Justice, Agency, and the Good: A Sartrean Approach to the Rawlsian Problem of Stability

Robin Pawlett-Howell

PhD

University of York

Department of Philosophy

February 2024

Abstract:

In this thesis, I adopt a Sartrean perspective to criticise and reformulate Rawls's response to the problem of stability. For Rawls, a conception of justice must be stable; that is, persons should be able to abide by the just organisation of social institutions. To achieve this aim within his framework, Rawls claims that persons can include justice within their conception of the good, defined in terms of a rational plan of life. In particular, Rawls argues that affirming justice within one's life plan expresses one's status as a unified self, secures the conditions for purposive agency, and enables the pursuit of one's major desires and interests. The culmination of these claims is the congruence argument: the Rawlsian conception of justice achieves stability because the shared sense of justice and the individual pursuit of a good life are congruent.

I criticise Rawls on three points. First, I contend that Rawls's position on rationality conflicts with his account of the unity of the self. Second, I demonstrate that Rawls commits to rationalistic voluntarism through his notion of life-planning. In contrast, I propound a Sartrean account of a non-localised, practical form of agency. Third, I argue that Rawls mischaracterises goodness; specifically, I show that a person's good cannot consist in their life plan since the process of life-planning is an extension of what the agent already takes to be good. Subsequently, I provide an alternative version of the congruence argument divided between, (i), a reflective endorsement of the Rawlsian society correlative to a particular self-conception and, (ii), the inclusion of justice within a fundamental project as a way of integrating and structuring one's other projects. In doing so, I offer a novel approach to exploring the issue of congruence *via* Sartre's existential phenomenology.

Contents

Chapter 5

Chapter 6

Chapter 7

Acknowledgements

This work was supported by the White Rose Arts and Humanities Council *via* a two-year full scholarship, taken over a three-year period. I recognise my great fortune in receiving that funding, and I hope to have done service to this opportunity.

First of all, I would like to thank my supervisors, Matthew Ratcliffe and Martin O'Neill. It is to their immense credit that I have completed this work. I am proud to have worked with them both.

Second, I would like to thank the community at the University of York Philosophy Department. My TAP member, Alan Thomas; because of his advice, I still aim for the standard of 'writing like Sartre'. Sam Dickson, we discussed a lot of philosophy over the years, and his insight has been invaluable. Ed Willems, Kendra Wegscheidler, Jacob O'Sullivan, and Rei O'Sullivan; I wish them only the best. Thanks also to Declan Hartness, Sean Hamill, Eleanor Byrne, Sarah Wood, Daryl Tyrer, Daniel Kim, Angelos Sofocleuous, Devon Howard, and Henry Knapper.

Third, I would like to thank the many students I had the privilege to teach while at York. I hope I repaid their insight and participation with an equal measure of enthusiasm. My thanks to Christian Piller, David Worlsey, John Blechl, and Chris Jay for facilitating this opportunity.

Fourth, I would like to thank my close friends. Joe Glendinning, I had the honour of serving as best man at his wedding. Jon Carrick, his advice made this project feel achievable. Thomas Bingley, I have relied on his friendship immensely. Robyn Stewart, she taught me how to write essays at undergrad; without her, I would have never pursued a PhD. Big Al, Danny Flynn, and Bede Batters, who have been with me from the start. Ishita Krishna, for the time we spent together. Cathy Davies, for getting me to care about social justice.

My deepest gratitude to Matthew Walton and Georgia Machen. Matty actually read my thesis, bless him, and he even claims to have understood it. Thank you, Matty.

My thanks also to Sol Clarke, Ben Walker, Marc Towse, Oscar and Nico Lihtola, Tracy and Richard Bingley, Conrad Wood, Joey, Ollie Drew, Haley Yocum, Isabelle Stenlund-Batters, Cameron Welsh, and the many others who have supported me during this time. If I have left anyone out, please know that I regret any omissions.

Finally, I would like to thank my family.

To my cousins Sean, Sam, Louis, Ashley, Heidi, and Ben; my brother Callum; my aunties Helen, Yvonne, and Tina; my granny Jean and grandad George; my uncle Mark; and my late uncle Steve. Thank you for being there. Special thanks also to my dad.

To my stepdad, it is because of you that I am studying Sartre; I appreciate you for reading my essays throughout the years. As for my mum, I can never repay your love, but such is its nature that repayment is never required. Thank you, for everything.

Author's Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Reduced and revised versions of chapter 7 were presented at 'The Second First Annual Boss Baby Symposium' and 'UK Sartre Society Conference 2022'. No other material has been presented or published elsewhere.

We are not examining the justice or the moral worth of actions from certain points of view; we are assessing the goodness of the desire to adopt a particular point of view, that of justice itself.

John Rawls

A Theory of Justice, 1971, p.568

What I did dimly sense was that one can't take a point of view on one's life while one's living it: it comes on you from behind, and you find yourself up to your neck. And yet, if you look round, you realize you're responsible for what you have lived – and that it's beyond repair.

Jean-Paul Sartre

War Diaries: Notebooks from a Phoney War, 1984, p.76

…The paradox of human life is precisely that one tries to *be* and, in the long run, merely exists. It's because of this discrepancy that when you've laid your stake on being – and, in a way you always do when you make plans, even if you actually know that you can't succeed in being – when you turn around and look back on your life, you see that you've simply existed. In other words, life isn't behind you like a solid thing, like the life of a god (as it is conceived, that is, as something impossible). Your life is simply a human life.

Simone de Beauvoir

Simone De Beauvoir: An Interview, Paris Review, 1965, p.37

Introduction

0.1 The Scholarly Background to Rawlsian Political Philosophy and French Existential-Phenomenology

In this thesis, my overarching aim is to cultivate a fruitful interchange between the Rawlsian approach to social justice and Sartre's existential phenomenology. Here, to anticipate this more general project, I shall demonstrate why these traditions have yet to enter into productive dialogue with one another, despite their shared philosophical heritage. I will also show that this interchange remains an open possibility that is worth exploring in more detail. To begin with, I shall provide a broad overview of the literature concerning both Sartre and Rawls. In doing so, I will explain the lack of scholarly work on Sartrean-Rawlsian philosophy. To round this discussion off, I will clarify how this thesis provides new grounds for a phenomenological approach to the Rawlsian conception of justice. Thus, I shall use this introductory analysis to anticipate one of the main contributions of this thesis – identifying room for a Sartrean-Rawlsian approach to justice – whilst also shedding light on some of the current scholarly work within this area of thought.

When analysing the literature on Rawlsian liberalism and Sartrean phenomenology, two things become clear. First, not much work has been done to bring Sartre and Rawls into contact. Second, when these thinkers or their respective traditions do meet, it is generally the case that phenomenologists criticise the Rawlsian approach to political philosophy. Concerning the first point, McBride (1980, p.31; 1991, p.15) highlights the lack of engagement with Sartre and Merleau-Ponty in Anglo-Saxon political philosophy, despite both thinkers being actively engaged in political discourse throughout their respective works:

> I would like to ask readers of books such as Dworkin's or Rawls's or even Nozick's to consider the references to Continental social and political philosophy to be found in them. One will not find any references in any of the…to Sartre, to Merleau-Ponty…I have checked this carefully to make absolutely certain of it, though the very thought that some such references might exist seems ludicrous on the surface of it to anyone who is familiar with the nuances of the current atmosphere…But is that not sad? (McBride, 1980, p.31).

To strengthen this point, Laborde's (2002, pp.131-146) analysis of the reception of Rawls's work shows that Rawlsian liberalism was often separated from other so-called "continental" bodies of thought. Instead, at least in French academic circles, Rawlsian philosophy was "first studied in departments of law and economics" and often confined to "disciplines perceived as bastions of 'apolitical' conservatism" (Laborde, 2002, p.141). Hence, the international impact of Rawls's work must be tempered in light of such nuances: Rawlsian scholarship was often viewed as a specialised area of study within continental academic circles, leading to a separation from other prominent areas of philosophy – including that of existential phenomenology. In this sense, Laborde's investigation regarding the reception of Rawlsian philosophy in Europe accords with McBride's description above: when considering the overlap between Sartre and Rawls, one is initially confronted by an atmosphere of mutual apathy.

Hidden just behind this atmospheric disinterestedness, however, several critical themes emerge concerning both Rawls and Sartre. From the outset, Sartre and his contemporaries (e.g. De Beauvoir and Merleau-Ponty) maintained a healthy distrust of liberalism both as a politics and as a theory of normative ethics. In their understanding, the traditional liberal notion of freedom was an abstraction, a myth, de-historicised and withdrawn from the practical context of day-to-day life. As a result, the French phenomenologists charged liberal political philosophy (more broadly, rather than the Rawlsian tradition) with exchanging practice for principles, thereby failing to make solid the grounds for meaningful political change. Though this line of thought can be teased out of Sartre's later work, it is particularly evident in Merleau-Ponty's discussion here, wherein the putative abstraction enacted by liberal thought is directly put to task:

> An aggressive liberalism exists…it can be recognized by its love of the empyrean of principles, its failure ever to mention the geographical and historical circumstances to which it owes its birth, and its abstract judgments of political systems without regard for the specific conditions under which they develop (Merleau-Ponty, 1969, p.xxiv).

Additionally, a brief examination of contemporary phenomenology reveals that this sentiment has carried on and has since been aimed squarely at Rawls's work. To further develop the themes above, Whiteside (1988, p.290), in exploring Merleau-Ponty's political philosophy, claims that "Rawls generates his principles of justice by conceiving them as the choice of perfectly rational, ahistorical, disembodied consciousnesses." Here, the issue at play is the '*original position*', i.e. Rawls's contract scenario wherein persons – taken from the contingencies of life that might normally affect their judgements – come to an agreement over principles of justice (discussed briefly in section 2.1 of this thesis). This same scenario is the target of Cross's critique, which uses Sartre and strands of communitarian philosophy to criticise Rawls's account of the self:

> If Rawls allowed that the self was an intersubjective self, partly constituted by membership of a community, then it would be possible to talk of common talents being used for common goals. However, Rawls cannot make this move because to do so would invoke the notion of the radically situated self (Cross, 2017, p.187).

As these references show, on the scant occasions when there is a dialogue between Rawls and existential phenomenology, it is critical rather than jointly productive. Yet, understood in a broader historical context, this general divergence ought to seem puzzling. As Bercuson's (2014) work demonstrates, there are Rousseauvian and Hegelian roots to Rawls's account of justice. That is, although these thinkers themselves are not phenomenologists, there is a discernible pathway from Rousseau to Hegel – and even other thinkers loosely connected to both traditions, such as Marx and Freud – that can be traced towards Rawls's (e.g. 2000, 2007) political philosophy. And yet, a similar pathway leads up to Sartre's existential phenomenology. Broadly speaking, the existentialists were heavily influenced by Kojève's (1969) Marxian reading of Hegel. Similarly, Darnell and Rohatyn (1992, p.255) write that Rousseau's "influence on Sartre is so obvious that it is rarely discussed." This is to the point where, in *Anti-Semite and Jew*, Sartre (1948, p.105) states his support for a form of "concrete liberalism" wherein the rights of persons, and presumably their status as citizens *proper*, depends on their "active participation in the life of the society." This description, though one would never tell from the secondary literature on Sartre and Rawls, is not a million miles away from a Rawlsian liberalism.

