances in moral psychology, Haidt finds the answer to his scepticism with the idea that the primary way people make decisions is with their intuition, i.e. their automatic, instantaneous judgments. So, to better encapsulate how moral judgment works, Haidt proposes a Social Intuitionist Model (Haidt 2001):

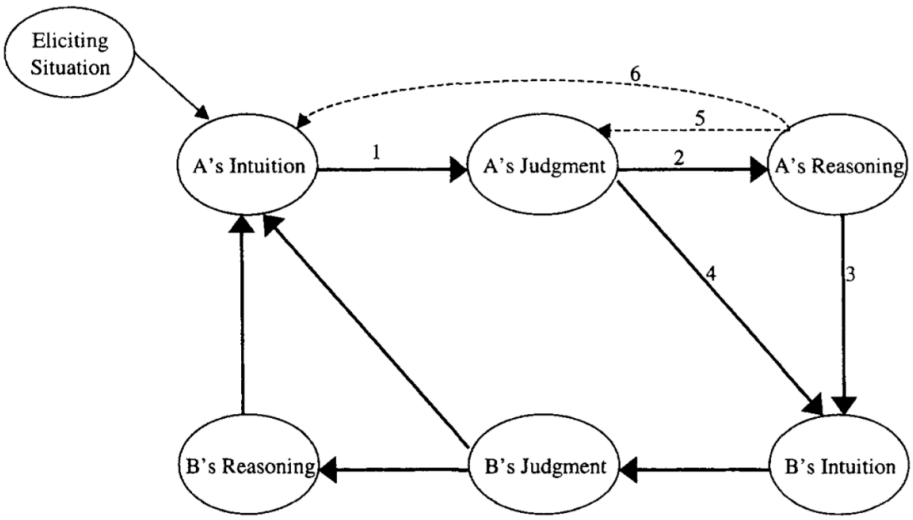

Figure 4 - Haidt's Social Intuitionist Model

This model, as shown, has 6 links. Haidt considers being the 4 key links to be:

1. Intuitive judgment.
2. Post hoc reasoning.
3. Reasoned persuasion (of others).
4. Social persuasion (of others).

And there are 2 more rarely used links:

5. Reasoned judgment.
6. Private reflection.

According to Haidt, this model accurately represents how people form their moral judgments. To explain this, let us consider an example of a morally eliciting situation. Suppose that there is a new policy under review that, when in effect, will ban all meat. This is based on a moral belief that has gathered a lot of support in our imaginary world, that animals should be treated with the same respect as humans. Hence, they ought not to be used as food sources. Suppose that Andy, a decision-

making agent who really enjoys his meat, says it is unacceptable and he would not support this policy. This model suggests that Andy makes his judgment instantly when being stimulated by an eliciting situation, based on his intuition alone (Link 1). Then, if left alone, Andy would stick to his judgment that this policy is unacceptable. Then, being a human, i.e., a social creature, as Andy is, he goes out to talk to his friends. One friend, Becky, questions Andy about his belief. Andy then says that it is unacceptable that the government is dictating what the citizens are eating. The model suggests that Andy's moral belief, that he reasoned to, *did not occur* prior to Andy's judgment, but *incurred* from Andy's desire to persuade Becky (link 2, 3, and 4). Haidt takes it that the automatic process, signified by link 1, is how people primarily make judgments, including moral judgments. Then, should there be a requirement for it, agents would initiate link 2 to appeal to another person's intuition via links 3 and 4. This is a justificatory reasoning process, as I will call them, to differentiate them from reflective reasoning processes. The key difference is that the former is a process undertaken to show why my judgment might be right, in service of convincing others, while the latter is a process to show why my judgment is true, in service of truth[8].

Critics of Haidt's model are quick to point out that there are cases in which people can override their intuitive reactions with reasoned judgments (Saltzstein and Kasachkoff 2004). Haidt (2001) does not deny this: he includes link 5 and 6 in his model to describe an internal reasoning process in case there is a lack of opposite view. However, he does not consider this process to be a major factor in how people judge situations before them. Mainly, Haidt (2012) questions the *frequency* of this overriding instance. The question is how often can reasoned judgment override intuitive judgment? Haidt uses a study conducted by Jennifer Lerner and Philip Tetlock, which shows that only under specific circumstances would people use their

[8] I will discuss this further in chapter Seven.

rational thinking to override their intuition before passing a judgment, to show the overriding frequency is minimal.

Lerner and Tetlock created an experiment where they ask participants to make a judgment on an eliciting situation. For example, consider the policy that bans all meat consumption. In this experiment, participants' belief about the anonymity of their judgment ranges from being completely anonymous, to being completely traceable, and will come under scrutiny by experts. They found that the less accountability is present, the more participants tend to pass their judgment quickly, intuitively. Only under three specific conditions will people form their judgments consciously and reflectively (Lerner and Tetlock 2003)[9]:

1. Decision makers learn before forming any opinion that they will be accountable to an audience.
2. The audience's views are unknown.
3. They believe the audience is well informed and interested in accuracy.

Lerner and Tetlock's conclusion here is that the foreshadowing of accountability plays a major part in when people use their reasoning to make judgments. People would use their justificatory reasoning, primarily, to explain or convince others of their judgments with post-hoc rationalisations, which is in line with Margolis' model presented above. It is only when the audience specifically requires one to use one's reflective rational thinking, with measurements put in place, that one would resort to it.

This appears consistent with observations of split-brain patients. Gazzaniga et al. (1962) observed a patient in 1962, who had gone through a procedure to have his corpus callosum (the nerve bundle that connects a person's left and right brain

[9] This will also be useful in my distinction between reflective reasoning and justificatory reasoning in Chapter Eight.

hemisphere together) cut. This was done to relieve this patient from convulsions, which he experienced frequently (up to 7-10 times per day) since a war injury he suffered in 1944. The procedure does affect one's physical skills, but the major point of focus is its impact on the patient's mental abilities.

Through testing with more split-brain patients, Gazzaniga found that there are primary functions to each hemisphere: the language centre of a person's mind is located in their left hemisphere, while the right hemisphere is responsible for initiating actions. Gazzaniga also found that, when patients with split-brains are stimulated to act via visual stimulus in their left field of vision, they act accordingly. But when asked why they acted as such, they tend to give a reason completely independent of the stimulus. The patient's left hemisphere was not stimulated, because the stimulus was not in their right field of vision; hence, it is not aware of the stimulus, but it gives an account to explain the action, anyway. From this, Gazzaniga concludes that the language centre is akin to an interpreter, which constantly hypothesises and describes actions that have already been made by the right hemisphere. This supports Lerner and Tetlock's theory that people primarily use their reasoning skills to produce post-hoc rationalisation.

With the primary conditions for people to engage in reasoning established, given Lerner and Tetlock stated these conditions are rare in real life, Haidt draws a conclusion from these studies that the social aspect of human life is crucial in moral judgment. Haidt suggests that moral reasoning holds a more *functionalist* approach: it exists *peripherally* to convince others to agree with one's view.

We do moral reasoning not to reconstruct the actual reasons why we ourselves came to a judgment; we reason to find the best possible reasons why somebody else ought to join us in our judgment. (Haidt 2012, p. 51)

A moral judgment differs from other kinds of judgments because it is not a simple subjective preference. It serves a unique, social function: to provide a basis to praise, or blame, *collectively*. So, in agreement with Lerner and Tetlock, Haidt believes

that deliberative thinking in moral judgment serves a role akin to a press secretary in a political institution: their job is not to make judgments, but to find evidence and arguments to support an intuitively made judgment to others (ibid). It is not news to point out that humans begin to wonder about morality in the context of a cooperative society. If a person is alone on an island, where the only rule is that there are no rules, that person would not waste their time thinking about morality.

Haidt's theory is thus sentimentalist in regarding the role of reasoning as minimal and not usually, or rarely, an integral part of making moral judgements. He presented a compelling argument, with empirical proofs, why people often reason in service of justifying themselves or to convince others. But I disagree with his conclusion that this justificatory role is the main role of reasoning in moral decision making. I next consider two views that provide arguments against this aspect of Haid's SIM model. My own evaluation and response will be developed in chapters Seven and Eight.

4.2. Kennett and Fine's argument for the conceptual necessity of reflective thinking

The first thing to criticise about models of moral judgement which are based on empirical research is that they do not discover moral truths. A descriptive model of how we in fact make moral judgements is not necessarily relevant to normative justification of moral statements. Although psychology is important and commendable, its truth is descriptive, not normative. But morality is not just a descriptive enterprise, it is also a normative enterprise. Moral rules give agents reasons to act in a certain way. Studies about morality are laden with questions not just about why we consider certain rules normatively binding, but also what properties ought specific rules have in order for us, as moral agents, to consider them to have that normative force. Any normative theory about morality needs to consider this. Kennett and Fine (2009) distinguish between three different questions (p. 80):

1. Which process, automatic or controlled, is most influential in the determination of moral attitudes or moral judgments?
2. Which process, automatic or controlled, best answers to our concept of moral judgment?
3. Which judgment has normative authority?

Moral philosophy benefits from psychology to establish a baseline, that is, such matters as the conditions, contexts, and mechanisms of how the human mind works. It makes sense that psychology can provide answers to the first question. But to produce a judgment that has moral weight, that is, normative authority, the task demands more than the execution of simple mental functions like pattern-matching or coming up with justifications. So, even if the first question can be considered answered with psychology, the second and the third cannot.

It is my understanding that Kennett and Fine think it is necessary to answer questions 2 and 3 above to satisfy at least one of our concepts for what we consider morally binding, for example, consistency. It is observable that what we consider moral dilemmas are largely constituted by inconsistency between moral judgments that a person holds. For one example, consider the Trolley Problem discussed in Chapter Two, where participants need to decide whether they ought to kill one person to save five others, with different levels of agential involvement in different scenarios. The part that we find morally puzzling is that we must stay consistent with the answers that we give, or else we find in ourselves an urgent need to have a justifying reason why we can make an exception. One other example is the puzzlement of veganism, especially when it revolves around the doctrine of equality. Followers of veganism cease consuming animal products because they believe that happiness and suffering are to be valued equally amongst all sentient beings. To go against this is to claim that not all lives matter equally, and that discrimination is

permissible, at least in some cases. But permitting discrimination is the basis for racism, which most people considered to be wrong. One can claim that speciesism is different from racism, and they ought not to be compared. But fundamentally, they both endorse some form of discrimination. So, to hold anti-racist sentiments and, at the same time, reject veganism is an inconsistency if, at the same time, one believes that all forms of discrimination is wrong.

The reason for this need for consistency may be that we take a moral judgment to be something we make as a person. When we judge an action in a moral sense, we consider it in terms of it being good. And when we judge someone morally, we consider their actions good; not just in their action's effects, but also in their wanting of such effects. This differs from judging one's taste in cuisine or aesthetics. The fundamental difference is one's moral characteristics, stem from one's wanting to do good, ought to persist. It has something to do with oneself, as a person, so it ought not to be fleeting. To condemn someone morally is to say they themselves are responsible for the result of their action as a person, not merely because they may lack some capacity to make rational decisions, or they may hold some distasteful preferences. Moral judgments are reflection on the person's agency, not merely our capacities. Therefore, the moral judgment that we make must be consistent for us to conceive ourselves as a coherent being, one with agency and an identity[10].

Based on this need for consistency, agency, and identity, Kennett and Fine suggest that Haidt's Social Intuitionist Model cannot satisfy some moral demands of our judgements or our concept of a moral agent. Haidt had used psychology to propose that our affective responses dominate the determination of our moral judgments. However, Kennett and Fine claim that a person's sense of agency would be oversimplified if we think of their actions as determined wholly by their affective responses like intuitions and emotions. To advance this claim, Kennett and Fine used Jones' (2003) distinction between our capacity to track and respond to reasons. One

[10] I will discuss this further in chapter Five.

can be a reason tracker when one acts as they have a reason to – e.g., I find something to eat because I am hungry. But on top of being a reason tracker, one can also be a reason responder, when one reflectively examines one's reason for acting – e.g., when I have an impulse to punch someone, I can examine my motives and reflect on them. For a judgment to correctly embody oneself, they claim that the process of responding to reason is necessary to separate between visceral, automatic responses and an agent's true desires. That means we must reflectively evaluate the motivation we take to explain our actions to make sure that they align to our true self. It is only through receiving our rationality's stamp of approval that desires can transform to normatively binding judgments. Therefore, if our affective responses were to claim dominance on our judgments, as in Haidt's model, this would undermine the sense in which a person's agency is embedded in said judgment.

To support this idea, Kennett and Fine draw attention to the way we hold someone morally accountable for their actions. They note that the way we attribute moral responsibility is sensitive to whether the agent can exercise their self-reflecting capacity. For example, we do not hold animals or small children morally responsible for their actions, no matter how undesirable the consequences of those actions are. We tend to excuse them in ways that refer to their conscious capacities to frame what they did as mistakes, because they do not know any better. A trickier case is psychopaths, who "display gross deficiencies in both reason tracking and reason responding", at least with respect to the "moral domain" (Kennett & Fine 2009, p. 86). It is hard to determine if they are morally good or bad people if they cannot feel anything towards morally motivating reasons (deficiency in moral reason tracking), and are thus unable to attribute any value to those reasons in their self-reflection (deficiency in moral reason responding). On the other hand, they claim that those who have reflective capabilities but lack the ability to take other's perspective in their reasoning, i.e., autistic people, can still be counted as moral agents. They conclude that the capability to make reflective judgments is essential in allowing agents to express their agency.

This view thus defends the idea that an adequate account of moral judgement must recognise a further role for reasoning than does Haidt's model. Reasoning, while it can be undertaken in the service of persuading others or providing post hoc rationalisations, is also implicated in distinguishing which actions express a person's agency, and which do not; and therefore, which actions are attributable to a person as a normative agent, and which are not. It is in this second use of reasoning that Kennett and Fine think the importance of reflective thinking lies; and that we ought to attribute normative authority to reflectively endorsed judgments. This is a plausible view, and I will expand this in chapter Seven with a further discussion in chapter Eight. For now, though, it is imperative to examine why, as far as expressing agency goes, reasoned judgments might not have the normative authority over sentimental judgments that Kennett and Fine attributed to it, through Jones' (2018) argument.

4.3. Karen Jones' trajectory-based model for rational agency

Jones' (2018) critique of Kennett and Fine identifies further ways that reasoning is involved in moral judgement, by arguing that reasoning processes themselves affect automatic decision making over time. While Jones shares Haidt's view that a moral judgment starts with an automatic, instantaneous judgment, she does not agree that rational thinking is limited to a peripheral role. Jones also does not agree with Kennett and Fine that a reasoned judgment can have an absolute normative authority over sentimental judgments, or, at least, not always, as sentimental judgments can also express past reflective endorsements. She suggests that conscious reflective deliberation can work passively, like a monitor in a security room, rather than having to be actively involved in the decision-making process to show its presence. To establish this, she points to the fact that judgments do not happen in a slice of time. Rather, they occur, one after another, in a continuous chain. This means that past judgments can influence present and future judgments. Hence, reasoning, regardless of whether it happens before or after a judgment is formed, is not isolated

to that case, and can still have an effect later down the line. Jones focuses on this temporal point of view to raise the instrumental role of reasoning, i.e., *conscious reflective deliberation*, in moral judgments.

While agreeing with Kennett and Fine that reflective endorsement is necessary to express agency, Jones does not think that a conscious, reflective, deliberation process is necessary for a judgment, especially affective, automatic ones, to exhibit an agent's reflective endorsement. This is because of two reasons: first, deliberation does not guarantee the expression of one's agency, and second, past reflective processes can, and do, modify our automatic responses. For the first point, Jones (2018) uses the case study, from a study by Arpaly (2003), of a dangerously thin anorexic, whose internal monologue against eating is of her mental disease and not of herself. This does not seem to be a problematic notion, as we can easily conceive of people who deliberate internally into acting in a way that will cause them regret later – when one is feeling scared, one often convinces oneself out of doing something by thinking that some associated risk is much higher than it really is. This, as Jones (2018, p. 267) called it, is a kind of "self-bullshitting", when agents reason in service of some pre-conceived judgment.

For the second point, Jones posits that there are at least two ways for conscious reflective deliberation to affect our automatic responses: by previous deliberations, or by the lack of its disapproval. For example, suppose that you saw an empty coffee cup on an empty bench on your way home; you were late, you hurried past without picking it up to put it in the trash bin, which is on the way. Later, you think about it, make a reflection on your actions, and conclude that next time, if the situation is similar, you would pick it up and put it in the bin, because it is the right thing to do. The next day, you see a plastic bag wafting in the wind while hurrying home; you catch it and you put it in the recycling bin. Because this shows the effect of a conscious reflective deliberation on your action, it still counts as a rational decision – even if, at the time, it is quite automatic.

This function extends to the modification of affective responses as well. For example, suppose we give out money to a homeless person because we feel sorry for them. But later, we learn that we were scammed, and that person is a rather well-off traveller that ran out of allowance for his trip. We reflected within ourselves to conclude that we would only give financial support to the homeless through official charities. This does not just affect how we act the next time we see a homeless person, but how we feel as well: we feel less sorry for their situation; or at least, not sorry enough to give them money immediately anymore. Jones calls this the self-regulation of affective responses, enabled by our advanced reflective capacities in adulthood. To Jones, this process of responding to the cause of our actions, or reason-responding, is what makes us rational agents. This differentiates us from mere reason-tracking beings, whose actions are determined solely by stimulus in their field of view. Resulting from this process is a trajectory of our development as an individual, which forms our agency.

Further, Jones (2018) points to the instance that reasoning can exhibit in another form: in their allowance for certain habits to continue. There are habitual things that we learn growing up, for example, to help someone get up when they fall. When we act in such a way, we do, at least sometimes, consider ourselves to be acting rationally. Here, we do not consider ourselves as such because we have reflected within ourselves, now or sometime in the past, to arrive at the conclusion that we ought to act in such a way. Primarily, we were taught to act that way, and we never ceased doing so. It is habitual, but we consider ourselves acting rationally because, after all this time, we have not found a reason to act contrary to it. This shows that we think this act is appropriate, or at least, we do not find it problematic, to act in that manner. In other words, it is consistent with our trajectory, our agency. Although this is passive, it nonetheless exhibits the effect of conscious reflective deliberation, as a lack of rejection, rather than an active approval.

Based on these two effects that conscious reflective deliberations can exhibit themselves in our judgments, Jones proposes that the role of reasoning is analogous

to a monitor, which tracks the progression of our automatic responses. If we continue to make reflections on our actions and continue to develop a trajectory of our agency by regulating our affective responses and reactions, conscious reflective deliberations will be present in our subsequent judgments. If our affective responses and reactions follow this trajectory, there is no reason for us to use conscious reflective deliberations. Hence, it can be said that our capability for conscious reflective deliberations do run passively, like a monitor, watching ourselves make autonomous judgments and intervening where necessary.

In this account, conscious reflective deliberation does not need to be actively utilised to make an impact on our judgments. It only needs to actively regulate and monitor our reactions and intuitive judgments and intervene when necessary. This gives reasoning a more prominent, albeit more passive, role in Jones' model compared to Haidt's. It is more prominent in that it has a wider effect than just under some specific conditions. But it is also more passive in that it is constrained by an agent's trajectory. Depending on whose trajectory, certain situations might not arouse deliberation at all.

The conclusion Jones draws here, which I mildly disagree with, is that a reflective deliberation process is not necessary in the expression of agency. The problem, I think, is the conflation between the kind of reasoning that enables self-bullshitting, i.e., deliberation in service of a preconceived judgment, that both Haidt and Jones allude to, and the kind that Kennett and Fine were referring to – the kind that enables agents to reflect internally. While I agree with the premise that affective judgments can sometimes express reflective endorsement, and agency to an extent, better than a reasoned judgment of the first kind can; I still believe that a reflective deliberation process of the second kind is beneficial in our moral decision-making. I think it is beneficial to the expression of agency if we do practice the second kind of reasoning before we act, as we often attribute deliberative actions to other's agency, more so than their habitual actions. On the other hand, Kennett and Fine's account is still quite vague. While I agree that the ability to reflectively reason has some effect

on our judgment on one's expression of agency; their account has not really demonstrated exactly how we express our agency in practicing controlled deliberative thinking. However, it is interesting that in Jones' argument here, she alludes to the intertwining between affective automatic processes and conscious reflective processes, that they can affect one another. With this, in the next three chapters, I will attempt to draw how these processes intertwine and jointly produce morally relevant qualities – in our case, the expression of agency.

5. Desires and Beliefs: the building blocks

For every theory of moral motivation presented so far in this thesis, whether they are sentimentalist, rationalist, or anywhere on this spectrum, I believe each of them has grasped something true; and that a successful descriptive theory of moral decision-making can incorporate these true ideas into one model. This thesis aims to move toward building such a model; one in which the motivational process is influenced by the interaction between emotional processes and rational processes; and recognises that extensively exercising this interaction is pivotal in the making of an agent's best judgment.

The model I draw in the following three chapters will illustrate how discrepancies between beliefs and desires cause our emotions and motivations, as well as how standards of rationality affect this process. In this chapter, I begin this development by presenting my understanding of beliefs and desires, regarding their general classification in 5.1, their respective senses of strength in 5.2, as well as further classifications between different kinds of beliefs and desires in 5.3. This is because in practical decision-making processes, we need to be able to rank which option is the best option to take; as we generally act in a way we believe will most likely be able to satisfy our strongest desire. A clear understanding of desires, beliefs, their senses of strength, and their classification will be necessary in the subsequent discussion on how they cause emotional experiences and what it means to be rational, in the upcoming chapters.

5.1. What are beliefs and desires?

The 20th century's moral philosophical debates have been shaped not just by moral philosophers, but also by involvements from scholars in the philosophy of mind. One notable contribution from scholars in the philosophy of mind that intensified the argument between rationalism and sentimentalism is the 'direction of fit' argument, provided by Gertrude Elizabeth Margaret (G.E.M) Anscombe (1963).

Anscombe provided a thought experiment that highlights the fundamental conceptual difference between desires and beliefs. She conceived of two distinctive kinds of attitudes that an agent can have about a state of affairs (α). Anscombe did this by examining differences between cases of inconsistency between the physical state of the world and the mind's representation of its state. The fundamental difference is that beliefs are susceptible to empirical data, while desires are not.

In Anscombe's experiment, she describes a man going shopping according to his shopping list, and an intelligence officer tailing him to record his actions. The analogy she wants to make is that, human desires (the man's shopping list in this experiment) remain unchanged in the case that the world does not conform to them (the man does not buy the things on the list). On the other hand, human beliefs (the intelligence officer's recording) are volatile to changes in the world (whether or not the man buys the things on the list). Another way to look at this is through a simpler thought experiment:

> *Supposed I want to eat two pieces of cake, and that I believe each piece contains 1000 calories, or 1kCal, which is under the daily limit I set for myself. But then my mom shows me the box, which says each piece contains 2kCal. This means I can only have one piece of cake if I want to stick to my daily limit.*

Here, as we can see, my previous belief that each piece only has 1kCal was replaced with a new belief (that each piece has 2kCal) by the information provided on the box. But the same cannot be said about my previous desire, that I want two pieces of cake. It is incorrect to say that the previous 'cake' desire disappeared or is replaced by any new desire; it is better to say that it persists and is in conflict with another desire (to stick to my daily calorie limit). Saying the previous 'cake' desire persists is more correct because that desire, instead of being compromised with the

desire to stick to my daily limit, can still win this conflict of desires and result in my having two pieces, anyway. In this example, our mental attitudes are conceived in two distinct types: beliefs, which have a "mind to world" direction of fit, and desires, which have a "world to mind" direction of fit (ibid).

Analogous to the thought experiment is the way that beliefs and desires interact with the world. When there are discrepancies between desires and the physical world, people are moved to change the world's current state. For example, my cake is not frosted, and given that I want icing on my cake, I am motivated to put some icing on it. But when there are discrepancies between beliefs and the state of the world, beliefs tend to change, at least when we are presented with substantial enough evidence; such as thinking I have icing in the fridge only to find that I have ran out of icing – in such a case my belief would change.

There are scholars who reject this classification. Platts (1979) argues against this classification of our mental attitudes on the basis that all desires involve some form of beliefs (p. 257). Desires inherently need an object – something to desire – whether it is a belief from observation or an imagination. So, Platt claims that it is questionable whether desires can be conceived separately from beliefs at all. However, a connection between these mental states does not negate that they can be conceived separately. For one, in saying that they are distinct mental states, I am not saying that they must exist independently from one another. They can also be subjected to one another. We can have beliefs about desires, like believing that I want my friends to be happy. Or we can desire to believe something, such as wanting to believe justice will always prevail, even if we do not believe it is so. In fact, these attitudes must be conceived distinctly from one another for there to be a space in which they can form a connection as such (Han 2017).

Compounding the separation, Anscombe (1963) pointed to the different types of knowledge yielded from interactions between attitudes and matters of fact. When there is an inconsistency between a belief and an observation, this results in factual knowledge. That is, facing the fact that each piece of cake has 2 kCal, I acquire the

knowledge that each piece is, in fact, 2 kCal. For Anscombe, this means the object of a belief is knowledge. But when there is an inconsistency between a desire and a factual situation, it gives a different kind of knowledge, which she calls practical knowledge. It is practical because it does not turn the mind to what currently is, but rather to what ought to be done. In other words, knowledge we gain from our beliefs is descriptive, while those gained from desires are normative. This separation between different types of knowledge is useful, as moral decision-making processes largely rely on practical knowledge, with some regard to factual knowledge. Hume's view that passion, or desire in our current terminology, is essential in moral matters also coincides with Anscombe's idea that desires are essential in practical knowledge.

5.2. The strength of our attitudes

Another distinctive feature about these mental attitudes is that they can seemingly differ in strength, in different ways. Distinctions in terms of strength are important to discuss because, as mentioned in the last section, practical knowledge is largely determined by our desires, and our desires do have a belief (from memory, observation, or imagination) as their object. In practice, regardless of how many desires we have, as humans, we only have one body with which to act; and "no matter how many different things you want to do, you in fact do one rather than another" (Korsgaard 1989, p.110). This generates a need to eliminate some desires in favour of another when we decide on which options we should follow. We determine what to do based on each option's expected utility – a function of how much we want the expected outcome, and how likely we believe we are to attain that outcome should we commit the act (Paul 2015). Here, the strength of our attitudes plays an important role in evaluating the practical value of each decision available to us when

we consider how to act in any given circumstances. So, the strength of our attitudes about any decision determines their ranking in our decision-making processes[11].

To describe the strength of an attitude, it is best to place them on a spectrum. Categorising attitudes into a binary of strong versus weak types seems unfit for our purpose of ranking attitudes. Suppose that we can conceive of attributes that can assign attitudes to either strong or weak types, and that the identification of these attributes is beneficial to our purpose here, the task of ranking attitudes within each category is still unsolved. Instead of using attributes that make an attitude strong or weak to form categories, binary or otherwise, I think using them to form a scalar system for attitudes, i.e., not weak versus strong but weaker versus stronger, is more fitting to our purpose. And to this end, the task of comparing attitudes to see which attributes make one stronger, or weaker, than another will be essential.

Each kind of attitude here has its own spectrum of strength – based on their direction of fit, that is, their point of reference. First, the strength of an agent's belief is characterised by how well it fits the world, i.e., how *certain* is the agent that of a belief's being true. Beliefs, as attitudes, can be conceived as an agent's acceptance that something is true (Cohen 1993). However, agents do not ordinarily simply accept or reject all beliefs, so that we have a binary between acceptance and rejection; we have doubts, and we often are unsure about something being true before accepting or rejecting it as true or false. The strength of a belief about α, therefore, can be understood as a spectrum, differentiated by the agent's certainty of α being true. This spectrum can range from an agent's being uncertain ($p=0$) of α being true, to an agent fully certain that α is true ($p=1$).