Given the above, a dialogue between Rawls and Sartre is not an implausible prospect, bearing in mind Sartre's early sympathy towards liberalism and the shared philosophical background of both traditions. However, contemporary existential phenomenologists engaged with Rawls have overlooked this possibility – with very few exceptions, such as Banerjee and Bercuson (2015). The claim I shall support now is that scholars have failed to foster a productive dialogue between Sartre and Rawls because of a misguided focus on Rawls's original position. To this end, recall that both Whiteside (1988) and Cross (2017) criticised Rawls's de-situated and de-historicised account of persons precisely as it is presented in the original position. More instructively, Kruks provides direct support for my analysis by situating this divergence as a broader trend intersecting both continental phenomenology and Anglo-Saxon political philosophy:

> …Anglo-American philosophy has, with some exceptions, remained militantly impervious to foreign importations.… Thus Rawls, to take the most influential example, simply tells us at the beginning of *Theory*...that

> when we formulate principles of justice, they must be acceptable to 'free and rational persons concerned to further their own interests', as if the notion of 'free and rational persons' were wholly unproblematic… (Kruks, 1990, p.2).

Both from the description provided and the reference to the early portion of *A Theory of Justice,* it is clear that Kruks is again referring to Rawls's 'original position'. What is less clear is *why* those within the existential tradition focus almost exclusively on this part of Rawls's account of justice. From a Rawlsian perspective, the notion of free and equal rational persons in the original position *is* unproblematic essentially because of the contract scenario's hypothetical nature; "this original position is not, of course, thought of as an actual historical state of affairs… It is understood as a purely hypothetical situation" (Rawls, 1971, p.12). Understood as an imagined scenario, the original position is only a procedure designed to tease out our moral convictions so that we can come to an agreement on justice. Consequently, existential phenomenologists criticise features of the original position that are not only already known by Rawlsians, but that are necessary for it to function as a thought experiment.

Given this additional context, it is evident that the dialogue between Rawlsian political philosophy and phenomenology has stagnated, and why it has done so. Both traditions have been talking past one another. Scholars from the phenomenological tradition, and I should add they are not alone in this trend, have all too often focused on a relatively small (though undeniably well-known) aspect of Rawlsian philosophy: the original position. Whilst "Rawls scholars", in response to the varied impact that *A Theory of Justice* had on European thought, "had to carve out a niche amid a diversity of schools and approaches" (Laborde, 2002, p.139). Yet, this divergence does not mean that phenomenology has nothing to offer Rawls, only that – if it is to provide something of value – the discussion as a whole must be redirected. It is misleading to suggest that Rawls and Sartre share no common ground, just as it is misleading to take the original position to be exhaustive of Rawls's political philosophy. For this dialogue to progress, there must be a meaningful shift in focus.

In this thesis, I propose that both traditions would benefit from a shared concern with the Rawlsian problem of stability, and I provide the groundwork to accomplish this aim. That is, phenomenology can be a fruitful lens through which we come to analyse the instantiation of Rawls's conception of justice and its impact on persons. Understood in relation to the prevailing phenomenological critique of Rawls, the move towards the problem of stability ought to seem relatively intuitive. As Hill (1999, p.854, emphasis added) puts it, "in various ways, explicitly or implicitly, political and legal theories make use of conceptions of what people are *actually* like." In the context of Rawlsian philosophy, no one is 'actually' like the parties of the original position – but the same cannot be said about Rawls's response to the problem of stability. Specifically, Rawls's account of persons is founded principally on his description of '*the good*' – that is, very broadly, the values and interests that shape each person's version of a good life – and its relation to the affirmation of justice *via* the congruence argument. As this thesis will demonstrate, Sartre's phenomenology is especially relevant to the congruence argument since it helps to expose the essential normative dimensions of self-awareness, thereby giving expression to the ways in which a person's conception of the good is formed and pursued.

As a final note, the turn towards stability also explains another decision I have taken in this thesis: my exclusive focus on Rawls's (1971; 2005) earlier work, *A Theory of Justice*, rather than *Political Liberalism*.[1] I take this approach because late Rawls would eventually abandon the congruence argument. According to late Rawls (2005, p.xvi), the aim of congruence would have meant that all persons in the well-ordered society must affirm the same "comprehensive doctrine" in order to incorporate justice within their version of the good. Yet such a requirement is an impossibility, given that society will foster a plurality of viewpoints and belief systems (Rawls, 2005, p.xvi). The account of congruence advanced in this thesis challenges that position. Instead, on the Sartrean-Rawlsian approach, to strive for congruence between justice and goodness does not mean that everyone accedes to the same comprehensive doctrine, only that justice transforms and regulates one's values and ends. Ultimately, this thesis not only addresses the silence between liberalism and existential phenomenology but also mends a dispute internal to Rawlsian philosophy by assessing and reconstructing the congruence argument.

0.2 Summary of Argument

This thesis has three main objectives. First, I aim to provide a detailed and plausible account of Rawls's primary response to the problem of stability: the congruence argument. To this end, chapter 1 establishes what the problem of stability *requires*, thereby identifying the conditions for a just society to be stable. As I clarify, the problem of stability is a practical issue of whether a conception of justice – that is, the distributive principles that organise societal institutions at the highest level – can be implemented within society. I argue from the Rawlsian perspective that the well-ordered society (i.e. a society organised according to Rawls's principles of justice)

[1] From here on, I will shorten Rawls's (1971) '*A Theory of Justice*' to '*Theory*'.

is only stable when it can generate supportive attitudes amongst its citizens.[2] On the back of this, I also show that justice as fairness – the name that Rawls gives to his own conception of justice – appeals to the status of persons as purposive agents to encourage the attitudes necessary for stability.

In chapter 2, I unpack the congruence argument independently; namely, Rawls's claim that – in the well-ordered society – one's sense of justice (i.e. the disposition to act as justice requires) is congruent with one's conception of the good as defined through a rational plan of life (i.e. one's version of a life well-lived).[3] In this discussion, I respond to Barry's (1995) criticism of Rawls: that 'doing what is right' should be sufficient motivation for persons to act in accordance with justice. In response, I show that whilst, for some people, doing what is right is sufficient reason to act justly, the congruence argument is nevertheless an appropriate stress-test for a conception of justice. Since stability is a practical problem, persons within the well-ordered society will have a diverse range of interests, values, frameworks of understanding, and so on. Understood in this way, congruence confirms that persons can affirm their sense of justice even in this complex environment. Additionally, since congruence shows that justice is harmonious with one's practical and personal interests, persons are assured their sense of justice is not rooted in envy, self-denial, subservience to authority, or deceptive ideological mechanisms.

My second major aim is to demonstrate that a person's conception of the good cannot be given in their rational plan of life – which is what Rawls uses to characterise a person's good and their status as agents – nor does it constitute a tenable account of agency. In chapter 3, I undertake a detailed analysis of Rawls's account of rational life-planning. I provide the empirical background to Rawlsian life plans and explain the mechanisms constitutive of one's life plan (e.g. the level of planning involved, the form of hierarchy, the principles of deliberation, and so on). Throughout this chapter, I also identify the roles that rational life-planning plays within justice as fairness. In doing so, I highlight that Rawls uses rational life-

[2] Rawls (1971, p.5) defines a well-ordered society as a society "in which (1) everyone accepts and knows that the others accept the same principles of justice, and (2) the basic social institutions generally satisfy and are generally known to satisfy these principles." In this thesis, my exclusive focus is on the Rawlsian conception of justice. As a result, I use 'well-ordered society' to refer to a society organised according to Rawls's principles, meeting the conditions just mentioned.

[3] Throughout this thesis, I sometimes use the terms 'goodness' or 'the good' to capture the idea that each person has their own conception of the good. Using these terms, Rawls looks to establish congruence between justice (i.e. affirming the Rawlsian principles of justice) and goodness/the good (i.e. pursuing one's own version of a life well-lived). In the context of the congruence argument, goodness is not a moral notion. Rather, it's a person's main values, interests, desires, purposes, etc. that constitute their version of a good life.

planning to ground equality amongst persons, define a happy life, secure the priority of liberty, and account for the unity of the self. Beyond this, I draw special attention to Rawls's endorsement of Royce (1908, p.169) and their shared position that a person, as well as their self-unity and conception of the good, is defined through their rational purposes given in a coherent life plan.

In chapter 4, I begin critically interrogating Rawls's account of rational life plans. I examine one of the main lines of criticism against Rawls: that life plans, given their global nature, are temporal abstractions unable to account for time-sensitive goods and virtues (I refer to this line of argument as the 'temporality critique'). I argue that the temporality critique fails to undermine Rawls's position on life-planning, which is better understood as a deliberative process, rather than an atemporal abstraction. I then advance my own criticism: that Rawls's account of rationality (*qua* 'rational' life plans) involves a form of evaluation abstraction. That is, within the congruence argument, persons are required to withdraw from their immediate interests, values, and so on, to identify with the good they would endorse had they had access to all the relevant information. Although I accept that a more modest account of rational life-planning can resolve this issue, this chapter draws out a tension between Rawls's conception of rationality and his account of the unity of the self. Additionally, I solidify the view that the congruence argument appeals to actual citizens and their internally rational plan of life.

In chapter 5, I argue that Rawls endorses an unworkable account of agency. Initially, I demonstrate that Rawls's account of life-planning and self-reproach – which I broadly define as a negative sentiment arising from the failure to treat oneself as a unified agent – commits him to a form of rationalistic voluntarism. Subsequently, I draw on Sartre's (2018) description of a non-localised, pre-reflective, form of agency (further elaborated upon in chapter 7) to show that self-reproach involves a range of normative phenomena that cannot be explained *via* a voluntaristic account.[4] Instead, the experience of self-reproach must be embedded within project-like structures, wherein one's sense of agency is not given as an isolated, localised act. In developing this line of argument, not only do I provide evidence that Rawls commits to a voluntaristic account of agency, but I also show why it is ultimately unworkable, even within the confines of his conception of justice.

[4] For Sartre's existential-phenomenology I rely primarily on his early work *Being and Nothingness*, first published in 1947. In this thesis, I reference the 2018 English translation by Richmond.

In chapter 6, I argue that the Rawlsian account of life-planning – even in its more modest, processual form – does not provide an adequate account of the good (i.e. characterising each person's version of life well-lived). As the previous chapters establish, Rawls's notion of rational life-planning depends on a deliberative, reflective, form of awareness. To advance my analysis, I draw on Sartrean phenomenology to tease out the epistemic and normative implications of reflective and pre-reflective consciousness. I put forward two main claims against Rawls's approach. First, that life-planning extends from what the agent already takes to be good and so cannot characterise it. Second, that one's view of oneself within a life plan involves a specific form of self-awareness, which does not exhaust its subject and is prone to misrepresenting it. Accordingly, I employ Sartre's account of pre-reflectivity and the epistemic errors associated with reflection to demonstrate that the life-planning approach fails to characterise goodness.