Beliefs' location on this spectrum of certainty is determined mainly by two factors: observability and coherency. Primary experiences like observation, memory, or imagination; or secondary experiences, which derive from someone else's

[11] Furthermore, understanding the strength of our attitudes will be pivotal in the discussion on the strength of emotions in chapter Six.

experience through stories, or other forms of media, can be the object of an agent's beliefs. Typically, beliefs from primary experiences tend to be stronger than those of the secondary kind – you believe there is a bushfire when you see it more than only hearing about it. However, there are exceptions, such as believing the Earth is spherical and not flat – the kind of knowledge attainable only secondarily unless you have the technical ability to verify it yourself. This is because, aside from observability, the probability of beliefs also differs with whether it is rationally coherent with a larger system of beliefs, i.e., with other beliefs that it can be true with, simultaneously. An agent can believe in something unobservable, over something observable, if said unobservable object is more rationally coherent with their other beliefs. Whether someone believes that the Earth is flat or spherical, for example, depends on their other beliefs. If one thinks "the Earth is flat" is more compatible with their other beliefs, e.g., "the horizon is flat", then they would be more inclined to believe that the Earth is flat[12].

On the other hand, the strength of a desire is based on the *saliency* of the desired state. While it is intuitive to conceive of the strength of an agent's desire by how much the agent wants the object of said desire, I think that it is not enough to reflect how context affects one's desire. Recall Platts' (1979) objection to the distinction between desires and beliefs above, because desires necessarily involve some form of belief. While I have rejected Platts' objection, arguing that these attitudes can be conceived distinctively from one another, it still holds true that contextual factors relating to someone's beliefs about a state of affairs α do affect their desires about α. The contextual factor provided by our beliefs about a state of affairs α, that is most relevant to the strength of a desire (*∂*) about α, is urgency. It is not problematic to assume, for example, our desire for safety increases when we believe that there is a visible threat, compared to when there is none. The spectrum of strength for desires must be able to take account of this contextual factor. To this

[12] I will expand this further in chapter Seven, where the discussion on rational thinking will be.

end, the appropriate question to determine the strength of an agent's desire for *∂* is "how important is it for me to satisfy desire *∂* right now?". The answer to this question is the level of saliency an agent has towards the satisfaction of desire *∂*[13].

5.3. Distinctive types of beliefs and desires

However, identifying the sense of strength of an attitude alone would still be too vague to allow development of some of the key concerns of this thesis. In particular, it would not help resolve the debate between Jones and Kennett and Fine, which hinges in part on which processes can latch on to an agent's true motivation – that is, which desires are intrinsically the agent's as opposed to being prompted by external factors. Kennett and Fine's idea that agents can only be held accountable for desires which they reflectively endorsed, even if they have other (stronger) desires that are prompted by external factors, implies that when an agent is motivated by a strong desire, said desire might be the strongest but not necessarily an agent's intrinsic desire. This shows that the distinction between weak and strong is not helpful in distinguishing between an agent's true desires, and her other desires, since it would be strange to claim that strong desires are ours and weak desires are not when their strength fluctuates.

The distinctiveness of the kind of desires that relate to an agent's self-conception is also stressed by De Sousa (1987). He distinguishes three kinds of desires:

[those that are] intrinsically subjective, those that we cherish constantly because they have moral or aesthetically objective import, and those that are morally significant not simply because of the value of their object, but because of what our having such desires means for our self-concept, our energy, our integrity. (ibid, p. 181)

[13] Furthermore, this sense of saliency will be crucial in determining our motivation in committing to an action. I discuss this further in Chapter 6.

As we can see, by having a connection to our own self-concept, a desire warrants its own category. De Sousa (1987) attributes this connection to a second-order desire, namely, the desire for having certain type of desires such as for beauty and goodness (p. 187). For example, diplomatic methods and strength might both be valuable in solving conflicts, but I am me in part because I value diplomacy over brute strength, and I would endorse the use of diplomacy to solve conflicts, should I reflect on the courses of actions I can take, over using brute strength. In Frankfurtian terms, this means I reflectively endorsed the use of diplomacy to solve problems, that I have a second order desire to use diplomatic methods (Frankfurt 1971). Korsgaard (1996, p. 191) also conceives of this reflective endorsement as identification, suggesting that there is a distinctive kind of desire which agents identify with, that such wanting contributes to what makes me *truly me*. In this thesis, I will refer to these kinds of desires as *agential desires*.

One other important characteristic of the agential desires, mentioned by both Frankfurt and De Sousa, is that this type of desires directs us continuously to satisfy them and not dismiss them. This characteristic is interesting to moral decision-making. It seems that desires that we consider "moral desires", such as the desire for justice, or for the good, demand continuous satisfaction – as it relates to what we consider an ideal *state of being*. When we speak of agential desires, we speak of an ideal identity, i.e., our ideal self, the kind of person we want to be (but might not yet necessarily be). If I say want to be a diplomatic person, I am expressing that the *state* of "being a diplomatic person" is something I find desirable, even if I am not yet a diplomatic person. An identity is a state of being, not an object. Analogously, when we speak of a wanting justice as a moral desire, we are expressing that a just society (state) is an ideal state. This solidifies the connection between moral desires and agential desires. Desires that are open to moral assessment tend to be agential desires. Therefore, this classification between non-agential desires and agential desires would be appropriate to our discussion on moral decision-making.

Another distinction between types of beliefs, that is important to decision-making, is the distinction between factual beliefs and procedural beliefs (Schwitzgebel 2019). This distinction is mentioned briefly in 5.1., corresponding to Anscombe's distinction between different kinds of knowledge, factual versus practical, from our attitudes. Factual beliefs correspond to the actual states of objects in the world, such as "there is a fire", in the form "A is". Beliefs of this type generally point to how things are. The second type, practical, procedural, or means-ends beliefs, relates to processes or a logical connection of different events, in the form of "if A then B"; such as "if there is smoke, then something must be burning".

These different types of beliefs have unique roles in our decision-making processes. Suppose we want to cut an onion into rings. Factual beliefs, like which shape the onion is already in, determine whether it would be possible for us to do so. If the onion is already halved from the root end, we can hardly make onion rings and it would be wasteful to try. And, supposing that the onion is still whole, procedural beliefs regarding which angles we should cut the onion from to make onion rings, would prompt us to not cut it from the root end. The distinction is important as procedural beliefs will be our main focus moving forward. However, it will be helpful to have a clear distinction between these kinds of beliefs, as well as the relationship between them in decision-making processes.

To sum up this chapter, beliefs and desires are mental states with their own distinctive functions. Beliefs are representations in our minds about the world around us, and desires are representations of how we want the world around us to be. When we see a whole onion, we believe that the onion is whole; but we might want the onion to be cut in half. If we do, we might develop a motivation to cut said onion in half. Moving forward, being able to keep in mind these types of attitudes will be beneficial to our task of defining emotions, motivations, and reasoning, and understanding how these attitudes and processes work together in moral decision-making. These different types of desires and beliefs interact with one another in a

complex way which gives rise to our emotions and motivations, as well as providing the framework in which our reasoning can exert its function.

6. Emotions: the signals

Now that we have a clear understanding of beliefs and desires, by drawing on recent work in empirical psychology I will describe how aspects of our emotional experiences can be explained in terms of comparisons between our beliefs and desires in this chapter.

One way to define emotions is to look at a common emotional experience: suppose you are stir frying your onion and it appears to be burning. In perceiving that, we would, in panic, take the pan off the heat in hope of salvaging whatever is edible. In such cases, we can report that we feel panic, and we reacted in such a way as to relieve that feeling. Reacting to the situation at hand involves an evaluative aspect (appraising and confirming the belief that the onion is burnt, as well as the desire for it not to be inedible), a phenomenological aspect (feeling panic), and a behavioural, or motivational, aspect (taking the pan off the heat) of an emotional experience (Scarantino 2021). These reported experiences are our focus in this chapter. I will discuss the evaluative and phenomenological aspects of emotions, using a contemporary study in philosophy of mind by Rainer Reisenzein (2009), in 6.1 to present how comparisons between beliefs and desires with various strength causes the diversity of our emotional experiences. With the link between our attitudes and emotions established, I will form the basic model of affective decision-making in 6.2, to illuminate further how agents are motivated by their desires and beliefs. This will be the basis model upon which I will draw the function of reflective thinking to develop my own model of moral decision-making, the Affective-Reflective Model (ARM).

6.1. Evaluative and phenomenological aspects of emotions

First, I will address the evaluative and phenomenological aspects of emotions. Reisenzein's (2009) beliefs-desires model for emotions posits that emotions are signals of important changes being made in our representational system (i.e., our

overall set of beliefs and desires) as we perceive states of affairs. This means our various emotional experiences are signals of different types of changes in our desires and beliefs systems. The diversity of emotions then can be interpreted as signals of multiple types of comparisons, which can be constituted by different types of beliefs and desires, as well as their contents, chronological order, or agents' certainty of their representational mental states. This would explain the evaluative aspect of emotional experiences – the emotions we have in observing states of affairs are our evaluations of these states against our representational system.

This description of emotions is consistent with a number of points already presented in this thesis. As Haidt (2012) points out, emotions contain cognitive steps. And as Margolis states (1987) emotions are a type of information processing, even if the feeling process typically happens instantaneously, and emotions form as we perceive things without much cognitive effort. This means the feeling process, if it does contain steps, must not bear a heavy cognitive load. Reisenzein's view here posits that the evaluation that forms emotions amounts to pattern-matching. And as pattern-matching does not require a heavy cognitive load (Haidt 2012, p. 41), Reisenzein's description of emotion is compatible with Haidt's and Margolis's.

Reisenzen (2009) argues that the phenomenological aspect of emotions, i.e., the feelings themselves, are signals produced by the evaluative process which forms emotions. Reisenzein (ibid, p. 9) gives some simple formulations of emotions, depicting which evaluative process produces which emotion. For example, feeling happy is a signal that the state of affairs we desire is the state of affairs we believe to be true; or vice versa for an unhappy feeling. To give a simple example: suppose I drive to a restaurant late at night and see it still is open. I feel happy because I desire it to be open, and I observe that it is.

I think this analysis can be applied to explain more complex emotions and how small differences in our representational system can affect the way we feel as well. For example, let us analyse anger and fear. In Reisenzein's terms, feeling angry about A would be a signal that we do not want A to be true, while we believe that A is true,

and we also believe that we could have done something to prevent A. Suppose that I drove to that restaurant to see it closed, while believing that I could have started driving sooner (but did not for whatever reason). Changes in these comparisons can bring about changes in our emotions. If we do not believe that we could have done something to prevent A, we will feel sadness and not anger. If I believe I could not have started driving sooner, perhaps just because the decision to go to that restaurant is a spontaneous one, I will only feel sad to see it is closed.[14]

As Reisenzein (2009) mentions, the feeling process signals *important* changes, not every change. This implies that the saliency of a situation plays a large role in the formation of emotions. I discussed salience briefly in Chapter Five proposing that the salience of a desire determines or acts as a measure of its strength. This link between saliency of desire, and emotion, consolidates the connection between our desires and our emotions.

And since saliency is subjective to agents, this connection between emotions, desires and salience would also help us understand the subjectivity of our emotions. People can have different emotional reactions as they perceive a state of affairs. In my example above, supposed I drove there with a friend. I can feel angry because I believe I could have done something to prevent being late, while my friend feels sad because they do not think I could have done any better. But in this case, we both feel something; what about cases where one person feels intense emotions, while others do not, as they perceive the same state of affairs? The answer to this stark difference in emotional experience, I believe, lies in the sense of saliency that is subjective to individuals and their desires.

[14] It should be noted that there are cases of irrational emotions. Take an extremely agoraphobic person for example, who would have an anxiety episode by simply stepping out of their front door. The account here suggests that the emotion itself might not be *that* irrational. Rather, it is the person's over-expanded beliefs about the possible danger from unfamiliarity (what we ordinarily judge as irrational to believe) and over-stimulated sense of danger-avoidance that caused these emotions. Of course, this idea is open to further examination, but alas, it is not our current focus here.

Saliency plays an important role here, as it might determine whether we feel emotions as all. Depending on our dispositions, history, experiences, identity, and all matters personal to us, what is considered salient varies from person to person, even if they perceive the same state of affairs. My feeling towards what I perceive can change drastically based on which aspect of seeing someone being hurt I am disposed to find important. If I am disposed to find human suffering important, I would have the desire to alleviate any suffering I can; and depending on whether I believe I can change this state of affairs, perceiving this would cause me to feel either anger or sadness. But if I am disposed to find human suffering not important, because I am psychopathic or sociopathic, I would not have any desire to alleviate other's pain, and I would be apathetic towards the person being hurt.

The role that emotions play that is important to moral decision-making is that, in experiencing an emotion, we focus our attention on specific beliefs, and their correlate desires, about the state of affairs we are perceiving. What we desire and what we consider to be salient to us are, for the most part, implicit in our emotional responses. When we feel happy in seeing Joe being happy, it implies that we do desire Joe's happiness, that Joe's happiness is important to us; while people who do not love Joe would not have experience of that happiness.

So, in using Reisenzein's beliefs-desires model for emotions, I have discussed the details of how our beliefs and desires cause our emotional experiences in this section. This will set up the upcoming discussion on motivation as another output of the same process that produced emotions.

6.2. Motivational aspects of emotion

Now that we are clear about how beliefs and desires cause emotions, we can move on to how beliefs and desires cause motivations to act. As we experience our emotions, some emotions motivate us to act, such as anger, while others do not, like contentment. This is because, in being motivated, there are (at least) two implied emotional states: a negative emotion about a belief at present, and a positive

emotion about an imaginative belief in the future, with the two belief states being connected by a procedural belief. Negative emotions are signals that our beliefs upset our desires, and positive emotions are signals that our beliefs satisfy our desires. In general, when we experience a negative emotion, we want to change what we believe is true – we want the world to be in a different state to the state it is in. The desire to change states of affairs is what motivates us to act[15]. Here we can see that motivation can be a result of comparing mental states, just as emotions are. This makes it possible to explain the motivational aspect of emotions by using our beliefs and desires. In acting to realise an end, it is implied that agents do desire that end. Even in cases where agents seem to act against their self-interests, there is something they want in the end result of said action, even if it is not the main focus of the action itself. One simply does not act if one has no desire for whatever the action can possibly bring about.

It is important here to keep in mind the relationship between the strength of our desires, and hence our motivations to act, and our beliefs. It is intuitive to think that the more strongly we desire something, the more motivated we are to satisfy that desire. Recall the distinction between desires for objects and desires for states of being in 5.3. Desires for states of being demand continuous satisfaction, and desires of those types are constitutive of our moral self-concept, such as being a kind person. So, it is safe to say that a desire like wanting to be a kind person is strong, and it demands continuous satisfaction; so, it would make sense that it motivates us always. But although it might remain present as a standing desire, the desire may not play a role in many particular decisions or often become occurrent. We do not go about our day trying to prove that we are a kind person every single day. Simply having a desire does not seem to necessarily produce motivations to act.