My third and final major aim in this thesis is to reconstruct Rawls's argument from congruence. In chapter 7, I tie up some loose ends regarding Sartre's account of pre-reflectivity. In doing so, I distance my contribution from substantive metaphysical doctrines to make my overall approach more palatable to Rawlsian liberalism. Using Ratcliffe's (2017; 2024) work, I argue that Sartre's account of pre-reflective agency is a plausible phenomenological thesis – which I treat as analogous to the psychological assumptions already operative in Rawls's theory – about the overall structure of consciousness. Additionally, since Sartre has also been charged with a form of voluntarism, I respond to these criticisms and address key points within Sartre's account of agency. Specifically, I demonstrate that one's pre-reflective awareness involves a basic, world-directed sense of possibilities. Similarly, I shed some light on Sartre's notion of global responsibility by highlighting that our projects involve a sense of ownership, are subject to acts of affirmation or acceptance, and are themselves embedded within broader projects.

Finally, in chapter 8, I provide a sketch for an alternative Sartrean-Rawlsian account of congruence. I open this chapter by examining Rawls's move away from congruence in his later work, *Political Liberalism*. Importantly, I clarify that it is possible – given the practical requirements of the congruence argument as detailed in chapters 1 and 2 and the use of phenomenology to respond to such requirements – to establish the compatibility of justice and goodness without imposing a substantive doctrine on all such citizens (*pace* late Rawls). Subsequently, I divide the Sartrean-Rawlsian congruence argument into two main parts. First, *reflection without rationalism*: persons can come to reflectively affirm justice on the basis of their own practical self-conception. As I highlight, we normally reflect on ourselves when we encounter some disturbance in our lives. In the case of congruence, the claim is that – when we reflect on the scope of justice – we discover that it fits with our view of ourselves as self-authenticating, practical agents. Second, *justice as a fundamental project*: justice can be affirmed as a project that regulates and structures my other projects. Hence, justice and goodness are congruent once the former consists in a project that integrates the values and commitments that I *am*.

Ultimately, this thesis fundamentally reassesses Rawls's congruence argument from a Sartrean perspective. I provide a detailed investigation into the working parts of Rawls's argument, scrutinise its commitments, expose its shortcomings, and then reimagine it through the work of Sartre. In doing so, I provide a positive contribution to the dialogue between phenomenology and liberal political philosophy. Even the critical portions of this thesis contribute to this overarching intention by locating the key areas where phenomenology *can* say something fruitful about justice as fairness. Beyond this, I also draw attention to the ways in which our lives are dynamic, multifaceted, complex, and ultimately impenetrable. Throughout this thesis, I uphold Sartre's sentiment that we exist through our commitments. As Sartre highlights, such commitments always slip through our fingers, carrying us off to some place new. In the same way, our respective versions of a good life – and the values, the purposes, the interests, the desires, that make it up – are similarly elusive. One cannot plan a worthwhile life; it is simply lived as such. This thesis looks to recognise this fact and consolidate its importance to the pursuit of social justice.

Chapter 1

Stability, Indeterminacy, and the Two Generative Properties of Justice as Fairness: The Stabilising Forces of the Well-Ordered Society

1. Introduction

This chapter answers two main questions. First, 'what is the problem of stability?' And second, 'what are the forces necessary to secure against instability?' I respond to these issues from a Rawlsian perspective, such that my analysis can be understood as *providing the background* to Rawls's (1971, pp.395-587) two main arguments from stability – i.e. 'the argument from moral development' and the 'congruence argument' respectively; both of which shall be investigated in chapter 2. In this chapter, I consider two candidate stabilising forces for justice as fairness: 'enforcement' and 'reciprocal attitudes.' I investigate *why* Rawls's account of stability is compelling to citizens who have yet to fully develop a Rawlsian sense of justice. Ultimately, my position is that the well-ordered society is stable because of what I call *the two generative properties of justice as fairness*: citizens will realise their status as purposeful and social beings by acting under the constraints of justice, and they will come to see their institutions as valuable independent of their private ends. In other words, justice as fairness is stable because it engenders within its citizens a fundamental re-evaluation of their attitudes.

This chapter proceeds as follows. Firstly, I provide context to this discussion by highlighting the importance of stability, clarifying what the problem of stability requires, and defining a stable society through Rawls's notion of equilibrium (section 1.1). In doing so, I situate the problem of stability as *a practical problem* for a conception of justice, which verifies that the demands of justice can be fulfilled and identifies the forces necessary for an ongoing state of equilibrium. Secondly, I put forward a candidate for the stabilising forces within justice as fairness: a Hobbesian-like system of enforcement (section 1.2). I do this by demonstrating how a system of enforcement resolves the main problems of stability and by providing textual evidence that Rawls saw the value in Hobbes's approach. Thirdly, I draw out a tension by contrasting Rawls's endorsement of Hobbes (1996) with his attempts to distance justice as fairness from *Leviathan* (section 1.3). Fourthly, I put forward an alternative account of stability along the lines of the two generative properties of justice as fairness, i.e. that citizens will transform their private ends and view just institutions as independently valuable (section 1.4). In doing so, I draw on McClennen's (1989) description of a non-Rawlsian society and use his analysis to consider the motivations necessary for its citizens to move towards justice as fairness. Finally, I point out an indeterminacy problem between the two respective accounts, enforcement and the generative properties (section 1.5). I resolve this indeterminacy by clarifying the role of enforcement within Rawls's work whilst affirming the two generative properties as the primary stabilising force within justice as fairness.

Ultimately, this chapter demonstrates that Rawls's (1971, p.572) response to the problem of stability depends heavily on the view that persons will want to change the structure of society because, in doing so, they will come to realise an essential part of themselves; "the desire to act justly and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire." Throughout this discussion, I highlight the broad features of Rawls's response to the problem of stability. Most prominent among these features is the attitudinal changes that justice as fairness is meant to engender, *via* the two generative properties of justice as fairness, and the appeal to the status of persons as purposive agents. The benefit of this analysis is that it prefigures Rawls's argument from congruence. That is, Rawls appeals to the relationship between justice and goodness to promote attitudes of reciprocal support between just institutions and citizens of the well-ordered society. At the broadest scale and permitting additional nuances, I detail the mechanisms that Rawls's argument from congruence leverages to establish a stable society.

1.1. The Practical Requirements of a Conception of Justice

Generally speaking, before a person commits to a course of action, they are confronted with the question of whether their decision is a plausible one, e.g. whether it can be achieved and sustained in the long-run. For instance, if someone makes a major financial decision, they may have to assess it in light of the peaks and troughs of the economy as a whole or, at a minimum, as a long-term consequence of their own personal finances. In any case, what this person is normally doing is considering the relative stability of their choices. If they have reason to think that their choice will be an unstable one, *then they also have reason to reconsider taking it*. At least in the Rawlsian picture, the same can be said about a conception of justice; if a conception of justice is unstable, "this fact must not be overlooked. For then a different conception of justice might be preferred" (Rawls, 1971, p.145). Here, whilst Rawls (1971, p.3) famously characterises justice as "the first virtue of social institutions", stability can be characterised as the second most virtue of institutions or the first virtue of a conception of justice. That is, stability is of utmost importance because it verifies that a conception of justice can be implemented and maintained in society.[5] Having agreed on what a conception of justice should look like, it is then necessary to assess whether its instantiation is *possible*; "for any conception of justice we might conjecture whether it is capable of being realized", and if so wonder, "what such a society would be like if it were governed by that conception" (Freeman, 2007, p.243).

Essentially, the problem of stability is a thoroughly practical one. When contextualised in relation to the claim 'ought implies can', answering to this problem is not necessarily designed to tell us *what we ought to do*, but whether what 'we ought to do' *can be done*.[6] To see this, contrast the issue of stability with Rawls's (1971, p.11) initial "aim" of providing a conception of justice which "generalizes and carries to a higher level of abstraction". Whereas the earlier stages of justice as fairness – in particular, the representational device of the original position – utilises abstraction by limiting the kinds and the amount of knowledge available, stability requires such knowledge to be reintroduced (just as being sure about what I 'can do' means possessing the relevant information). For example, suppose the serial ordering of the principles of justice (i.e. the priority Rawls gives to the liberty principle over the difference principle) is established through the original position. In that case, we can ask whether citizens will still support such an ordering when their material needs are not fully met and enquire as to what psychological mechanisms permit them to deal with this disparity (Rawls, 1971, p.543). Adjustments to the initial conception may or may not be necessary in light of the problem of stability. However, what is initially determined through the abstract must be proven compatible with the practical; *I only ought to do what I can do*.

According to Freeman's analysis – which follows and clarifies Rawls's position on this matter – there are two main issues that together comprise the problem of stability. First, *the assurance problem*, which requires that a conception of justice ensures that its citizens feel as if the laws of justice will be followed (Freeman, 2007, p.246; Rawls, 1971, p.270). Second*, the disruption problem*, which requires a conception of justice to demonstrate that whenever disruptions do come about, "general compliance with the norms of justice" will be reaffirmed (Freeman, 2007, p.246; Rawls, 1971, p.144). Broadly put, these conditions stipulate

[5] There is a diachronic element to stability. A conception of justice is stable if it can be implemented *and* sustained through time.

[6] I will explain this more in chapter 2. Broadly speaking, my point is that once we have the design for a conception of justice, it must be submitted to the tensions and dynamics of social life.

requirements for implementing a conception of justice within society; answering them constitutes the *desiderata* for an account of stability. Thus, a conception of justice is stable when persons feel that the norms of justice will be maintained, and once the basic structure of society can respond to disruptions to social life. Conversely, a conception is unstable if the norms of justice are not adhered to or if they degrade when disruptions occur. If citizens lacked such assurances under the given conception, then this would indicate that they suffer from a distrust of the social institutions they live under. Similarly, if a conception of justice could not respond to disruptions by reaffirming the rules of justice, then there would be sufficient reason to doubt the longevity of that conception, as well as the normative force of its rules. In both instances, the issue of stability is at stake.

To further this discussion, I shall take Freeman's categorisation for granted, in order to apply these problems to the technical definition of stability that Rawls provides in *Theory*. For Rawls (1971, p.457), instability is defined *via* deviation from a state of "equilibrium", meaning that "a movement away from it [equilibrium] arouses forces within the system that lead to even greater changes." In contrast, a system is "stable whenever departures from it, caused say by external disturbances, call into play forces within the system that tend to bring it back to this equilibrium state" (Rawls, 1971, p.457). As Rawls (1971, p.457) clarifies, the "relevant systems" concern "the basic structure of society" – a structure which includes "the political constitution; the legal system of trials, property, and contracts; the system of markets and the regulation of economic relations; and the family" (Freeman, 2007, p.464). Hence, to say that a well-ordered society is stable is to claim that the institutions comprising its basic structure continue to return to a just arrangement regardless of changes to the fabric of those institutions/institutional practices. In fact, Rawls (1971, p.458) expects that "society will (…) contain great diversity and adopt different arrangements from time to time", the key question in regard to stability concerns whether these divergences are "tolerable" – that is, whether they return to an acceptable state of equilibrium in relation to the prevailing conception of justice.