[15] Perhaps it is worth noting here that this is not limited to changing an undesirable state. Some do act to maintain a status quo. But, in my view, the motivation to maintain the status quo can also be seen as motivation to either prevent or correct occurrences that disrupt the status quo, making these occurrences the states of affairs that we aim to change.

So, for us to have a motivation to act to satisfy an agential and moral desire, we must also believe, to some extent, that these desires are frustrated, and that we can do something about it. This would mean our motivation is an output of the same process that causes emotions. Recall Anscombe's (1963) claim, discussed in Chapter Five, that when there are discrepancies between objects of our belief and their related desire states, people are necessarily moved to make an attempt at changing the outside world to "fit" with their desired state of affairs. But, saying that we will necessarily be led to act when our beliefs upset our desires might be too strong of a position, as it implies that emotions necessarily lead to action. We can, and often do, fail to act even if there is a wild difference between our desires and our beliefs, for instance, because of interferences either from other desires, or we believe that our desires are impossible to be satisfied, e.g., depression.

I support a weaker position, that when we believe that our desires are frustrated, we will simply have a motivation, that is, a disposition, to act. A discrepancy between our desires and beliefs induces a motivation, but it does not necessarily move us to act. An important relative factor between motivations and emotions is their force: the more strongly we feel about something, the more motivated we are to act. Stronger emotions, like being enraged at something, motivate us much more effectively than weaker emotions, such as simply being irritated. However, this does not mean that we will necessarily act on our strongest desires, as the strong view suggests. It is possible for an agent to have a strong desire for something, but also have an accrued number of weaker desires, which in sum can have a higher motivational force than the strong desire.

This weaker position here is compatible with Hume's view, for it still allows that desires are necessary for motivation. It does not go against Hume's idea that having a motivation presupposes having a desire. It is even more so compatible with Smith's Humean dispositional view on desire, that desires cause dispositions for action, but do not necessarily cause action, presented in Chapter Two. Moreover, in conceiving beliefs as a partial cause of emotions and motivation, we can more easily conceive of

ways in which reasoning affects our emotions, which will be discussed further in Chapter Seven.

6.2.1. The motivation model

Putting these ideas about motivation together, we can arrive at a rough model to demonstrate the process of how beliefs and desires lead to action as discussed so far:

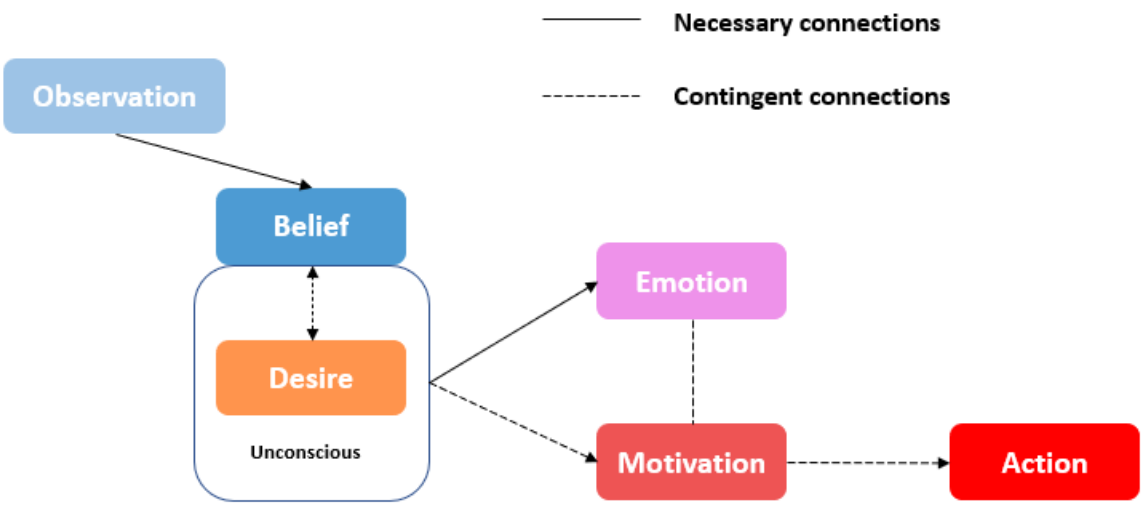

Figure 5 - The basic motivational model

Aside from what we have discussed so far, I add three elements in this model to provide more clarity on the conditions of, and connections between, its elements. The first element added to our base desire-belief-emotion-motivation is "Observation", to indicate another element of this process: it is context dependent. For every process, we do not necessarily conjure up all of our desires and beliefs; they are brought up by their relevancy and saliency, which is dependent on contexts, provided by the situation we are facing, what we are observing, and sometimes, what we are primed to pay attention to. Note here that "observation" is used in a loose sense: states of affairs need not enter the comparative process, as belief states, strictly via our sight; they can enter via any kind of senses, or even memory recall.

The point here is that belief states need to be instigated by some or other features of our context. This prompts them to be compared to related desire states. This means comparison processes are triggered base on context. Therefore, "observation" is added to indicate this dependency.

The second element added is dotted and continuous lines to indicate connections between other elements of this process. There are two necessary connections and four contingent connections. The necessary connection between observation and belief indicates that when we observe something, it is necessary that we think it is true, albeit in a shallow sense. We might have a sense of rejection towards what we are observing, e.g., that we are being cheated on when we see our lover in bed with someone else, thinking there must be some other explanation. But at the very least, we believe that we are seeing it, so, observations necessarily cause some beliefs in us. And if we have an emotional experience accompanying this observation, it necessarily means that the situation has some salience to us, we have some desire relating to what we are seeing. The emotion plays the role of signalling this to us.

There are three contingent connections in this model. The first is in the comparing process; we only compare our beliefs to related desires if we have those desires. Otherwise, there would be no comparing process because we have no desire related to the state of affairs in our beliefs; in other words, the state of affairs here have no salience to us. This would result in a "neutral" attitude, or that we do not feel any emotion or motivation. The contingent connection between the belief-desire comparison and motivation indicates that while the comparison between our desires and beliefs necessarily causes emotions, it does not necessarily induce motivation. Motivation, in various degrees, is necessarily induced only if our beliefs are inconsistent with our desires, but not necessarily if our beliefs do not upset our desires. The contingent connection between emotion, motivation and action is there to recognise that there is a connection regarding the strength of desire between them. A desire might be strong enough to cause emotions, but not necessarily

enough to cause motivation; and it needs to be even stronger to cause action. So, it is neither necessary that we are motivated if we feel emotions, nor that we act if we feel motivation.

The third element added is the indicator that the comparison between our desires and beliefs happens unconsciously, that is, automatically and unreflectively. Let us put this model through a basic example of me pulling my hands away from the fire because I feel pain. Supposed that I feel pain, instigating belief A: "I am in pain". Belief A then, assuming that I find not being in pain important to me, is compared to desire α "I want to not be in pain". I then would feel fear and the motivation to do something to relieve my pain, e.g., pulling my hand away from the fire, as the result of comparing belief A and desire α. If there is nothing else in this process that would prevent it, I would act on said motivation. In my experience, the act of pulling my hand away from the fire as I feel pain from the heat is a reaction: I react to the sensation of pain with a flash of fear and instinctively pulling my hand away, without much thought put into it. This means while I might not be conscious of the comparative process itself, I am conscious of its products, which are emotions and motivations, phenomenologically.

6.2.2. Desires and motivations

There are some properties of this process that I want to emphasise. The first is that the sense of saliency, which is deterministic of this process, is subjective but is also vulnerable to external factors. External factors include situational contexts. For instance, suppose that the pain I feel is from performing hard exercises, rather than having my hands near an open flame. This would result in a quite different response to the felt pain. In this case, belief A, that I am in pain, would not be salient to me, as it is expected when I begin to exercise. As such, the pain is already reflected in my beliefs, so no emotion or motivation to cease exercising are generated. Internal factors include my preferences and can also dictate the course of this process. Suppose that I am a masochist, and I am putting my hand close to the fire. Belief A is

salient, desire α might be brought up, and I do have a feeling of fear for my safety, but desire α^X: "I want to feel pain" would also be involved in the process. Given that I am a masochist, my desire for pain is stronger than my desire to be safe, resulting in a stronger pleasurable feeling, compared to fear, in having my hands near an open flame, experiencing the burning pain.

Secondly, via context, saliency of the state of affairs in an agent's attention can be primed. Once an emotion is felt, the desires that partly cause it become more readily accessible for subsequent processes. Haidt (2012) draws on an experiment by Zhong and Liljenquist (2006), stating:

> *[the experiment] has shown that subjects who are asked to wash their hands with soap before filling out questionnaires become more moralistic about issues related to moral purity (such as pornography and drug use). Once you are clean, you want to keep dirty things far away. (Haidt 2012, p. 71)*

If these subjects felt positive about being clean, then in our model, that feeling was the result of believing that they are clean, and having a desire to be clean. Here, subjects are primed, or disposed, to find cleanliness important. Then, a subsequent observation was made about issues on moral purity. Morally purity may be opposed to being morally impure, foul, or dirty. Those who have a lingering positive sense of being clean from the previous comparative process would have the desire to be clean, and not to be dirty, readily accessible for comparison. This may mean it is more salient than other desires that are also salient, but were not recently brought up in the person's attention. This leads the subjects to feel a clearer sense of disgust towards pornography or drug use, explaining why they are more moralistic in their judgments about those issues (Haidt 2012). This property of the motivation process points to a problematic implication: emotions and motivations can be primed. I will talk more about this in the next chapter.

The third property of interest is that observations, that is, perception and beliefs, do play an important role in determining the strength of our desires, emotions, and motivation. As discussed in relation to priming, not every desire possibly related to our beliefs will necessarily be brought up. Desires with more obscure connections to the state of affairs in our attention might not be brought up. We can fail to recognise the relevancy and saliency of the situation to some desires that are important to us if it the frustration of these desires is not brought to our attention. So, to compare the strength of our desires to their fullest extent and make our best decision, it is important to recognise, as much as possible, the effects of our actions, not just those our observations first latch on to.

So far, I have drawn the basic model of decision making, from observation to motivation. In doing so, I have drawn the connection from our attitudes to our motivation, and highlighted some important properties in this basic model to decision-making processes. Upon this model, I will demonstrate the function of reflective thinking in this thesis, that is, to expand our awareness about our actions and improve our decision making by enabling our affective judgments. We will explore this notion in the next chapter.

7. Reasoning: the corrective process

In the last chapter, I have outlined how our beliefs and desires determine our emotional experiences and motivations. This provides a rough outline of the motivational process, from when we observe a stimulus to being motivated enough to act. In this chapter, I will attempt to show what is the role of reasoning, or rational deliberations, in this process. In doing so, I complete the development of my model of moral decision-making: the affective-reflective model or ARM.

7.1. Being rational

Rationality, as Bratman (1987) puts it, is (p. 412):

> *[…] roughly, norms that enjoin or reject certain combinations of attitudes. These include norms of theoretical rationality that enjoin both consistency and coherence within one's beliefs. And these include norms of practical rationality that apply to intentions.*

The process of assessing combinations of attitudes, i.e., different types of beliefs and desires, is the reasoning process. It is quite clear that beliefs and desires are involved in the reasoning process. To demonstrate how they are involved, consider a scene from the TV show "The Good Place":

Donna is Eleanor's mother. Donna was a terrible mother to Eleanor and walked out on her when she was 16 years old. At 32, Eleanor found Donna and learnt that Donna has made a new family with Dave under a different name, Diana. In this new family, Diana is an attentive partner and a loving mom, a complete turnaround from the Donna that Eleanor knew. Eleanor decides to tell Dave, thinking that this is just another of Donna's scams; only to learn that Dave already knows, because Donna told him everything. Eleanor is surprised, but not convinced and insists that Donna must be running some sort of scam, and there is no possible way that she has changed. When asked why Eleanor would insist on that line of thinking, she

confesses that she does not want to believe that Donna has changed; because if Donna has, then Donna was always capable of change, and that means Eleanor was just not worth changing for.

Here we can see Eleanor's reasoning process: she combines different beliefs and assesses which one is false and should be eliminated. There are two points in this case interesting to our present purpose: the demand for consistency and coherency between Eleanor's attitudes, and how desires guide her reasoning process.

First, let us break down what are Eleanor's attitudes:

- Belief 1 (B1) - Donna has changed.
- Belief 1' (B1') – Donna has not change.
- Belief 2 (B2) - Donna can change.
- Belief 2' (B2') - Donna cannot change.
- Belief 3 (B3) – Donna can change for someone if they are worth her changing for.
- Belief 3' (B3') - No one can make Donna change.
- Belief 4 (B4) - Eleanor was not worth changing for.
- Desire 1 (D1) – For Eleanor to be worthy enough to make Donna change.
- Desire 2 (D2) – For Eleanor to believe that Donna cannot change.

If Eleanor perceives B1, there are two possible parallel lines of thought:

- Scenario 1 (S1) = B1 -> B2 -> B3 -> B4: If Donna changed, then Donna can change, then there is someone Donna deemed worth changing for, and that someone is not Eleanor.
- Scenario 2 (S2) = B1' ->B2' ->B3': If Donna did not change, then Donna cannot change, then no one can make Donna change.

Here, regardless of the truth of each premise, the arguments can be considered rational. The logical connections between premises in each line of reasoning are consistent and coherent. What separates being rational from being irrational is whether the combination of attitudes in question can be logically and consistently held simultaneously.

This idea of logical consistency is analysed further by De Sousa (1987) , by proposing six principles that we use to classify an attitude as rational. While Bratman finds rationality in combinations of attitudes, De Sousa finds it within the attitude itself; their ideas have a clear difference in object. Rather than assessing the rationality of a belief or desire in terms of its coherence with other beliefs and desires, we can say that specific beliefs or desires are themselves rational, or irrational. Nonetheless, an attitude's being rational, or irrational, still involves a sense of consistency or coherence. De Sousa formulates rationality's principles around the idea that each attitude has its own formal object: "the point of believing is to believe what is true. Similarly, the point of wanting is to want what is good" (p. 159). This means it would be irrational to believe what is false, and to desire what we believe is bad for us. In other words, De Sousa's idea refers to rationality of an attitude in the sense that includes the answer to the question 'is this attitude rationally consistent for me to hold?' i.e., is it consistent with believing what is true and wanting what is good. Therefore, De Sousa's principles can also be construed in such a way as that they apply to the combination between the attitudes in question and their respective rational attitudes. In "The Good Place" case, since the point of beliefs is to represent what is true, a rational attitude would be "what I observed is true". B1' is inconsistent with what Eleanor observed, the connection between B1' and what Eleanor observed is deemed irrational, therefore rejected. In De Sousa's terms, this would mean that B1' is irrational (for Eleanor to hold); in Bratman's terms, the combination of B1' and what Eleanor observed is irrational. With this conversion in mind, Bratman's and De Sousa's views are compatible with each other.

Amongst the principles of rationality proposed by De Sousa (1987), there are four principles of rationality that are important to our understanding of how rationality can be used to assess desires and beliefs. One, rationality assessment implies a standard of success; in Bratman's terms, there is something that belief or desire being assessed must be consistent with. Two, rationality assessment can be minimal in the sense that it can include cases where attitudes can be consistent with one another, under specific circumstances and assumptions, without requiring these attitudes to be true. This gives space for theoretical thinking. S2 presented above is one such case; another would be when we entertain the idea that the earth is flat because the perceivable horizon is flat. Three, rationality constraints are best thought of in terms of the rejection of irrationality. Finding an attitude rational does not necessarily make every other attitude with the same formal object irrational (p. 163). Rationality only rejects irrational cases, such as if we pair B1 and B1' (Donna changed, and at the same time did not change), or B2' and B3 (Donna cannot change and also she can for someone she deemed worthy to change for). And finally, there are two types of rationality, a cognitive type, and a strategic type; the frustration of one does not necessarily mean the frustration of the other. An attitude can be cognitively irrational (e.g., inconsistent with other attitudes) but strategically rational. Onions are not cubical; but if onions are cubical, cutting them would result in onion squares and not onion rings. And, obviously, this is also consistent with Bratman's distinction of theoretical and practical rationality. Strategic, or practical, rationality will be the major focus in the next section, as it is the standard of rationality most used in deliberation regarding motivation. With this as a foundation, it will be made clear how proper practical reasoning can enhance our affective process in the making of my model.

7.2. Practical reasoning and the Affective Reflection Model

As presented, rational thinking examines mental states for inconsistency. According to Hume (2009), our reasoning can affect our conduct, via our desires, in one of two ways (p. 700, elaboration added):

1. "[Reasoning] excites a passion by informing us of the existence of something which is a proper object of it."
2. "[Reasoning] discovers the connexion of causes and effects, so as to afford us means of exerting any passion."

Generally, I agree with Hume's assertions. As I understand it here, he means to attribute an 'analyst' role to our reasoning. We reason to analyse procedural beliefs in order, firstly, to assess whether a state of affairs is upsetting our desires, perhaps in an obscure way. Secondly, reasoning examines whether acting according to a procedural belief is consistent with realising our desires. Both of these roles are performed in relation to rational assessments of procedural beliefs. And as I have discussed in the last subsection, rational assessments look for inconsistencies in our attitudes. In practical reasoning, the kinds of inconsistencies our reasoning looks for are between causes and effects. So, when properly exercised, reasoning can tell us what the effects of our actions are likely to be; and from that, we can assess to what degree we desire those effects, which therefore alter our affective judgments and our motivations. Or as Hume puts it, reasoning excites our passions. In short, reasoning alters our motivations primarily by its ability to assess for inconsistencies between the causes and effects of our procedural beliefs.

Rational thinking processes look for consistencies in our procedural beliefs, such as "being in pain" and "being in danger". Suppose that I have a desire for safety, and that I think being in danger necessarily follows being in pain, so that if I am in pain, the desire for safety is excited. On the other hand, rational thinking can also be used to cut these connections between causes and effects, stopping our desires from being triggered by irrelevant factors. For example, my being in pain as I am lifting weights in a controlled environment, where I am not in danger. This informs me that the pain I am experiencing is not frustrating my desire for safety. In either direction this process takes, the connection between beliefs and a corresponding desire is reaffirmed.

This function can also relate to first and second-order desires. In Chapter Five, I mentioned Frankfurt's (1971) notion of reflective endorsement and second-order desires. Our first-order desires take the form of "I want X"; our second-order volitions take the form "I want to want X". Suppose that I want my society to have better support for homeless people (desire *∂*), and that this rationally requires me to pay more tax. I may have a desire to pay less tax, but to be consistent with my desire *∂*, I have to want to pay more tax. If reflecting on this inconsistency leads me to endorse the desire to support the homeless over the desire to pay less tax, then I will form a second-order volition to want to want to pay more tax. This second-order volition might connect to elements of my self-conception, for example, reflecting that I want to be a compassionate person.

However, beyond Hume's claim, I do not think reasoning is irrelevant in determining moral principles for actions. Hume's claim was about how our passions help us in distinguishing virtues from vices – by inciting pleasures or pains (2009). This is consistent with Reisenzein's desires-beliefs model for emotions presented in the last chapter. We believe that an action (A) is virtuous because we have a morally relevant desire (α) that A can bring about, therefore we feel a positive emotion about enacting A. In this case, rational assessment can be applied to the model I have developed as:

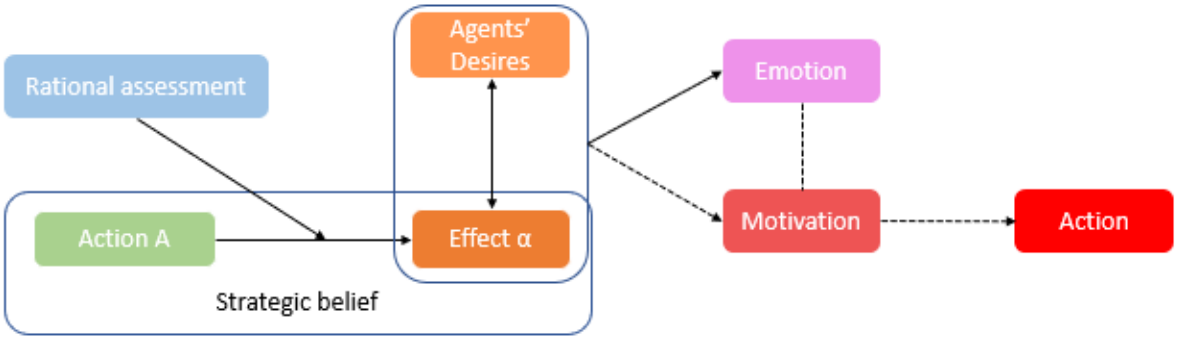

Figure 6 - Rational assessment in the basic model for motivation

Judgments from rational thinking here take the form of what I will call a possibility verdict – a judgement of whether it is realistic to achieve what we desire based on what we believe is true. At face value, possibility verdicts do not have much moral significance. Rational examination for means-ends inconsistencies can produce a verdict that some desire is irrational in a sense, but this only informs us if we are desiring something we believe to be impossible; like holding on to the desire to fly while believing we cannot. This kind of desire could affect our emotions and motivations, but not in a major way. In terms of first-order desires, a possibility verdict might have a more noticeable effect. If we only want to fly as a passing thought, a momentary desire, believing we cannot do so might snuff out the desire altogether. But suppose that the desire to fly is consistent with another desire that is important to me, perhaps the desire to break boundaries, so I form a second order desire – that I want to want to fly. In this case, believing I cannot fly does not serve as a final judgment, but as a starting point to realise my desire to fly. The reasons constituting the belief that I cannot fly are seen as problems I need to address in order to realise my desire to fly.

However, moral principles are not just about distinguishing virtues and vices; they are also concerned with pursuing virtues and avoiding vices. In making moral principles, rules, or regulations, the aim is to encourage virtuous (and prohibit vicious) ways of acting. And we come to know which ways of acting are virtuous, or vicious, by the effects they cause. So, in making a judgment about which ways of acting we should encourage, or deter, a level of certainty about these connections between causes and effects is required. These connections can be obscure to us, or false connections can be assumed by us. Practical reasoning examines our procedural beliefs to dismiss or reinforce connections between causes and effects. In properly examining these connections, we can increase the level of certainty in our judgment that one course of action is virtuous, and the other is vicious, through being certain about the effects they cause. Hence, the practice of reasoning is instrumental in determining what we should do.

Perhaps an example is warranted here. Suppose that I am hungry, and I want to satiate that hunger. To do so, I have three options: I can get dinner from Chicken Fillet (**A**), Korean Friend Club (**B**), or McDaisy (**C**). Eating at either one of these three restaurants will satiate my hunger (***∂***).

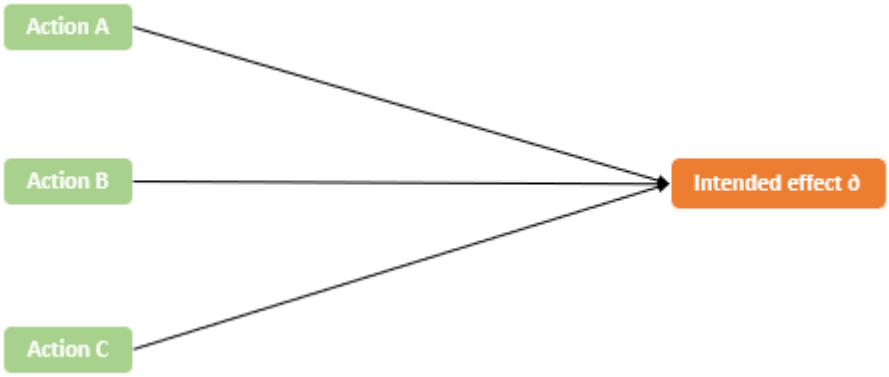

Figure 7 - The non-monogamous relationship between causes and effect (1)

Now suppose that I start to think more about these choices and draw more connections between these actions and their causes. Through some simple searches, I find out that, aside from serving food, Chicken Fillet is a business that is committed to support fair wage (**Ω**), and they also donate to gay conversion therapy practices (**Φ**). Korean Friend Club prices are pretty low, which means I can save money (**β**), and they are also committed to paying fair wages (**Ω**). McDaisy has pretty good reviews, so their food must be tasty (**α**), and they are cheap too (**β**).

This set of the effects of my choices about actions can be represented as:

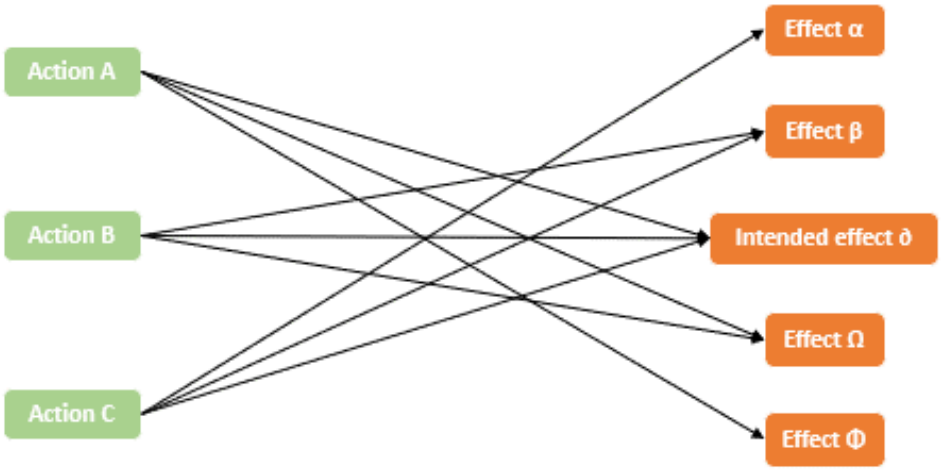

Figure 8 - The non-monogamous relationship between causes and effects (2)

- Action **A**: getting dinner at Chicken Fillet.
- Action **B**: getting dinner at Korean Friend Club.
- Action **C**: getting dinner at McDaisy.

- Effect ***∂*:** Satiate my hunger.
- Effect **α**: Have tasty food.
- Effect **β**: Save money (cheap).
- Effect **Ω**: Support fair wage
- Effect **Φ**: Support a business which donates to gay conversion therapy.

- **A -> *∂*+ Ω+Φ**: Eating at Chicken Fillet would satiate my hunger, support fair wage, and support a business which donates to gay conversion therapy.
- **B -> *∂*+β+Ω**: Eating at Korean Friend Club would satiate my hunger, help me save money, and support fair wage.
- **C -> *∂*+α+β**: Eating at McDaisy would satiate my hunger with tasty and cheap food

Based on the established connections, agents can then properly respond emotionally to these courses of actions (A, B, and C). These responses are indicative of their desires, and the course of action which satisfies their desires the most, both quantitatively and qualitatively, is the rational course of action to take.

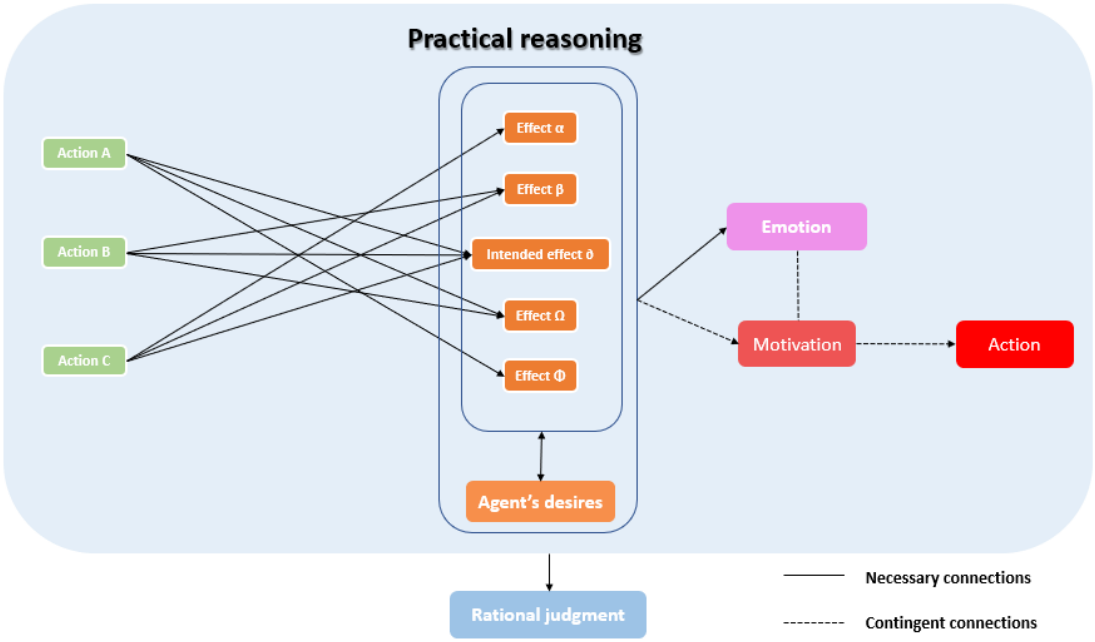

Figure 9 - The Affective Reflection Model (ARM) in decision-making

Suppose that, aside from ∂, I also find α, β, and Ω desirable, while Φ is repulsive to me. By fully establishing the connection between causes and effects, therefore enabling myself to have an appropriate emotional reaction to A, B and C, I can see that option B is a rational decision to make.