In sum, an unstable conception of justice will deviate from a state of equilibrium when it inadequately responds to the respective problems of disruption and assurance. For example, if a given conception failed to respond to the problem of disruption, then changes to social cooperation will not realign to the norms of justice. Instead, it will hit a fail-state where citizens and institutions move further away from the norms until their behaviour no longer recognisably fits with the justice principles. Yet, a society does not need to be in perfect equilibrium for it to count as stable. Disruptions will inevitably occur within society. However, in the short term, and under the right conditions, such disruptions are permissible: they may result from natural disasters, a readjustment of institutional practices, etc. It is only when such disruptions lead to significant deviance from the norms that the conception of justice is considered unstable. Inversely, for a conception of justice to be stable, it must invoke mechanisms that respond to the problems of assurance and disruption by re-establishing the equilibrium state; that is, such mechanisms – "forces" as Rawls (1971, pp.456-458) occasionally calls them – are necessary to secure a stable society. But what are these forces within justice as fairness? What ensures, on the broadest level, that the well-ordered society will be stable? Here, I refer to the 'broadest level' because my concern is with the *conditions of success* for Rawls's two arguments from stability, i.e. what mechanisms they invoke to ensure equilibrium-state within the well-ordered society. I shall answer these questions from a Rawlsian perspective throughout the following sections.

1.2. A Broad Convergence Between Rawls and Hobbes

One way of characterising the forces that ensure stability in the well-ordered society is by considering the application of the first principle of justice. I take this approach partly because the operative parts of a stable society are the conception's state of equilibrium and the forces that maintain it. By enquiring into the first principle, it is possible to work backwards from the conception of justice and explore what forces the equilibrium-state demands. Put differently, I am concerned with the practicality of the first principle – and for the purpose of simplicity, only the first principle – within the context of the well-ordered society. With this in mind, the first principle of justice and its accompanying ordering principle requires the following distribution:

> Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all…liberty can be restricted only for the sake of liberty (Rawls, 1971, p.302).[7]

To refine things further, I shall consider the question of stability as it pertains to *a closed system* (an individuated society), *a state of equilibrium* (societal norms and institutional practices organised according to the liberty principle), and *the two problems of instability* (the problems of disruption and assurance respectively). I will call this the 'liberty society.' Based on this

[7] This is the original rendering of Rawls's (1971, p.302; 1987, p.5) liberty principle. To note, this definition changes from requiring 'the most extensive system of liberties' in *Theory*, to a "fully adequate" scheme of liberties in *Justice as Fairness: A Restatement*. To remain consistent throughout this thesis (which is based on *Theory*), and because its content does not affect my argument overall, I have kept to the original version.

qualification, the liberty society is unstable when disruptions to social cooperation or a general mistrust held between citizens and institutions lead to growth in the abuse of rights and liberties. Here, I am concerned with the forces necessary to avoid this unstable outcome, or perhaps more accurately, the overarching mechanisms that the Rawlsian conception of justice relies on – wherein the liberty principle is broadly definitive – to ensure a dependable and continually returned to state of equilibrium. Let me begin by considering an intuitive solution.

At least *prima facie*, a broadly Hobbesian approach would solve the problem of stability within the liberty society. In Hobbes's (1996) work, instability is construed as a threat to sociality as well as individual citizens' psychological and physical well-being, such that the primary function of the sovereign is to protect and secure general compliance with the rules.[8] In fact, the principal concern of the liberty society seems to be a built-in expectation of a Hobbesian conception: the exercise of power and authority to secure rights and liberties, resulting in members of society being assured that other citizens will comply with the rules, fortifying the society against any potential disruptions. In other words, we have our closed system (the liberty society), our equilibrium-state (a comprehensive list of rights and liberties) and a potential response to the two problems of instability (disruption and assurance, respectively) through a coercive sovereign. In fact, Freeman cites Hobbes as providing an archetypal response to the problem of instability:

> Hobbes argued that nearly absolute political power was necessary to resolve these problems. To provide all with the assurance that others will comply with justice [the assurance problem], and to insure that society can withstand internal conflicts and disturbances [the disruption problem], the Sovereign required *de facto* powers unlimited by legal restrictions and regulations (Freeman, 2007, p.246)

According to this brief analysis of Hobbes, the force that ensures against the respective problems of disruption and assurance, producing a dependable return to equilibrium, *just is* an authoritative system of coercive measures. Citizens will support the Hobbesian conception of justice primarily because it is enforced, and they will be assured primarily because disruptions are continually and directly resolved.

On the face of it, the above is an acceptable account of stability within the liberty society. Instability as a practical problem is resolved by assuring citizens that their rights and

[8] A simplified interpretation of Hobbes's work is relied upon for this discussion, it is not intended to be a detailed presentation. My reading is both brief and generally charitable: I am concerned with Hobbes insofar as he relates directly to Rawls and insofar as he provides a *prima facie* plausible response to stability.

liberties will be protected by a background system of enforcement. Even if the operation of coercive power is not always direct, citizens will be comforted by the fact that – when and where it is necessary – the rules will be enforced. What is required for stability in the Hobbesian society is the ongoing sense that issues will be decisively resolved by the sovereign whenever appropriate. To provide an example: a sovereign could determine whether the claim to assisted suicide constituted a basic liberty or whether it conflicted with any other putative rights, such as the right to life. On this basic understanding, civil disputes that arose as a result of this 'final say' on euthanasia would then be resolved through a system of legal restrictions, regulations, and ultimately punitive measures. The upshot is that stability is founded upon the functional powers of a coercive sovereign, which possesses both the authority and the capacity to resolve disturbances to social order – in other words, it is the sovereign (typified by a framework of enforcement) that maintains an ongoing state of equilibrium within society.

The second thing to note is that Rawls has expressed direct approval for this Hobbesian approach. This is outside, I should add, the secondary evidence that supports this Hobbesian-Rawlsian convergence. For example, Ladenson (1980, pp.139-140) argues that rational persons situated behind a veil of ignorance would justify a coercive (Hobbesian) sovereign precisely on the basis of stability. As Ladenson (1980, p.140) puts it: "all rational people under the veil of ignorance would acknowledge the necessity of the state, and thus would not object to the exercise of coercive power". However, I will not rely on this point as I think Ladenson misconstrues the problem. To capture the broad strokes of my counterargument, there is a clear distinction between the justification of the sovereign within a hypothetical scenario and the instantiation of this sovereign as a response to the problem of stability. One should not use the former to establish the latter since stability is *essentially* a practical issue. That being said, this does not depreciate the fact that Ladenson rightly highlights the overlap between Rawls and Hobbes, nor does it nullify Rawls's explicit endorsement of Hobbes within the more appropriate context of the organisation of institutions:

> In the absence of the authoritative interpretation and enforcement of the rules…an arrangement is unstable. By enforcing a public system of penalties government removes the grounds for thinking that others are not complying with the rules… the existence of effective penal machinery serves as men's security to one another. This proposition and the reasoning behind it we may think of as Hobbes's thesis (Rawls, 1971, p.240).

Here, what Rawls refers to as "Hobbes's thesis" is expressed by the analysis I have provided so far, i.e., a system of coercive measures helps secure against the problems of disruption and assurance, leading to an ongoing state of equilibrium defined by the norms of justice. For example, a citizen could claim that their disorderly conduct was only temporary or that it was justified by some other party's actions. Since (in the absence of enforcement/a sovereign) there is no authoritative system to ratify such excuses as *legitimate reasons*, or indeed to punish illegitimate excuses, general mistrust would be fostered between citizens and institutions. In other words, not only would the absence of enforcement seem to permit disruptions to occur, but it would also lead to a lack of assurance in the conception of justice. Hence, a system of enforcement seems to stand as one of the primary forces that ensure stability within the well-ordered society; indeed, it is one of the only forces that Rawls mentions directly. Subsequently, even if Rawls (1971, p.222) does not wish to impart *absolute* powers to a regulatory system of punitive measures – as, for example, he endorses a "constitutional democracy", rather than a Hobbesian sovereign – it is clear from the above that enforcement is pivotal to *Theory*'s response to stability. In the next section, I aim to bring this position into question by highlighting some theoretical tensions within Rawls's work. Having done so, I will use this internal discord to point towards a more palatable and nuanced reading of *Theory*.

1.3 Distancing Rawls from Hobbes

To take stock of the discussion so far, I have defined stability through the Rawlsian notion of equilibrium, identified two main problems related to instability through Freeman, and demonstrated that Rawls (much like Hobbes) utilises a system of enforcement when responding to these problems. There is, however, a tension that arises out of this analysis. In particular, the Rawlsian-Hobbesian convergence contrasts starkly with Rawls's deliberate attempts to distance justice as fairness from Hobbes's legacy. As Rawls (1971, p.11, n.4) puts it, whilst Locke, Rousseau, and Kant all share in a tradition similar to his own, Hobbes's *Leviathan* is excluded from this connection because it "raises special problems." However, as Gricc (2007, p.371) rightly points out, Rawls never actually explains what these "special problems" are or why they warrant Hobbes's broad exclusion from his work. The resulting issue is this: if a broadly Hobbesian position is required to deal with the problem of stability, why does Rawls feel the need to distance himself from it? What precisely are the 'special problems' that he has in mind?

An explanation of this discordance does not lie in Rawls's (1971, pp.521-522; 587) discussion of Hobbes but in his treatment of the so-called "private society" that occurs later on in *Theory*. Admittedly, when Rawls identifies thinkers associated with the private society – in particular, Plato, Hegel, and Adam Smith – he does not mention Hobbes; however, this list is qualified within the same footnote: “the notion of a private society, or something like it, is found in many places” – and I will show that Rawls (1971, p.521, n.3) certainly takes this to be true of Hobbes. Furthermore, the notion of a private society is broadly defined by Rawls (1971, p.521) as exhibiting two main features. First, individual citizens or associations are taken to be discrete entities with their own private ends. Second, institutions are understood not to possess any value independent of these private ends. For Rawls, this form of society is undesirable, and there is evidence to suggest that it is undesirable *partly because it takes enforcement to be its primary stabilising force*:

> Since the members of this society are not moved by the desire to act justly, the stability of just and efficient arrangements when they exist normally requires the use of sanctions. Therefore the alignment of private and collective interests is the result of stabilizing institutional devices applied to persons who oppose one another as indifferent if not hostile powers…It is sometimes contended that the contract doctrine entails that private society is the ideal (…) But this is not so, as the notion of a well-ordered society shows (Rawls, 1971, p.522)

My point is that this characterisation of the ’private society’ posited by Rawls fits with the ‘liberty society’ I described in the previous section. To see this, recall that citizens in the liberty society possess an extensive set of rights and liberties. Given the simplified Hobbesian approach, such rights-claims are recognised *via* a system of regulation and enforcement. For Rawls, this approach means that citizens’ private and collective interests only converge on the successful execution of punitive measures. The upshot is that persons are conceived as fundamentally separate from one another, with the sole mechanism for stability *scaffolded around* their private interests. In accepting this analysis, I suggest that one of the ‘special problems’ that Rawls has with Hobbes concerns precisely the issue of stabilising forces. In fact, this interpretation is corroborated further by Rawls in *Lectures on the History of Moral Philosophy*:

> Hobbes’s social contract establishing the sovereign does not involve a shared end…the state’s institutions are a common end only in the sense that they are a means to each individual’s separate happiness or security. Those institutions do not specify a form of public political life that is to be seen by citizens as right or just in itself … The society of *Leviathan* is a kind of private society (Rawls, 2000, p.365).