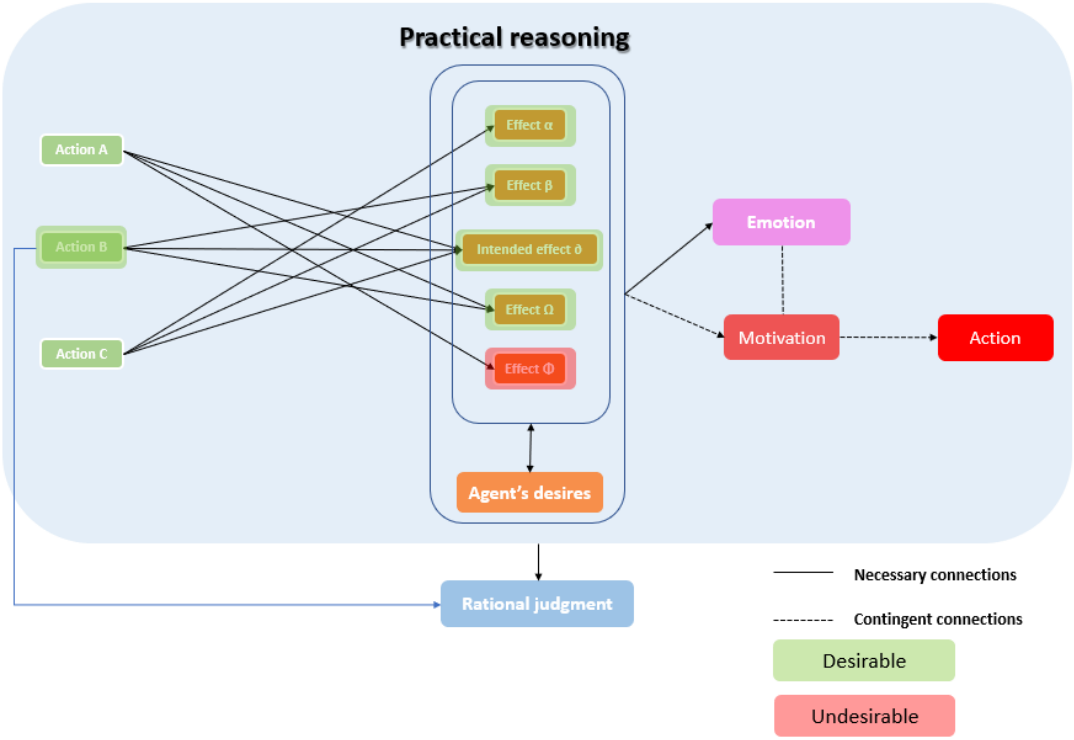

Figure 10 - Example of the ARM's function

Here Frankfurt's idea of a second order volition can be incorporated. For Frankfurt's idea, perhaps I have a second order volition regarding Ω (supporting fair wage); that is, I want to support fair wage and I want Ω to be my motivation. However, as this example indicates, agents are not always motivated by only one reason. Agents have multiple desires, and each course of action can simultaneously satisfy or frustrate a number of their desires. In this sense, having a second order volition can be applied in a weaker sense – that our second-order volitions can be satisfied if its object is included in the reasons which motivated us. B (eating at Korean Friend Club) can satisfy our second order volition regarding Ω because Ω is one of its effects.

In this sense, making connections between our means-ends beliefs and our desires, evaluating where these beliefs fit with our desires, and comparing our desires to see which course of actions can satisfy us the best, is giving rational consideration to these procedural beliefs. In drawing these connections, we can have appropriate emotional reactions to each way of acting, based on a full understanding of their effects. The option which triggers the strongest positive emotional reaction would be

the one that causes the highest level of desire-satisfaction (and conversely, the lowest level of desire-frustrations); thus, making it the rational judgment.

This is a more accurate representation of many dilemmas than two options that directly contradict one another, such as the Trolley Problem mentioned in chapter One. Of course, moral dilemmas we face in real life situations are much more nuanced than either the Trolley Problem or our example here suggests. For example, suppose that we want to build shelters for the homeless, and the only two companies who agree to take on our project are one who abuses tax loopholes, and another who lobbies against free healthcare. Should we commit to either of their services, which contribute to their wealth and support their conduct? Or should we hold out for another contractor, even if it means homeless people have to suffer more by staying out on the street longer, and possibly die? But for now, the example at hand will be sufficient for our discussion here.

While these are rational comparisons because they involve the assessment of consistency between causes and effects, as well as effects and desire-satisfaction, the comparisons are largely dependent on emotional processes. Rational comparisons are dependent on the values determined by our desires. The strength of our desires is often made conscious to us through our emotional responses, as discussed in 6.1. So, it is clear that emotional processes provide pivotal input to the rational process. Without these inputs, i.e., with neither desires nor emotions produced by them, the rational process cannot be completed, nor will agents have any inclination to act, as greyed out below:

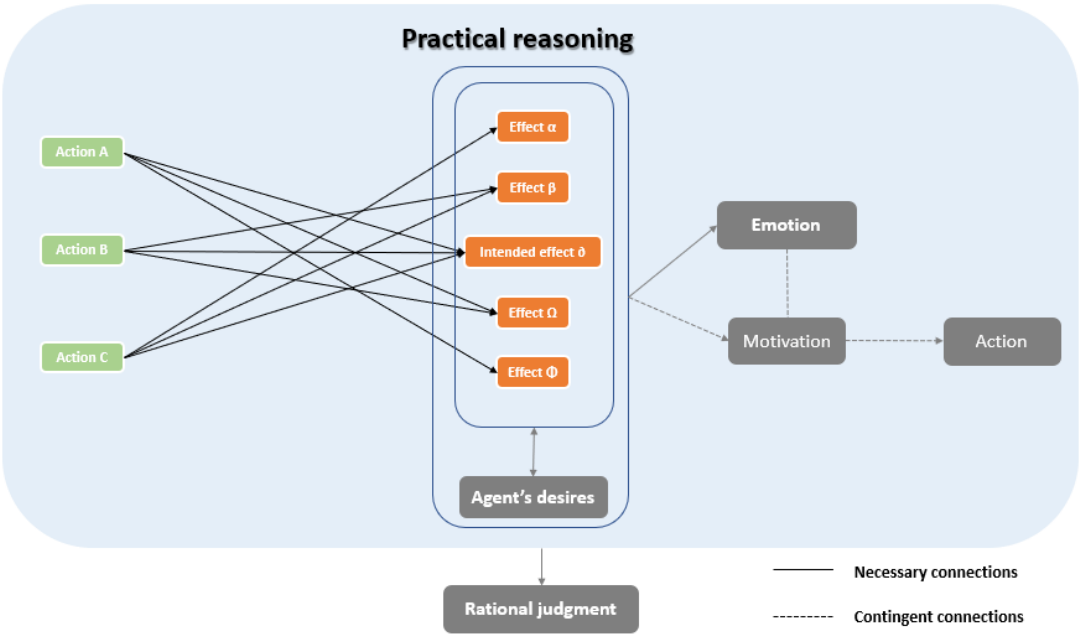

Figure 11 - The ARM without input from automatic affective processes

The making of an agent's best judgment is caused by the cumulative causation effect from the proper practice of both our reasoning and emotional responding. A cumulative causation process involves the interaction between mutually dependent factors, or in this case, processes (Argyrous 2011). It is part of practical reasoning that agents provide inputs to the feeling process. And the feeling process then produces emotions, giving them insights about their desires. These insights are, in turn, pivotal inputs to the practical reasoning process, which produces the agent's best judgment. In other words, acting on the course of action fuelled by the strongest (combination of) desires is acting on the rational, all things considered, course of action that we ought to take. Therefore, as far as practical decision-making is concerned, a rational judgment is an emotional judgment.

With the ARM fully drawn in this chapter, I have presented how the affective and the reflective processes can jointly cause an agent's best judgment. With the basic model drawn in chapter Six, I have discussed the function of reflective thinking, that is, to discover and assess the connection between causes and effects, and how reflective thinking can alter our affective judgments and motivation. This indicates

that reflective thinking enhances our affective responses, that our desires are better reflected in our judgments when we do reflect on them.

8. Reflective reasoning and morally desirable qualities

My aim in this thesis has been to develop a model of moral decision-making that moves away from the simple sentimentalist/rationalist dichotomy I summarised in Chapter Two; recognises the advances in understanding emotion outlined in Chapter Three; and incorporates what I regard as the successful aspects of others' approaches outlined in Chapter Four. Through the Affective Reflection Model (ARM) presented the previous chapter, we can see that both the affective responding and rational reflecting processes are parts of each other, and jointly cause an agent's best judgment. The ARM is a model that describes the intertwining of emotions and reasoning in decision-making processes, with emotions showing us the desirability of an action based on their probable effects, which reasoning discovers. This desirability becomes the standard of rationality for our decision-making. The model's purpose in this thesis is not to replace nor refute ideas presented in chapter Two to Four of this thesis – rather, the model is used to reconstruct these ideas in a unified framework. In this final chapter, I show how my model meets the above desiderata by looking at its overlaps with, and differences from, the accounts of Haidt, Kennett and Fine, and Jones; along with some remarks on the Sentimentalism versus Rationalism tradition presented in chapter Two.

First, in 8.1, I will discuss how the ARM can make clear the distinction between justificatory and reflective reasoning processes, which are usually conflated when Sentimentalists like Jones conceive of reflective deliberation. In reasoning, we assess procedural beliefs to see the extent of our actions' effects. But Haidt's SIM provides a notion problematic to this – agents can reason in service of a preferred judgment. Agents can, and often do (as Haidt noted), engage in justificatory reasoning processes, in which they assess their procedural beliefs just enough to make it seem like their preferred judgment is the rationally justifiable one. This calls for the separation between justificatory reasoning and reflective reasoning, the latter being a

process in which agents aim to assess their procedural beliefs to the fullest extent possible, i.e., to understand our actions fully.

In 8.2., I will discuss how the ARM can expand on the notions of agency and accountability raised by Kennett and Fine and Jones. Acting in accordance with agential desires, I want to point out here, is being motivated by agential desires. And in aligning our actions with the satisfaction of these agential desires, we need both our reflective reasoning capacity, to draw the fullest extent of our actions' effects, and our affective responding capacity, to identify whether these effects satisfy or frustrate our agential desires. The notion of accountability is better conceived using our intention, connected to our attitudes, as argued for by Sinhababu (2013). This will also capture nuances about different levels of accountability regarding different levels of intentions (Foot 1967). The importance of reflective reasoning is especially evident in higher moral decision-making processes that may affect collectives of people. Only in aiming to draw all cause-effect connections we can determine who is affected by the action in question; thereby taking into account their affective responses, making up the desirability of said action.

8.1. Justificatory reasoning versus reflective reasoning

In using the ARM to understand the roles of emotions and reasoning, we can identify a conflation about reasoning processes that is pivotal to understand the disagreement between the accounts of Haidt, Jones, and Kennett and Fine, presented in chapter Four. The conflation in question is between justificatory reasoning processes and reflective reasoning processes. While as Haidt says, sometimes we use reasoning in the service of justifying some emotional response we already have, I have shown in Chapter Seven that reasoning can also generate new emotional responses, or alter our overall emotional responses to a set of options for action. This implies that there is a kind of reasoning process, a purely reflective kind, which we must conceive of distinctly from justificatory reasoning. It is distinctive in that the desire underpinning the judgment that drives it is the desire for truth. Jones (2018,

p.267) mentions a "bullshitting" kind of reasoning, which one uses in service of ego needs, i.e., *the desire to be right*. In contrast, we can conceive of a reasoning process in service of truth – one which is motivated by curiosity, i.e., *the desire for truth*. While justificatory reasoning seeks to draw means-ends connections favourable to a pre-determined judgment, reflective reasoning seeks to draw *all* means-ends connections.

In a more social context, exchanges between multiple justificatory processes can achieve a similar effect to a reflective process, as Haidt's SIM implies. People can reason to support their pre-conceived judgments. Haidt (2001, pp. 820-822) likens this justifying role of reasoning to lawyers (reasoning) arguing for their clients (judgments) in court cases, and suggests that the role of reasoning is not analogous to a judge – a view echoed by Jones (2018) and Arpaly (2003). This is an appropriate analogy if we refer to the proper role of lawyers in the court system, which is not to win a case, but to raise all related considerations and evidence to contribute to a fair judgment. Analogously, a social justificatory process can endorse a judgment by relating it to the satisfaction and frustration of any (moral) desires we may have. This implies that, for Haidt, reflective reasoning does not play a major role in the making of our moral judgement because a social justificatory process can achieve similar effects.

Justificatory reasoning is also common in political debates. Most debates about policies concern what we consider being good for society at large, where what being good looks like differs between people. Often, people link a desire for something to the desire for their idea of what good is. Some link social equality to their idea of good, some link economic growth to their version of good. Once they do, they might gain a dogmatic vision, which demands the satisfaction of the desire they link to the desire for the good, no matter what it takes.

However, this social justificatory process still has an important drawback when we recall that agents' beliefs are on a spectrum of certainty. In justificatory reasoning, often agents only reason "enough" to support their preferred judgment – enough in

the sense that they cease to reason further when they have successfully made means-ends connections that are beneficial to their preferred judgment. The case of a dangerously thin anorexic person, used by Arpaly (2003) and Jones (2018), provides an obvious example here. There are multiple ways that anorexia can affect a person's judgment, but in this case, let us assume that the choice of whether to eat a piece of cake or not is between the desire to lose weight and the desire to live. Recall in 5.2, where we have discussed our attitudes' respective spectrums of strength, and that beliefs do come in a spectrum of probability. The anorexic person can see that abstaining from eating *can* lead to their death, but it does not mean that they think doing so *will* lead to their death. They can engage in reasoning more to determine the accurate probability of the consequences of their action. But they might just stop at reasons which encourage their preferred judgment. We can play out this stopped-short reasoning process such as "eating will make me fat, while not eating only might lead to my death – one is a causal connection, and one is a contingent connection. Causal connection is more reliable than contingent connection, so I ought to trust it. Besides, I would rather die than be fat."

Here, agents deliberate to find a sufficiently rational course of action to take, one where they can satisfy the desire that they have, while not (necessarily) frustrating any other major desires they might have. As Jones (2018, p. 267) puts it, agents are bullshitting themselves in a Frankfurtian sense – reasoning in service of ego needs, to deliver the answer which they antecedently wanted. We reason enough to persuade others or enough to justify a decision to ourselves, and no more. This use of reasoning is consistent with Haidt's SIM.

What seems to go wrong in these kinds of examples is *ignorance* - that the reasoning process ceases at a point where the person has justified their desire, while undertaking additional reasoning would show that the action they have decided for is not likely to best fulfil all of their relevant moral desires. Recall in 6.2.1. that saliency can be primed. This means, depending on context, some desires might be primed to be more important to satisfy than others; hence, increasing the intensity of

emotional experiences produced by these primed desires. For example, the desire for ideological freedom and the rejection of control are often primed in political debates, so much so that some would risk the health and well-being of others in order not to frustrate their desire. In this case, to support their judgments, they only need to reason enough to doubt the probability of the opposing judgment, that the well-being of others is not really at risk, or the risk is not high enough for them to sacrifice themselves. In both this example and the anorexic example above, agents are choosing to remain ignorant – which I presume is not a desirable quality for a moral judgment.

It is not problematic to assert that being ignorant is an undesirable quality of a moral judgment, or a morally good agent. To correct this kind of justificatory reasoning, agents need to offset it by a reasoning process with the aim of making all means-ends connections available to them – a reflective reasoning process. Only then can we have accurate emotional responses to a set of options for action, and see if we really do desire something enough to justify the negative effects it may cause. And it is only by reasoning reflectively that our judgments can have the qualities demanded in a moral judgment; qualities that I will discuss in the next section.

8.2. Why should we reason reflectively?

With the separation between justificatory and reflective reasoning processes, and that justificatory reasoning can lead to ignorance, made explicit, we can now discuss why reflective reasoning is beneficial to the making of a moral judgment. Using the ARM, I argue the reflective reasoning process is necessary for moral decision-making processes, albeit not in the overriding way, as Kennett and Fine might suggest.

The ARM presented in chapter Seven can further elaborate on the expression of agency in our judgment, a notion that is mentioned in both Kennett and Fine (2009) and Jones (2018). The primary concern raised by their arguments, as I take it, is

whether, and how, the reflective deliberation process is necessary in guiding our actions so that they exhibit our rational agency. Kennett and Fine's (2009) answer is that reflective deliberation is necessary, we must respond or reflect on our desires and not merely act upon them to exhibit our rational agency[16]. They hold that acting without reflectively deliberating is unsatisfactory towards holding ourselves as persons (pp. 85-86). Kennett and Fine use Frankfurt's idea of reasoning to form second-order volitions that separate our agential desires from non-agential desires[17]. They see the making of second-order desires as indicative of a person's agency in their decision-making, thereby explaining our practice of only holding agents capable of reasoning accountable for their actions (Kennett & Fine 2009, p. 86).

While I agree with Kennett and Fine that reflective thinking is important in our practice of holding people accountable, I find that relying solely on Frankfurt's idea of reflective endorsement to form second-order desires to claim normative primacy for a reasoned judgment lacking. It is not that the Frankfurtian picture is wrong – it is rather insufficient to explain our practices surrounding accountability here. Drawing the connection between the ability to reflectively endorse and the ability to express agency, and therefore the ability to be held accountable, is right. However, the account is insufficient to capture the entire expression of agency when it neglects the input of our affective process – which opens Kennett and Fine's account to criticisms like Jones'.

Jones (2018) agrees that rational agency is a necessary component of moral judgments, but adds that affective responding can also exhibit rational agency. She argues that reflective deliberation can play a regulatory role in our affective responses, hence rational agency can be expressed through them. The difference between Jones and Kennett and Fine on this point is that Jones does not see rational

---

[16] Here they used the terms *reasons for action*; I understand this term to mean reflectively endorsed desires. When we are asked to provide a reason for our actions, we refer to some end that we want to achieve – that is, a desirable end.

[17] Recall the distinction between agential desires and non-agential desires we discussed in 5.3.

judgments, which are supported by a deliberative process, to be the most indicative of a person's agency, as we often do practice justificatory reasoning which might not be expressive of our agency. Here, I take it, that Jones conflates the justificatory reasoning process and the reflective process when she spoke of a deliberative process. I agree that an endorsement from a justificatory process does not necessarily express our agency. But I think, and partly agree with Kennett and Fine, that our agency is better expressed in our judgments after a reflective process.

Here, I want to argue that the reflective reasoning process expands our expression of agency, conceived as our intention, by expanding our *awareness*. The notion of intention can better be used to capture which moral decisions do, and which do not really express a person's agency (Davidson 1971). In saying that I intended to do A with known effects α and β, I also assert that I intended for those effects to happen. And because I intended for them to happen, I would be responsible for effects α and β. If either α or β were a part of a causal chain that results in an unforeseeable harmful effect *∂*, then it would be unreasonable to hold me accountable to those effects. If I kick a ball to score a goal, and I do score a goal, then the act is attributed to me, to my agency. But suppose that someone blocks it with their face, and is injured. Then, causing that person's suffering would not be something I am held accountable for.

Sinhababu's (2013) account solidifies the connection between our attitudes and our intentions. According to him, intentions are constructed by our desires and beliefs in the following formulation (p. 680):

1. Agent A desires that φ.
2. Agent A believes that S will obtain, and that agent A's B-ing in S would make φ more likely.
3. If agent A were to believe that S obtained, the desire and belief would, without further practical reasoning, produce motivational force causing agent A to initiate B-ing.

This account proposes that we construct an intentional motivation with a central desire, coupled with a conditional means-ends belief, to produce motivational force when the condition for that belief is fulfilled. Taken with Davidson's idea above, this shows that agent's intentions, and therefore their agency, can be shown through their desires and how they believe they can achieve their desires – i.e., that reflective endorsement is not always necessary for the expression of agency. It is the agent's awareness of the effects of their actions, as well as their affective responses to them, that determines their intention and express their agency. Reflective reasoning expands this awareness, which enables more affective responses, therefore expands the expression of one's agency. The coupling of desires and procedural beliefs here is in line with the ARM in chapter Seven, which also relates it to the production of emotions and motivation.

Conceptualising the expression of agency by using the notion of intention also captures nuances about a different level of accountability that Kennett and Fine's account does not. As presented in 7.2., causes and effects do not have a monogamous relationship – an effect can have many causes, and an action can cause multiple effects. In this sense, forming a second-order volition seems pragmatically irrelevant for holding others accountable if the action is the same. While I agree that reflective endorsement is an important notion in expressing agency, it is hard to tell what it is exactly that people are endorsing in their actions. Suppose that I seek to punish a person who caused harm to me. If others wish to hold me accountable for my action of punishing (harming) said person, they must be able to discern whether the desires that motivated my action were desires for justice, or for vengeance, for there is a morally significant difference between them. But my reason for doing so, and whether it was endorsed by a second-order volition, is available to me alone. It seems to be hard, if not impossible, to hold each other responsible for our actions if we have multiple reasons to act a certain way, and which reasons are reflectively endorsed is only discernible by the agent.

Investigating one's intentions as their expression of agency, constructed by our desires and beliefs, can solve the aforementioned problem with Kennett and Fine's account in Chapter Four, that their account is unclear how our agency is expressed through reflective reasoning. While it might not be clear which desires an agent reflectively endorses in their action, we can discern their intention by investigating the alternative options available to them. Consider this part of our example in chapter Seven:

- A -> ∂ + Ω + Φ: Eating at Chicken Fillet would satiate my hunger, support fair wage, and support a business which donates to gay conversion therapy.
- B -> ∂ + β + Ω: Eating at Korean Friend Club would satiate my hunger, help me save money, and support fair wage.

Here, since ∂ (satiate my hunger) and Ω (support fair wage) are both present in A and B, choosing between them is choosing between Φ (support a business which donates to gay conversion therapy) and β (help me save money). Suppose that the agent is fully aware of these causes and effects, and chooses A, we can say that they intentionally choose Φ, which means this choosing speaks directly to their agency, hence subjects them to the accountability-holding by others. Alternatively, suppose that the agent here lacks the proper capacity to reason for whatever cause, and cannot draw the connection between A to Φ. Then they would not be held responsible for Φ. Pragmatically, this would be a plausible way to hold people accountable for their actions, based on their capacity to reason. So, without relying on the concept of second-order volitions, the analysis of a person's intention can also show a person's agency in their judgments.

However, there is a small expansion I must make with Sinhababu's account here regarding the role of desire in our intention. In how he constructs intentional motivation, the model requires agents to find an action's effect to be desirable for

the action to be intentional. But, again, an action can have multiple effects. This necessarily means that an action, while it satisfies some of our desires, can also have undesirable consequences. With Sinhababu's construction of intention above, it would seem that an end is unintentional if it is undesirable. But this seems unintuitive. Let us change our example above and only consider the choice between Ω (support fair wage) and β (help me save money). Supposed that I am not in any financial crisis, and I want to support fair wage, but I want to save money more – so the rational choice would be choosing β. In choosing β, then, it would seem that I also intentionally choose to not-Ω, not unintentionally like Sinhababu's construction would suggest (because not-Ω is undesirable). Simply by being aware of my own actions' undesirable consequences, and choosing to enact the action which causes them, even in pursuance of my desires, would make every consequence that I am aware of intentional.

But perhaps choosing to not-Ω here does not have the same sense of being intentional as choosing to β. Here, I do not desire to not-Ω intrinsically, but as a "collateral damage" in choosing to β. To explain this, we can use the distinction between direct intention and oblique intention Foot (1967) made, discussed in chapter Two[18]. In accordance with this distinction, choosing to β would be my direct intention, and choosing to not-Ω would be an oblique intention. Incorporating this distinction would enable us to add further nuance to the practice of holding others responsible for their actions, and its significance to our moral judgments, that Kennett and Fine draw our attention to. Using intention, along with the recognition that there are different levels of intent to show a person's agency, can help us make sense of different levels of accountability we hold others to. An example of this is different degrees of legal liability for murder we hold others to: one intentionally brings about the death of their victim in first degree murder, and one (obliquely)

[18] There is a long tradition relating to the Doctrine of Double Effect, but I will not investigate it in this thesis. Here I am merely using Foot's distinction of the two kinds of intention to make a connection to moral accountability.

intentionally puts others in danger's way (reckless endangerment) in second degree murder. This, I think, can capture the nuances about our practice of holding people accountable based on their capacity to reason better than Kennett and Fine's account, which relies on the Frankfurtian idea of forming second-order desires, can.

Acting in accordance with our agential desires means acting consistently with desires that are constitutive of our self-concept. And in aligning our actions with the satisfaction of these agential desires, we need both our reflective reasoning capacity, to draw the fullest extent of our actions' effects, and our affective responding capacity, to identify whether these effects satisfy or frustrate our agential desires. In this sense, our reasoning improves our emotions in their ability to exhibit our agential desires in our actions accurately. This is compatible with Jones' idea that reflective reasoning can regulate our emotional responses. This is largely because going through these processes, we will be able to establish a clearer hierarchy of desires for ourselves, i.e., decide which desires ought to be prioritised, and our improved pattern-matching of indicators that potentially frustrate high-level desires.

My account presented here is compatible with Kennett and Fine (2009) because it shows the importance of reflective reasoning by using the notion of accountability in one's action. But unlike their account, I am not raising the importance of reasoning by the making of our second-order volition. It is not that the Frankfurtian picture is wrong – it is rather insufficient to explain our practices surrounding accountability here. Drawing the connection between the ability to reflectively endorse and the ability to express agency, and therefore the ability to be held accountable, is right. However, the account is insufficient to capture the entire expression of agency when it neglects the input of our affective process – which opens Kennett and Fine's account to criticisms like Jones'. The role of reasoning in the expression of our agency presented here is to draw out the border of our intention – by expanding the scope of our affective responses and help us in determining that, in committing to act in a certain way, what are the intended (but not necessarily desired) effects and

what are not. However, this does not ascribe to reasoning processes the normative authority that Kennett and Fine (2009) do.

The account presented here is also compatible with Jones' (2018) trajectory model, in that our reasoning also contributes to the shaping of our habits, therefore improving the expression of our agency in our automatic affective responses. For Jones, this means that deliberative reasoning might not be necessary. However, we can use reasoning processes in a more active way than just as a monitor, as Jones suggested. As mentioned throughout this chapter, socially, means-ends connections can change based on the person at the receiving end of our actions. The subjective nature of other people's desires might be a contextual factor that we missed and failed to take into account. As monitors, our reasoning kicks in when we observe something that potentially contradicts our desires. We only reason as we need to; if our actions do not seem to be problematic to us (which they often do not), we rarely reflect on our actions. But this observation can be that our actions have already caused someone to suffer.

The moral significance of drawing as many means-ends connections as possible is even clearer when we consider that moral discourses are interpersonal processes. A decision to act in a certain way has moral saliency when the effects of said action affect others, and not just the agent who made said decision. This often takes the form of decisions that result in policies or laws, when the decision applies to society at large, or it would set a precedence that subsequent decisions must follow. Sometimes we even make moral judgments about others' personal choices – such as not exercising or what others consume. But those judgments are often on the social effects of those choices – for some, discouraging an unhealthy lifestyle is a moral concern. So, in making a moral decision, we must accurately draw means-ends connections to the furthest extent possible. The understanding of our action here not only draw the extent of their effects but also *who* are affected by our actions. This enables all those who are truly affected by said judgment to contribute to the

making of it, by determining whether an effect is desirable or undesirable, perhaps through a democratic process.