It is worth noting here that the two indented quotes of this section both use the idea of a ‘private society’ – a term used by Rawls (2000, p.365; 1973, p.521; 522; 587) in a limited sense, around five times in *Theory* and only once in the *Lectures*. In light of this, it seems appropriate to align the Hobbesian approach to stability with Rawls's apprehensions regarding the private society in *Theory*: that a society stabilised by a system of regulations and punitive measures treats individuals as atomised entities, concerned only with their own ends and the ends of their immediate associations. Hence, there must be additional nuance to the first-gloss response to the problem of stability – i.e. a system of coercive measures – and to Rawls's apparent endorsement of Hobbes's thesis. If Rawls's mistrust of private society partly concerns the stabilising mechanisms it employs, then the broad account of stability required for the well-ordered society cannot solely depend on a Hobbesian-like system of enforcement; there must be an alternative approach.

1.4. Appealing to and Impacting the Attitudes of Persons

To resolve the tension of the previous section, I will extract an alternative understanding of Rawls's approach from McClennen's (1989, pp.3-30) investigation into the problem of stability. In his article, McClennen (1989, p.8) asks us to imagine a liminal stage wherein citizens live in a society that is not yet well-ordered – i.e. its institutions do not conform to a particular conception of justice and its citizens have not yet come to any agreement on this matter. As a corollary, citizens within this society "have yet to develop a full-blown sense of justice" (McClennen, 1989, p.8).[9] The challenge for Rawls is to explain the transition from this liminal state to the well-ordered society. McClennen (1989, p.8, emphasis added) posits that Rawls does *not* recommend "that they [the liminal citizens] put into place some institution that will *of itself* stabilize their relations with one another." Indeed, this coheres with my analysis so far: a system of enforcement scaffolded around private citizens, which 'of itself' stabilises society, is precisely the state of equilibrium that Rawls wishes to avoid. Instead, as McClennen (1989, p.8, emphasis added) puts it, "the invitation" to such citizens from Rawls "is to adopt an arrangement *whose stabilizing effect comes through its impact on their attitudes*." In this way, citizens within the liminal society are encouraged to "take steps that will result in a reconstruction of themselves as purposive beings – a transformation of their affective ties to one another and hence of their final ends" (McClennen, 1989, pp.8; 15-19).

This alternative account of stability is one founded on *reciprocity*. It is an approach that appeals to the attitudes of citizens in relation to other members of society, the institutions comprising its basic structure, and the principles of justice. The upshot is that justice as fairness

[9] Though I will discuss this in chapter 2, a 'sense of justice' is (broadly) the disposition to do as justice requires. In this case, citizens in the liminal society have not developed the sense of justice correlative to Rawls's conception.

reverses the two main features of private society. Instead of consisting in private ends, citizens are encouraged to adapt their ends to incorporate the notion of the well-ordered society as a collective endeavour. For McClennen (1989, p.6), this is reflected through rational choice theory. As he puts it, disruptions to the agreement required for a society to be well-ordered consist of "defection" and "renegotiation", such that a society is stable when individuals approach this problem "holistically rather than incrementally" with respect to their own ends (McClennen, 1989. p.12; 19). In this analogy, promoting the likelihood, strength, and value of this agreement – which here is achieved through the 'opening up' of private ends or a holistic approach to rational choice – is tantamount to establishing the stability of a given conception of justice. More specifically, as an agreement, citizens do not want to impose stability upon themselves. Instead, a conception of justice is stable when it coheres with the status of citizens, to borrow from McClennen, *qua* purposive beings.

The upshot of this reading, in regard to the well-ordered society, is that Rawls (1971, p.454, 474). is not concerned with quashing de-stabilising forces *per se*, but with establishing that "those taking part in these arrangements desire to do their part in maintaining them" – or even more generally, that citizens possess "a willingness to work for (or at least not to oppose) the setting up of just institutions." As such, disruptions are ironed out because citizens are motivated to maintain just institutions even when their private benefit is not immediately evident (resolving the problem of disruption). Similarly, by recognising the value of institutions and justice as a collective endeavour, citizens will feel assured that the norms of justice are sufficiently compelling and generally followed (resolving the problem of assurance). In sum, my position is that Rawls's account of stability depends on what I will call *the two generative properties of justice as fairness*:

1. Justice as fairness encourages citizens to assess their own situation and final ends from within the constraints set by justice so that they can realise their status as purposive beings.
2. Justice as fairness encourages citizens to maintain and appreciate just institutional arrangements independent of their private ends, so as to realise the value of reciprocity and collective agreement.

My intention with these properties is to express the stabilising force of justice as fairness – 'reciprocal attitudes', as it might be called – in its broadest terms. Rawls's account of stability is successful insofar as it secures these two generative properties. If these attitudes prove sufficiently appealing, then Rawls can go on to secure an ongoing state of equilibrium within the well-ordered society. Moreover, it should be noted that the interpretation I provided above has the additional benefit of prefiguring Rawls's two arguments from justice. In particular, it clarifies that the challenge for Rawls is to demonstrate that the justice principles can be incorporated into one's system of ends, that they transform such ends, that they recognise the conditions of sociality, and that they realise one's status as a purposive being. Though I will develop this account in more detail in the next section, I have offered a brief sketch of the two generative properties of justice as fairness – i.e. a candidate for the stabilising force of the well-ordered society. From this, I will now resolve the tension between this account of stability (i.e. the two generative properties) and Rawls's endorsement of Hobbes. In doing so, I complete the background to Rawls's arguments from stability.

1.5. Completing the Background to Rawls's Account of Stability

At this point, I have provided two accounts of the well-ordered society's stabilising forces: enforcement and reciprocity – the worry now is that it comes out indeterminate as to which constitutes the primary stabilising force within the Rawlsian framework. To see this, imagine how this current account of stability (which utilises both enforcement and reciprocity) would be received by citizens within the liminal society. Without being clear as to how these stabilising forces relate to one another, the citizens will be at a loss. Naturally, they would ask: 'how can a conception of justice both 'stabilise itself' through a system of enforcement and have its stabilising effect come through its impact on our attitudes? And if it can take such an approach, how do we know that the latter claim is not made in bad faith?' Replying that 'we need both approaches' fails to say anything qualitatively significant, e.g. it does not clarify how to approach the notion of reciprocal support, whether it should be given priority when it comes to stability, and so on. Essentially, the apprehensions of the citizens within the liminal society are justified because the Rawlsian conception of justice comes out indeterminate in relation to two markedly different accounts of stabilising forces.

To start by summarising my position, I propose the following answer: enforcement is a necessary but insufficient condition for stability within the well-ordered society because it is a formal requirement for realising the liberty principle. Furthermore, when it comes to the problem of stability, priority and independence must be given to the two generative properties of justice as fairness; that is, to generating the appropriate attitudes towards the conception itself. The upshot of this is that an account of justice is stable in the Rawlsian sense when it is a practical conception that fits the broader points we can make about society and persons.[10] This fittingness then translates to the maintenance of the conception (i.e. equilibrium) and specifies some of the main connections between the principles, its institutions, and members of society. My contribution is vindicating this account through the two generative properties of justice as fairness, thereby outlining the background to Rawls's response to the problem of stability and clarifying the role of enforcement. As such, my account should appeal to Rawls and Rawlsians.

To show that enforcement is necessary for the stability of the well-ordered society, primarily because it gives content to the principles of justice, I will again use the liberty principle as an example, and in particular, the putative legal rights that it confers onto citizens. I will say, as Hohfeld (1917, p.717) does, that rights are claims which entail duties. When I exercise a right to legal aid, I am enacting a claim against my government that demands the provision of a qualified attorney. The government, or whatever relevant party might be in play, then has an obligation to fulfil this claim. Importantly, this obligation must be enforced and executed through a regulatory system of coercive powers. To recognise this necessity, recall that in the absence of such regulations, the proliferation of excuses would rapidly increase; there needs to be a system that ratifies excuses as permissible reasons. Conversely, without this form of enforcement, the liberty principle is rendered unintelligible and ineffective. If there are no systems of regulation and enforcement, then whether I have a claim to a qualified attorney would be altogether unclear (e.g. how are we to know what 'qualified' means, in what legal cases does this right obtain, and so on). This connection between liberty and enforcement is one that Rawls highlights directly:

> Now the connection of the rule of law with liberty is clear enough. Liberty, as I have said, is a complex of rights and duties defined by institutions…But if the precept of no crime without a law is violated … what we are at liberty to do is likewise vague and imprecise. The boundaries of our liberty are uncertain … To be confident in the possession and exercise of these freedoms, the citizens of a well-ordered society will normally want the rule of law maintained (Rawls, 1971, pp.239-240)

As a brief aside, which will serve to expand on Rawls's point here, Hermann criticises this approach *precisely because* of its minimalistic impact on the issue of stability:

> The argument seems to be that procedural integrity produces legal legitimacy which results in social stability. It is quite doubtful that

[10] Chapter 2 expands on this point in light of Rawls's congruence argument.

> procedural regularity in itself contributes to anything more than the maintenance of the status quo. (Hermann, 1974, p.1417, n.37).

Here, I accept that enforcement gives content to the liberty principle partly through regulative consistency, such as treating similar legal cases similarly (Hermann,1974, p.1416). Additionally, Hermann is correct – though, I think, trivially so – that procedural integrity primarily serves to maintain the *status quo*. What is far less clear is why, on the account that I am advancing, this is a problem for Rawls. After all, the regular application of legal procedures is what citizens will come to expect in regard to their basic rights and liberties. Understood in this way, if we take the role of enforcement to be one of providing content to the principles of justice, then the minimalist effects on stability that Hermann highlights are both expected and easily explainable. Hermann is right to point out that enforcement does little for stability, but only because the question of whether justice as fairness is stable *extends well beyond a system of coercive measures*. This does not mean that enforcement is not a necessary assurance since, without it, the principles of justice would risk being rendered contentless.

There is a more positive way of putting this point: the citizens of the liminal society will prefer a conception of justice that is stable because of its transformative effect on their private ends, because it will remodel their relationship to social institutions, and because it will allow them to realise their status as purposive beings. Again, the notion of enforcement is necessary to this overall picture – for example, because it buttresses the terms of fair cooperation – but this will not be suitably appealing to citizens who do not currently possess a full-blown sense of justice. As Rawls (1971, p.455) puts it, "however attractive a conception of justice might be on other grounds, it is seriously defective if…it fails to engender in human beings the requisite desire to act upon it." In this way, the Rawlsian approach to stability must be attitudinal in nature: it must engender a meaningful change in the beliefs, dispositions, and behaviour of persons. Hence, the indeterminacy between different stabilising forces is resolved once priority and independence are given to the two generative properties of justice as fairness. It is the 'liminal citizens' (or the citizens of the well-ordered society) themselves that are responsible for the stability of justice as fairness – but that is Rawls's appeal, that in this responsibility, they will find something of unique value.