In properly assessing our means-ends reasoning, we can actively seek contextual factors which might cause our habitual judgments to result in the frustration of our desires or the suffering of others. Reasoning is still worthwhile to do, even if it has affected our automatic judgments via past reasoning. That is to say, to think before you act, while not always possible, is always a good thing.

9. Conclusion

In this thesis, I presented the Affective Reflection Model (ARM), a descriptive model of moral decision-making. I used it to demonstrate the moral significance of emotions and reasoning, by revisiting arguments made in moral philosophy, philosophy of mind and moral psychology through the lens of the ARM. With the moral significance of emotions and reasoning being presented using the ARM, it becomes clear that their roles in moral decision-making processes are not only mutually inclusive, but also pivotal in the satisfaction of a key demand for a moral judgment, the expression of agency, brought to attention in modern discourse by Kennett and Fine (2009). And as a further point, the presented mutual inclusivity of emotions and reasoning made it possible to conceive of these capacities in a unifying, and necessary, way for moral decision-making processes; this presents an alternative to the long-standing tradition of Rationalism versus Sentimentalism, which conceives them as opposing forces.

We began by examining the arguments of Hume and Kant as a starting point in **Chapter Two**. Hume argued that it is our passions that guide our moral judgments. Particularly, moral judgments imply motivation, and passions are what we use to determine what is virtuous and what is vicious, in a way that reasoning cannot, since it only pertains to matters of fact. Countering Hume, Kant asserted that, for moral appraisals of a person's character to be valid, these appraisals must be made on actions which agents committed to with their own free will (Rohlf 2018); a will that is uncontrolled by transient desires. For Kant, our human desires are a source of control – sometimes we succumb to our desires even if doing so warrants moral disapproval from others. Being a free agent requires acting in accordance with deliberative choice, rather than one's desires. This counterargument to Hume from Kant shaped the continuation of the Rationalism versus Sentimentalism tradition in modern discourses. The tradition of aligning moral demands with either emotions or reasoning, and use those demands to discredit the other capacity, echoed through

modern arguments: in Smith's (1987) claim that motivations are not intrinsic in moral judgments, therefore emotions are unnecessary in moral thinking; Shafer-Landau's (2003) defence against Mackie's (1977) argument against moral realism; and Green's (2009) argument that emotions latch on to morally insignificant factors which can skew our judgments. Notions of motivations and self-control from Hume and Kant's era are used throughout these arguments.

A major shift of this tradition, which is also important to this thesis, is how emotions were re-conceptualised as cognitive processes – much like reflective reasoning – by psychologists. This shift in the conceptualisation of emotions is presented in **Chapter Three** through three studies. The first is Ekman's (1999) formalisation of emotions as cognitive processes, followed by Damasio's (1994) observation of patients with damages to their Ventromedial Prefrontal Cortex, which shows the dependency of cognitive processes on affective responses. These two studies established that the line between reflective and affective capabilities in human beings is not as clear as it is often assumed in philosophical debates. Margolis (1987) put emotions and rational thinking, in terms of automatic affective responses and conscious reflective deliberation, in a linear "cognitive ladder". This constitutes two significant changes to the tradition of Rationalism versus Sentimentalism – the "rebranding" of emotions and reasoning as *automatic processes* and *controlled reflective processes*; and the arguments switched their focus to the necessity (or otherwise) of the controlled, reflective process.

From this development, hybrid accounts in moral decision-making started to form, which incorporate features of both processes to some extent in order to describe the functions and necessity of controlled reflective processes in moral judgments. We have examined three accounts of this kind in **Chapter Four** – Haidt's (2001) Social Intuitionist Model, Kennett and Fine's (2009) argument for the necessity of rational reflective thinking for the expression of agency, and Jones' (2018) Trajectory-Dependent Model of (Human) Rational Agency. While each theory

advanced the debate in important ways, in these hybrid accounts, terms used in the rationalist versus sentimentalist tradition still persists.

I believe that Sentimentalists and Rationalists have captured something true in their respective arguments. I think that they have adequately captured essential functions of our mental capacities in our moral decision-making processes. However, preserving a mutually exclusive sense in conceiving the functioning of these processes in moral decision-making is uncalled for. To advance the descriptive discussion on the roles that our mental capacities play in moral decision-making further, I have demonstrated how these cognitive processes work together in a way that is not mutually exclusive, and that they are capable of being mutually beneficial to one another. This is my aim in developing the Affective Reflection Model (ARM), as an evolution from the linear cognitive ladder proposed by Margolis, to a circular feedback process.

To develop the ARM, conceptions of desires and beliefs, which were examined at length in **Chapter Five** – including their core differences, classifications, and their own notions of strength – are the building block of the ARM. Built upon these conceptions of our attitudes is the formulation of the functions of emotions and reflective reasoning. In **Chapter Six** we discussed the function of emotions by Reisenzein's (2009) account, which saw emotions as signals of important changes in our attitude system, which is an automatic process. From this description of emotion, I drew a basic model of motivation. Then, in **Chapter Seven**, I examined the function of the reflective reasoning process in this basic model of motivation and set up the ARM. Reasoning examines the rationality of our attitudes – of whether, and to what extent, our beliefs are logically consistent, our desires are simultaneously satisfactory, and our procedural beliefs can satisfy our desires (Bratman 1987, De Sousa 1987).

Then, I drew attention to what I call the non-monogamous relationship between causes and effects to draw the ARM. In making a practical judgment, we are making a judgment on which means we ought to take to attain some ends (Paul 2015). In this process, our attitudes are examined rationally to enable agents to produce true-

to-self emotional reactions, which then become inputs for rational decision-making processes. Reasoning, at the same time, enables identification of the different potential means to our ends and their likely consequences, which can prompt further emotional reactions. These processes function in a feedback loop, or a Circular Cumulative Causation process – mutually dependent components interact and cooperate in their functioning (Argyrous and Stilwell 2011). Exercising controlled reflective thinking enhances our emotional reactions, and better emotional reactions produce better input for rational reflective thinking.

Through the ARM, I made two important inferences important to the role of emotions and reasoning in moral decision-making in **Chapter Eight**. The first notion was the distinction between justificatory and reflective reasoning. The first kind of reasoning is the use of reasoning to support a preconceived judgment or egotistical needs; while the latter being the use of reasoning in the pursuit of truth. This distinction, and therefore their respective functions, is often conflated in Sentimentalist accounts. In separating them, I have made explicit how the practice of reflective reasoning can expand an agent's awareness of their actions, of the effects that their actions cause. In deciding to carry out an action, while being aware of the effects that their action causes, agents show that they intended the effects they are aware of (Sinhababu 2013). Intention is one's expression of one's agency in their judgment. This is the foundation of my second point, that the expansion of an agent's awareness about their actions, which enables more affective responses by agents, the expression of agency is increased in their final judgment. This, as I argued, is the role of reflective reasoning, in conjunction with emotions, in the expression of agency in one's judgment.

Having presented how the expression of agency, a morally significant demand in one's judgment, is jointly caused by both one's affective responses and one's proper practice of reflective reasoning, I have demonstrated an alternative way to conceive the roles of emotions and reasoning plays in moral decision-making. This, I hope, will be able to pivot the Rationalism versus Sentimentalism tradition from

seeing the role of these mental processes in an antagonistic sense to a mutually inclusive and collaborative sense.

Bibliography

ANSCOMBE, G. E. M. 1963. *Intention,* Ithaca, N. Y., Cornell University Press.

ARGYROUS, G. & STILWELL, F. J. B. 2011. *Readings in political economy : economics as a social science,* Prahan, Tilde University Press.

ARPALY, N. 2003. *Varieties of Autonomy. Unprincipled Virtue,* New York, Oxford University Press.

BALTZLY, D. 2019. *Stoicism* [Online]. Metaphysics Research Lab, Stanford University. Available: https://plato.stanford.edu/archives/spr2019/entries/stoicism/ [Accessed 4 June 2021].

BRATMAN, M. 1987. *Intention, Plans, and Practical Reason,* Cambridge, Harvard University Press.

COHEN, L. J. 1993. What Has Probability to do with Strength of Belief. *In:* DUBUCS, J.-P. (ed.) *Philosophy of Probability.* Netherlands: Springer

DAMASIO, A. R. 1994. *Descartes' error: Emotion, reason and the human brain,* New York, Putnam.

DAVIDSON, D. 1971. Agency. *In:* BINKLEY, R., BRONAUGH, R. & MARRAS, A. (eds.) *Agent, action, and reason.* Great Britain: University of Toronto Press.

DE SOUSA, R. 1987. *The rationality of emotion,* Cambridge, Massachusetts, M.I.T. Press.

EKMAN, P. 1999. Basic emotions. *In:* DALGLEISH, T. & POWER, M. (eds.) *Handbook of cognition and emotion.* Chichester: John Wiley & Sons Ltd.

FOOT, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review,* 5**,** 5-15.

FRANKFURT, H. G. 1971. Freedom of the Will and the Concept of a Person. *Journal of philosophy,* 68**,** 5-20.

GAZZANIGA, M. S., BOGEN, J. E. & SPERRY, R. W. 1962. Some functional effects of sectioning the cerebral commissures in man. *Proceedings of the National Academy of Sciences of the United States of America,* 48**,** 1765-1769.

GREENE, J. D., CUSHMAN, F. A., STEWART, L. E., LOWENBERG, K., NYSTROM, L. E. & COHEN, J. D. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition,* 111**,** 364-371.

HAIDT, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review,* 108**,** 814-834.

HAIDT, J. 2012. *The righteous mind: Why good people are divided by politics and religion,* New York, Pantheon Books.

HAN, B.-C. 2017. *The agony of eros,* Cambrige, Massachusetts, M.I.T Press.

HELZER, E. G. & PIZARRO, D. A. 2011. Dirty liberals! Reminders of physical cleanliness influence moral and political attitudes. *Psychological science,* 22**,** 517-522.

HUME, D. 2009. Part I - Of Virtue and Vice in General. *A Treatise of Human Nature : Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects.* Auckland, New Zealand: Floating Press.

JOHNSON-LAIRD, P. N. & WASON, P. C. 1977. *Thinking : readings in cognitive science,* Cambridge, New York, Cambridge University Press.
JOHNSON, R. & CURETON, A. 2021. *Kant's Moral Philosophy* [Online]. Metaphysics Research Lab, Stanford University. Available: https://plato.stanford.edu/archives/spr2021/entries/kant-moral/ [Accessed 4 June 2021].
JONES, K. 2003. Emotion, Weakness of Will, and the Normative Conception of Agency. *In:* HATZIMOYSIS, A. (ed.) *Philosophy and the Emotions.* Cambridge: Cambridge University Press.
JONES, K. 2018. Towards a Trajectory-Dependent Model of (Human) Rational Agency. *In:* JONES, K. & SCHROETER, F. (eds.) *The Many Moral Rationalisms.* Oxford: Oxford University Press.
KAUPPINEN, A. 2018. *Moral Sentimentalism* [Online]. Stanford Encyclopedia of Philosophy. Available: https://plato.stanford.edu/archives/win2018/entries/moral-sentimentalism/ [Accessed 30 August 2021].
KENNETT, J. & FINE, C. 2009. Will the real moral judgment please stand up? : the implications of social intuitionist models of cognition for meta-ethics and moral psychology. *Ethical Theory and Moral Practice,* 12**,** 77-96.
KORSGAARD, C. M. 1989. Personal Identity and the Unity of Agency: A Kantian Response to Parfit. *Philosophy & Public Affairs,* 18**,** 101-132.
KORSGAARD, C. M. 1996. *The Sources of Normativity,* New York, Cambridge University Press.
LERNER, J. S. & TETLOCK, P. E. 2003. Bridging Individual, Interpersonal, and Institutional Approaches to Judgment and Decision Making: The Impact of Accountability on Cognitive Bias. *In:* SCHNEIDER, S. L. & SHANTEAU, J. (eds.) *Emerging perspectives on judgment and decision research.* New York: Cambridge University Press.
MACKIE, J. L. 1977. *Ethics : inventing right and wrong,* Harmondsworth, New York, Penguin.
MARGOLIS, H. 1987. *Patterns, thinking, and cognition: A theory of judgment,* Chicago, University of Chicago Press.
OAKLEY, J. 1992. *Morality and the emotions,* New York, Routledge.
PAUL, L. A. 2015. What You Can't Expect When You're Expecting. *Res Philosophica,* 92**,** 149-170.
PLATO (ed.) 2003. *The Republic,* London: Penguin Books.
PLATTS, M. D. B. 1979. *Ways of meaning : an introduction to a philosophy of language,* Boston, London, Routledge & K. Paul.
REISENZEIN, R. 2009. Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research,* 10**,** 6-20.
ROHLF, M. 2018. *Immanuel Kant* [Online]. Stanford Encyclopedia of Philosophy. Available: https://plato.stanford.edu/archives/sum2018/entries/kant/ [Accessed 30 August 2021].

SALTZSTEIN, H. D. & KASACHKOFF, T. 2004. Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique. *Review of General Psychology,* 8**,** 273-282.
SCARANTINO, A. A. D. S., RONALD. 2021. *Emotion* [Online]. Metaphysics Research Lab, Stanford University. Available: https://plato.stanford.edu/archives/sum2021/entries/emotion/ [Accessed 30 August 2021].
SCHWITZGEBEL, E. 2019. *Belief* [Online]. Metaphysics Research Lab, Stanford University. Available: https://plato.stanford.edu/archives/fall2019/entries/belief/ [Accessed 30 August 2021].
SHAFER-LANDAU, R. 2003. *Moral realism: A defence,* Oxford, Oxford University Press.
SINHABABU, N. 2013. The Desire-Belief Account of Intention Explains Everything. *Noûs,* 47**,** 680-696.
SMITH, M. 1987. The humean theory of motivation. *Mind,* 96**,** 36-61.
ZHONG, C.-B. & LILJENQUIST, K. 2006. Washing Away Your Sins: Threatened Morality and Physical Cleansing. *Science,* 313**,** 1451-1452.