In this way, legal norms will fluctuate, and systems of enforcement will readjust, but what ensures the stability of justice as fairness is that *its citizens realise parts of their nature that would be otherwise inaccessible to them* (say, in a private society). This is the higher-order incentive that Rawls offers for his conception of justice. As he puts it in *Theory*, "properly understood, the desire to act justly derives in part from the desire to express most fully what we are or can be, namely free and equal rational beings with a liberty to choose" (Rawls, 1971, p.256). In this way, the stability-conducive attitudes of justice as fairness are generated precisely because of this appeal, and the conception as a whole – not just the formal requirement of liberties – is secured in turn. Again, as Rawls writes when discussing the Kantian interpretation of justice as fairness:

> …if a person realizes his true self by expressing it in his actions, and if he desires above all else to realize this self, then he will choose to act from principles that manifest his nature as a free and equal rational being (Rawls, 1971, p.255).[11]

Hence, though enforcement is necessary to give content to the distributive principles, justice as fairness secures the stability of the well-ordered society by appealing to the attitudes of persons and the ultimately transformative effects of justice (*via* the two generative properties).

I will close this chapter by specifying the way in which the two generative properties of justice as fairness prefigure Rawls's arguments from stability (discussed in chapter 2). Broadly, the two properties generate a change in the dispositions of persons so that they view their institutions as independently valuable and assess their own situation *via* the lens of justice. Concomitantly, the advantage of these attitudinal changes is that persons realise their status as purposive beings and open their private ends to the shared ends of the community. Subsequently, Rawls's aim with these more specific arguments is that, as far as possible, persons become disposed to support the conception of justice through mutual reciprocity. More specific questions naturally arise on this understanding: What does it mean to be a free and equal person, and how does justice as fairness substantiate and then realise that account of the self? How do we know that justice will enable a widespread change in attitudes? What are the aspects of justice as fairness that ensure reciprocity? To this end, Weithman (2024, pp.1-22) offers three "stimulus conditions" that promote the two generative properties of justice as fairness: institutionalisation (the general effects of living under just institutions), psychological principles (the effects of moral education), and "the tendency to return good for good" (the fact that justice as fairness cares for citizens' own pursuit of a good life). The remainder of this thesis focuses on this latter point, Rawls's congruence argument. Given this focus, in

[11] For the most part, the Kantian features of justice as fairness are set aside in this thesis. Instead, I focus exclusively on the notion of life-planning – which Rawls uses to characterise persons within his argument from congruence, and which finds favour in non-Rawlsian, non-Kantian, circles also (see, chapter 3). It is also worth noting, as I address directly in chapter 8, that Rawls comes to abandon the Kantian aspects of his earlier work.

conjunction with my analysis here, Rawls's account of stability aims to accurately describe and appeal to the capacity for formulating a conception of the good to show that justice as fairness satisfies the condition of reciprocal support. Hence, the stability of the well-ordered society rests, perhaps above all, on what Rawls says about the nature and dispositions of persons.

Chapter 2

Justifications, Institutions, and Assurances: Defending the Rawlsian Argument from Congruence

2. Introduction

Having laid out Rawls's broad account of stability *via* the two generative properties of justice as fairness, I now want to explain the more specific dimensions of Rawls's argument, i.e. *how* justice as fairness encourages citizens to develop attitudes of support towards its institutional arrangement. Throughout the rest of this thesis, my primary focus is on the congruence argument – which substantiates the second of Rawls's two arguments from stability (the other being the argument from moral development, which I shall briefly discuss in section 2.1). In the indented paragraph below, I summarise the congruence argument in broad terms. Building from this summary, my aim in this chapter is to investigate the details of Rawls's account of congruence before critically engaging with it as this thesis develops:

> We each have an interest in our life going well; supposing that a conception of justice as fairness has been agreed on, can it also be included within our version of the good life? If so, then justice and goodness are congruent, and persons have reason to affirm justice as a regulative desire. Subsequently, attitudes supportive of justice will be generated by the Rawlsian approach.

This chapter proceeds as follows. In section 2.1, I highlight the main assumptions underpinning Rawls's congruence argument. Namely, that the choice of the principles of justice has already been explained and supported, that its institutional arrangement has already been established, and that persons already have an effective sense of justice. In section 2.2, I then respond to Barry's (1995) argument – namely, that congruence is unnecessary since persons should be motivated to uphold principles of justice by their willingness to do what is right – from a Rawlsian standpoint. Specifically, I argue that Barry misrepresents the congruence argument, which is meant to assure citizens that they can affirm justice throughout all parts of their lives, even when its relation to moral conduct is not immediately evident. Finally, in section 2.3, I examine Rawls's responses to the problem of congruence (i.e. whether justice and goodness are compatible). Specifically, I detail how *Theory* responds to five main worries; that justice, (1), is founded on false beliefs, (2), requires subjugation to authority, (3), is rooted in envy, (4), undermines the powers of the self, and (5), disregards the value of the community. Ultimately, I provide an overview of the congruence argument, including its underlying assumptions, and defend it from Barry's claim that it is unnecessary for the stability of the well-ordered society.[12]

2.1. The Assumptions of the Congruence Argument

Rawls's argument from congruence – which shows that, in the well-ordered society, the good of justice is compatible with citizens' respective versions of the good – does not start *de novo* but instead takes several key assumptions for granted. To see why this is the case, recognise that in order to assess the stability of a conception of justice, one must assume that the account has been sufficiently worked out and that the institutions necessary to realise it have come to fruition within society. Rawls's intention is to consider the implications of an accepted theory of justice in order to analyse how it will bear out in society and whether it will be able to sustain a state of equilibrium throughout its instantiation. The upshot of this is that Rawls builds upon the earlier parts of his conception, meaning that the main conclusions of Parts I, II, and the start of Part III of *Theory* are presupposed by the argument from congruence, which appears mainly at the end of Rawls's monograph. In this section, I will briefly explain these assumptions to demonstrate how they build into the congruence argument – starting with Part I of *Theory*.

2.1.1. The Justificatory Stage of Justice as Fairness

In the justificatory stage of *Theory*, Rawls (1971, p.15) follows a broadly contractarian approach to layout and justify his conception of justice. To achieve this, Rawls models our firmly-held moral convictions through a hypothetical contract agreement called 'the original position'. In this situation, persons are placed behind a "veil of ignorance" that strips them of knowledge of anything other than the general information necessary to come to an agreement on principles of justice (Rawls, 1971, p.12). Moreover, such persons are hypothetical representatives of actual persons who possess – what Rawls (1971, p.50) calls – the two "moral powers": the sense of justice and the capacity to formulate a conception of the good. In such a scenario, Rawls (1971, p.302) argues that reasonable and rational citizens thus defined (*via* the two moral powers) will affirm the two principles of justice. The first of these principles is the 'liberty principle', discussed in chapter 1, which gives citizens equal claim to a fully extensive

[12] Contextualised in relation to the previous chapter, my aim is also to highlight some Rawls's main thoughts on the nature and disposition of persons, i.e. that they are capable of formulating and pursuing a conception of the good, that they have a foundational interest in this good, and that they will want to be assured that their sense of justice is self-authenticated (e.g. against the list of worries detailed above).

set of rights and liberties, and which is given priority over the second principle. The second is the difference principle, which requires "social and economic inequalities… be arranged…to the greatest benefit of the least advantaged… and…attached to offices and positions open to all" (Rawls, 1971, p.302).

Whilst this provides only a brief characterisation of Part I of *Theory*, it is necessary to recognise that Rawls resolves the problem of justification by arguing that his two principles of justice will be chosen by hypothetical persons in an appropriately constructed contract scenario. As a result, when considering whether his conception of justice is stable, Rawls presupposes that his principles have been fully explained and justified. To see this, recall from chapter 1 that stability is a practical problem, which necessitates the introduction of substantive information incompatible with Rawls's use of the veil of ignorance; as Rawls (1971, p.398) puts it, "when we come to the explanation of… the stability of a conception of justice, a wider interpretation of the good is required." This "wider interpretation" includes the fact that "we have a tendency to love those who manifestly love us", the fact of shared "social values", the "good of activities", and our pursuit of independently defined "final ends" – though there is doubtless more that could be added to this list (Rawls, 1971, p.398). Hence, through the original position, Rawls looks to establish a "free-standing" conception of justice *before* considering whether such a conception will be stable; as Mendus (1999, p.74) describes the problem of stability, "the aim is to show why, having been persuaded by the justification, the agent would be motivated to act."

2.1.2. The Institutional Stage of Justice as Fairness

In addition to justification, the question of which institutions/institutional arrangements satisfy the principles of justice must already be (broadly speaking) settled prior to the congruence argument. Whilst the specifics of this institutional arrangement will change over time, it is nevertheless necessary to have a template for what such an arrangement would look like, so that Rawls can describe in more detail – among other things – his conception's state of equilibrium. To this point, Rawls's congruence argument also assumes Part II of *Theory*. Here, Rawls (1971, p.195) illustrates "the content" of the two principles of justice by enumerating the institutions necessary for their realisation, as well as the "duties and obligations" that these principles and institutions give rise to. To this end, Rawls (1971, p.274) suggests that a "property-owning democracy" – organised according to the principles of justice – would provide "a social minimum" that allows all citizens to "exercise their basic rights effectively and realize fair opportunities, thereby achieving individual independence" (Freeman, 2002, p.279).

Regarding the connection between Part II of *Theory* and the problem of stability, there is less to be said overall. Having already supported his conception of justice, Part II essentially serves as a more or less concrete example of the institutions and political processes necessary for the well-ordered society. That being said, there are some relevant claims here for stability; for example, the fact that Rawls (1971, p.363) views civil disobedience – narrowly defined as "a public, nonviolent, conscientious yet political act contrary to law" – as a "stabilising device… [within] the special case of a nearly just society." Similarly, it is in Part II of *Theory* that Rawls (1971, p.337) states, "given the value of a public and effective sense of justice, it is important that… the duties of individuals be simple and clear… [this ensures] the stability of just arrangements." From this, we can assume that effectively laying out the duties of persons within the well-ordered society is required for stability, as well as correctly characterising the institutions needed to realise justice as fairness in the first place. However, throughout *Theory*, Rawls (1971, p.246) is principally concerned with justice in the realm of "ideal theory", meaning that he assumes general compliance with the requirement of justice – with the only expectation being narrow cases of civil disobedience. In this way, Rawls is not concerned with laying out in detail the precise institutions that justice as fairness requires, but with providing an "an Archimedean point for assessing the social system without invoking *a priori* considerations" (Rawls, 1971, p.261).

2.1.3. The Argument from Moral Development

This brings me to the final major presupposition that underpins Rawls's argument from congruence. For the most part, this assumption is laid out within chapter VIII, Part III, of *Theory*, meaning that Rawls is addressing the problem of stability directly (which compromises the main topic of Part III). Put in broad terms, the assumption is this: *persons within the well-ordered society have, and expect others to have, an effective sense of justice*; that is, very broadly, the desire to act in accordance with the requirements of justice. To this end, Rawls describes how persons within the well-ordered society develop their sense of justice. In particular, he argues for a pedagogical approach to moral development, wherein persons come to hold increasingly complex and firm connections to the principles of justice. Given its importance to the problem of stability, I shall reconstruct this argument below before providing some general comments:

1. Supposing that a conception of justice is understood as generally correct and justified, we can trace how its schema leads to individuals developing a sense of justice – that is, the "moral motivation to do what rules of justice require" (Freeman, 2007, p.249; Rawls, 1971, p.477; 505).

> *1a*. Justice as fairness promotes the development of a sense of justice primarily through an educative process. In particular, the three-stage moral development is divided between, and in order of, "morality of authority", "morality of association", and "morality of principles" (Rawls, 1971, pp.462-479).

2. If we can establish (*1*) through (*1a*), then when the development of this moral capacity is "sufficiently strong" and entrenched, it should:

> *2a*. Act as "regulative desire" and countermand "temptations to violate the rules."
>
> 2b. Be publicly known and stand as "a condition of human sociability" (Rawls, 1971, p.495; 497; 561).

3. As a result of (*2a*) and (*2b*), individuals will propagate a return to a state of equilibrium (i.e. the just arrangement of institutions) whenever disruptions occur within society.

4. As a result of (*2a*) and (*2b*), individuals are assured that others will comply with the rules of justice since they can be seen as expressive of a common nature, as well as a shared goal (e.g. in the social union of social unions).

C. Therefore, justice as fairness will be stable.

Here, the moral-educative process within the well-ordered society progresses from the family *via* the morality of authority, to specific communities and friendship *via* the morality of association, to the relationship between persons and institutions more broadly *via* the morality of principles. As a result, each stage of moral development prefigures the next, up to the point wherein "the conception of acting justly, and of advancing just institutions…[has] an attraction analogous to that possessed before by subordinate ideals" (Rawls, 1971, p.473; 514). In many ways, the educative process is *additive*: where possible, each stage contributes and reinforces the subsequent stage. As a result, Rawls claims that the sense of justice is a *supremely regulative commitment* – that is, the "highest-order…desire", which shapes and regulates the other desires that one has (Rawls, 1973, p.565; Freeman, 2003, p.294). Therefore, there is a greater chance of stability within the well-ordered society – according to the two generative properties of justice as fairness – since persons will have a sufficiently strong sense of justice that motivates them to uphold and act in accordance with just institutions and obligations.

The first part of Rawls's argument of stability – which I will call *the argument from moral development* – can be understood primarily as a work in moral psychology. That is, Rawls (1971, p.p.494) looks to affirm the "*deep psychological fact*" that we develop attachments to institutions correlative to how they affect us."[13] In light of this, the psychological aspects of the sense of justice can be formulated in this way: it is both something that we *already have* – since without it, no form of sociality would be possible – and something that we are *capable of strengthening* – such that appeal of justice as fairness is that it promotes this desire through its principles, through an educative process, and through the organisation of social interaction.[14] Hence, the final assumption underpinning the congruence argument is that we all have some notion of 'doing what is right' (i.e. a sense of justice) and that – through the moral education of the well-ordered society – this notion of 'doing what is right' will come to bear on the principles of justice.

In summary, three major assumptions underpin Rawls's argument from congruence. First, the principles of justice have already been explained and supported through the original position. Second, the institutions and obligations necessary to realise justice as fairness have already been arranged. Third, persons have an effective sense of justice, established through a pedagogical system of moral development. In section 2.2, I will highlight the general appeal of Rawls's congruence argument before unpacking it in more detail. Specifically, I will argue that congruence is, (i), a requirement for the sense of justice to be a supremely regulative desire, and (ii), a test of whether justice is harmonious with certain general facts about human nature. To do so, I will first consider an argument against congruence; namely, that the motivation to do 'what is right' should be sufficient to maintain the stability of the well-ordered society.

2.2. The Test of Congruence and the Problem of Moral Motivation

Broadly speaking, Rawls's argument from congruence holds that – within the well-ordered society – a person's sense of justice is congruent with their conception of the good, understood

[13] For an account of evolutionary psychology and evaluative judgements of this kind, see Street's (2005, pp.109-166) paper 'A Darwinian Dilemma for Realist Theories of Value'.

[14] Some scholars, such as Hill (2014, pp.200-216), have captured this idea by distinguishing between the sense of justice and the capacity for the sense of justice. It is presumed, in this way, that we all have the *capacity* for a sense of justice without having it take the particular form that Rawls outlines.

as the values and ends that shape their version of a life well-lived. In other words, the remainder of part III of *Theory* claims that acting as justice requires – e.g. affirming and participating in just institutions in the well-ordered society – can be a meaningful part of one's "plan of life" or at least harmonious with it (Rawls, 1971, p.566). [15] To understand this argument, it is necessary to come to grips with why Rawls included it in the first place. After all, it may seem initially odd; the congruence argument is primarily worked out within the realm of ideal theory and assumes that persons already have – and expect others to have – an effective sense of justice (Rawls, 1971, p.567). This encourages a plausible line of enquiry: if persons already possess an effective sense of justice, *then why is this not sufficient motivation for citizens to support just institutions*? Why include the congruence argument at all? To this point, Barry argues that a sense of justice is *all that is required* for citizens to maintain the stability of the well-ordered society:

> It is…quite natural to say that the thought that something is the right thing to do is what motivates us to act rightly. I take this to be precisely the proposition that Rawls objects to… [That is, by providing the congruence argument], Rawls commits himself…to the ancient doctrine that no act can be regarded as rational unless it is for the good of the agent to perform it (Barry, 1995, p.884).

There are two main points here. First, Barry's position is that agents can be sufficiently motivated to act on their sense of justice *simply because* it is the right thing to do. [16] Second, according to Barry, Rawls departs from this first point with the congruence argument by maintaining that justice must be compatible with citizens' conceptions of the good. In general, the argument is that Rawls commits a crucial error by moving beyond his account of moral development. By doing so, he overestimates the role that ostensibly non-moral considerations (e.g. our respective versions of the good) have in moral motivation. For Barry, this move is either superfluous or harmful since – in his view – a sense of what is right is both necessary and sufficient for persons to act justly.

Here, it should be noted that Rawls (1971, p.477; 512, 569) *does* reject the idea of a "purely conscientious act", i.e. "the desire to do what is right… simply because it is right". *However*, on close attention to the passages where Rawls (1971, p.477; 477, n.15; 569)

[15] Throughout this section, I will generally follow Freeman's (2003, pp.277-315) interpretation of Rawls and build on his response to Barry. Moreover, it should be noted that the notion of a 'rational plan of life' is pivotal to Rawls's argument from congruence, as he uses it to define a person's conception of the good (to this end, recall the onus on Rawls to characterise persons from chapter 1 accurately). As a result, this idea gets its own treatment in the next chapter and much of this thesis is dedicated to assessing its plausibility.

[16] To note, Barry supports this first point through Scanlon's (1982) work on rationality.

advances his rejection, it is clear that he distances himself from a decidedly *narrow* viewpoint. More specifically, Rawls (1971, p.477) is concerned with "the sense of right…[as] a desire for a distinct (and unanalyzable) object" – a view he attributes to Ross (1930, pp.157-160; 1939) and rational intuitionism more broadly. With this in mind, Barry's claim must be that the congruence argument generalises the implications of this critique by providing a broadly instrumentalist view of moral motivation, i.e. one acts rightly to realise one's own interests and desires. Put differently, for Barry, Rawls extends his critique of a purely conscientious act towards a rejection of a conscientiousness *tout court*, hence the requirement of congruence for a conception of justice.

On this claim of generalisation, Barry's critique misrepresents the congruence argument. Against Barry, it is clear that Rawls does have a notion of 'doing what is right because it is right'. For example, Rawls (1971, pp.573-574) describes the sense of justice as "a striving that contains within itself its own priority", whilst writing that, "for the sake of justice a man may lose his life where another would live to a later day." To be clear, then, Rawls's rejection of rational intuitionism does not lead to a rejection of conscientious *tout court*; as Freeman (2003, p.282) puts it, "Rawls does not deny that people can act for the sake of duty – quite the contrary… we can and do act simply for the sake of justice." Instead, Rawls criticises rational intuitionism on a rather specific charge: *it fails to be sufficiently informative*. In particular, Rawls (1971, p.478) claims that if acting rightly is to be explained by the grasping of unanalysable non-natural property, then there is nothing to add as to *why* we should do what is right; "on this interpretation the sense of right lacks any apparent reason, it resembles a preference for tea rather than coffee." In this way, the Rawlsian account of moral motivation is principally concerned with the reasons for doing what is right. Subsequently, the congruence between justice and goodness must be understood as *one of those reasons*, allowing for the fact that some persons may be moved by their sense of duty alone.

This seems to be the crucial point of the congruence argument: Rawls (1971, p.569) is not concerned with the idea that persons should *de facto* do the right thing but that *they have reason to do the right thing when other options are available to them*. Indeed, this is part of what it means to possess a sense of justice that is 'supremely regulative'; that is, a desire that overrides and organises other desires. As Rawls writes:

> Since …[the] principles are regulative, the desire to act upon them is satisfied *only to the extent* that it is likewise regulative with respect to other desires… Other aims can be achieved by a plan that allows a place for each,

> since their satisfaction is possible independent of their place in the ordering. But this is not the case with the sense of right and justice… (Rawls, 1971, p.574, emphasis added).

With this in mind, the sense of justice is a desire that regulates the structure and the particulars of one's life plan according to the demands of justice. Given this, acting in accordance with justice cannot be explained *via* unanalysable moral properties, for it has to be an overriding motivation that always gives reasons for one to act justly. One must be assured that one can "abide by reasons of justice in all of one's actions" (Freeman, 2003, pp.290-291; 299). To this end, the strongest possible assurance is that justice is congruent with one's conception of the good. *Contra* Barry, then, the congruence argument does not look to show that persons do what is right because it is good, only to show that "being a good person (and in particular having an effective sense of justice) is indeed a good for that person" (Rawls, 1971, p.577). It is motivation by means of assurance, that the sense of justice is "compatible with human nature or our good" (Freeman, 2003, p.283). In Rawlsian terms, the broad thrust of the congruence argument is this*: it is rational to be just.*

There is another way of putting this response to Barry. Namely, the congruence argument functions as a stress-test for justice as fairness. In this way, even supposing that the majority of persons are moved by their sense of justice alone, the congruence argument remains a relevant factor in determining the stability of the well-ordered society. That is, persons may act rightly by affirming their sense of justice, but this does not mean that this disposition is beneficial or even possible (Freeman, 2003, pp.286-287). As I put it in chapter 1, it remains to be seen if what ought to be done, can be done. Subsequently, Rawls must consider whether justice fits within the complex of a human life, including all of the social, interpersonal, and normative dynamics that follow suit. If justice can be affirmed as part of persons' respective conceptions of the good, then this verifies the practical possibility of justice as fairness.

Given that the congruence argument is a form of verification – that persons within the well-ordered society can regulate all their other desires according to justice – I shall end this section by emphasising *why this form of assurance is necessary*. Here, Freeman offers a plausible answer to this problem:

> It may be that we want to do our duty of justice for its own sake; still, if these moral sentiments are grounded in illusions, defeat our primary purposes, prevent us from realizing important human goods, or require ways of acting that are not in our nature consistently to perform, then surely this

> is relevant to the justification of a conception of justice (Freeman, 2003, p.287).

More specifically, Freeman (2003, pp.286-287) points out five main concerns of the congruence argument, which I will summarise here – whilst pointing out where they are discussed or implied within *Theory* – before analysing them further in the next section:

> *Contingency and Ideology*: The sense of justice may be the result of convention, or worse, the result of prevailing power structures that serve to affirm false values to misrepresent the conditions of society (Rawls, 1971, p.514).
>
> *Authority*: The sense of justice may arise out of deference to authority and the abdication of personal responsibility. For example, my disposition to act justly could be founded on my fear of punishment, or it could be that – as the argument from moral development seems to suggest – it is merely a learned product from my early childhood (Rawls, 1971, p.514).
>
> *Envy*: In society, there may be an unequal distribution of resources. As a result, some citizens may envy others owing to their advantages and therefore call for a strictly equal distributive arrangement. So long as it is plausible that such cases can arise due to tendencies in human behaviour, then it is possible that the sense of justice (in this example, represented by the call for equality) similarly has its basis in undesirable characteristics such as envy (Rawls, 1971, p.539).
>
> *Self-abnegation*: From a broadly Nietzschean perspective, the aim towards justice requires sacrificing the higher powers of the self. In this way, the sense of justice may be imposed on persons to stem their capacity for self-realisation, rather than promote it (Rawls, 1971, p.499)
>
> *Community*: Communal life is crucial to human flourishing. If the sense of justice is a purely individual activity or disposition, then this may be incompatible with the community as a whole. Hence, persons may be unable to affirm justice as a shared end or even realise the value of acting justly (Rawls, 1971, p.520).

In summary, the congruence argument aims to demonstrate that persons can affirm the sense of justice as supremely regulative within their conception of the good. Concomitantly, this acts as a form of assurance to persons within the well-ordered society – that their sense of justice is not a result of, or connected to, the five concerns laid out above – thereby consolidating its stability: “the greater the lack of congruence, the greater the likelihood, other things equal, of instability with its attendant evils" (Rawls, 1971, p.576). In the next section, I will briefly investigate Rawls's specific answer to each of these concerns.

2.3. Congruence within the Well-Ordered Society

So far in this chapter, I have established the assumptions underpinning the congruence argument and argued that congruence functions as a form of assurance that justice can be affirmed as a supremely regulative desire, even given certain tendencies and dispositions generally associated with human nature. For simplicity, I have followed Freeman in summarising the main concerns correlative to the problem of congruence: contingency, authority, envy, self-abnegation, and community. With these points in mind, I shall briefly explain Rawls's answer to these problems below, before summarising the congruence argument in general terms.

2.3.1. Contingency and Ideology

As Rawls (1971, p.547) accepts, the "problem of social justice" can be solved "by eliminating in thought the circumstances that give rise to it." For example, feudal systems tend to assume some form of natural order, wherein "the basic structure is said to be already determined, and not something for human beings to affect" (Rawls, 1971, p.547). As such, there is a concern that justice as fairness could be rooted in ideological mechanisms which coerce citizens into adopting certain attitudes or belief systems. This is not the case in the well-ordered society. Instead, the original position requires that justice as fairness is confirmed by facts regarding society that are publicly known (e.g. that social structures tend to change over time). As a result, the well-ordered society cannot be stable "by promulgating false or unfounded beliefs", for one of these facts is that "men are not powerless to shape their social arrangements" (Rawls, 1971, p.547). In the Rawlsian approach, the principles of justice are a product of agreement, rather than an *a priori* natural order.

2.3.2. Authority

Here, Rawls (1971, p.500) writes that "the morality of authority when conceived as a morality for the social order as a whole" tends to "demand self-sacrifice for the sake of a higher good and to deprecate the worth of the individual." Subsequently, the concern is that deference to authority undermines citizens' sense of self-authorship and disproportionately requires citizens to sacrifice their ends in service of this authority. In contrast, the educative process of justice as fairness extends well beyond the morality of authority, which "has but a restricted role in fundamental social arrangements" (Rawls, 1971, p.467). Instead, each stage of education prefigures the next (and *vice versa*) until citizens possess a sufficiently nuanced understanding of the principles of justice (Rawls, 1971, p.515). In being formulated additively, the educative process is designed to adapt to the "limitations of human nature" (Rawls, 1971, p.515). Since persons are capable of understanding moral notions independent of an authorial figure, justice as fairness aims for these more advanced capacities, both *via* the later stages of moral development and by caring for citizens' respective conceptions of the good.

2.3.3. Envy

Again, Rawls (1971, p.539) anticipates this problem directly: "the sense of justice… [may be] a reaction-formation: what was originally jealousy and envy is transformed into a social feeling." Here, part of Rawls's response involves the plausible claim that whilst "persons have opposing interests and seek to advance their own conception of the good", this is not "the same thing as their being moved by envy and jealousy" (Rawls, 1971, p.540). To support this observation, Rawls draws an analogy to children competing for their parent's affection. In such instances, whilst these children may not have a full grasp on their sentiments, it seems plausible to say that "their social feeling arises from resentment, *from a sense that they are unfairly treated*", rather than from envy *tout court* (Rawls, 1971, p.540, emphasis added). More broadly, Rawls (1971, pp.538-540) responds to the problem of envy by claiming that "conservative writers" erroneously conflate envy and resentment, whereas the latter remains relevant to social justice because of its relationship to unfair treatment. In the Rawlsian view, the question is not whether justice as fairness is grounded in affective dispositions, but whether the relevant dispositions are present for justifiable reasons. "To insist upon equality as the two principles of justice define it is not to give voice to envy" (Rawls, 1971, p.538).

2.3.4. Self-Abnegation

From a broadly Nietzschean perspective, morality and frameworks of social justice are paradigms designed to undermine (rather than encourage or foster) the powers of the self. *Theory* argues against this view from two converging viewpoints. First, Rawls (1971, pp.414-432; 528) presupposes the truth of the "Aristotelian principle" and its "companion principle"; namely, that persons have a fundamental interest in pursuing their realised capacities (Aristotelian principle) and tend to appreciate it when others exercise their realised capacities (companion to the Aristotelian principle). Second, Rawls provides a Kantian interpretation of justice as fairness, wherein persons are defined as "free and equal rational beings." According to Rawls, the point of convergence between these viewpoints is that we all have a fundamental interest in realising our nature as free and equal rational beings. Hence, Rawls (1971, p.255) claims that if a person wishes to exercise their realised capacities, and if the two moral powers (rationality and reasonableness) constitute the highest of those capacities, then citizens will want to act from the principles of justice. For Rawls, this is best expressed *via* the original position. Since the parties of the original position are abstracted from their contingencies, and since they are defined as reasonable and rational citizens, then the principles that they choose will be best expressive of their nature as free and equal persons. Thus, the two principles of justice – chosen within the constraints of the original position – are those intended to manifest and encourage the powers of the self, not lead to their diminishment.

2.3.5. Community

Broadly, the 'communitarians' – a monicker generally attributed to thinkers such as Sandel (1988), Taylor (1985; 1989), and MacIntyre (1984) – argued that liberal philosophy overlooked the way in which persons are inextricably situated within communities. Emphasising how participation in communal life is required to make political discourse intelligible, this strand of thought positioned the value of community as central to the issue of social justice. In *Theory*, Rawls (1971, p.526) anticipates this critique both in his description of the sense of justice and his notion of a "social union of social unions". Concerning the former, Rawls (1971, p.495) highlights how the sense of justice is fundamentally established on the background of the community; "a capacity for a sense of justice built up by responses in kind would appear to be a condition of human sociability." On this view, persons would not have a sense of justice if it was not for its role in communal life, such that to say that justice as fairness 'generates its own support' (see, chapter 1) means that it does so through a nexus of political, personal, and cultural communities. This communitarian thrust to justice as fairness is also seen through the description of the well-ordered society as a "social union of social unions" since "the successful carrying out of just institutions is the shared final end of all the members of society, and these institutional forms are prized as good in themselves" (Rawls, 1971, p.527). Indeed, sociality and communal life are so fundamental to Rawls's project that the pursuit of justice can only ever be conceived as a shared endeavour, where all persons are expected to share in the value of just institutions.

2.4. The Importance of the Good

Above, I have outlined some of Rawls's specific responses to the problem of congruence within the well-ordered society. To grasp their place within the congruence argument, recognise, for example, the unpleasant feeling of discovering that one's life was organised according to a lie or that one was denied the good of the community. More often than not, the Rawlsian argument from congruence responds to these concerns by turning them on their head, e.g. justice as fairness not only recognises the value of community, but persons will also come to value the shared end of justice. Yet there is a broader point at play here, which allows Rawls to group seemingly divergent concerns under one banner: *everyone has a foundational interest in formulating and pursuing their version of the good life*; the more specific claims that Rawls makes (contingency, authority, envy, self-abnegation, and community) substantiate the working parts of *Theory*'s response to this interest in the actual complex of a human life.[17] As this summary shows, justice as fairness is not rooted in false beliefs, deference to authority, or unjustified envy, but instead upholds the value of community and affirms the powers of free and equal persons. In doing so, the congruence argument aims to demonstrate that it is rational to affirm justice as a regulative desire *from the individual point of view* (rather than that of the original position or the shared view of justice):

> Thus what is to be established is that it is rational… for those in a well-ordered society to affirm their sense of justice as regulative of their plan of life…. that this disposition to take up and to be guided by the standpoint of justice accords with the individual's good (Rawls, 1971, p.567).

To be clear, this more general view shall be the primary topic of the remainder of this thesis. My concern is not with the specific issues that might compromise Rawls's argument from congruence (e.g. how *Theory* responds to envy) but with the more general claim that goodness and justice are congruent within the well-ordered society. The benefit of this approach is that Rawls's general thesis is based on what he says about the nature of persons, meaning that the particulars of his argument are either fully substantiated or at least clarified by this analysis. For example, Rawls's response to the problem of self-abnegation is that *Theory* recognises the status of persons as free and equal. As I will show in the following chapters, the formulation and execution of a rational life plan define at least part of this status. More broadly, for the congruence argument to function as a stress-test for a conception of justice, Rawls must adequately classify the diverse interests, purposes, etc. of persons within society. Yet, for this stress-test to function, Rawls must adequately characterise persons and their capacity to formulate and pursue a conception of the good. Here, the notion of rational life-planning again proves crucial.

In sum, persons are motivated to act justly given the confirmation that their sense of justice – as a desire regulative of other desires – is congruent with their good; that is, their

[17] Another way of understanding this point is that, even under the realm of ideal theory, *individual versions of the good will inevitably differ*. Hence, a conception of justice must be tested to see whether it is compatible with a diverse range of viewpoints and practical interests (Freeman, 2003, p.284).

version of a life well-lived. This claim that justice and goodness can – and, according to the early Rawlsian view, should – be rendered congruent with each other will be the main topic of this thesis. Specifically, I shall examine how Rawls *defines* 'the good' and analyse whether that definition is plausible given the broad account of stability and this summary of the congruence argument. In chapter 3, I will unpack Rawls's notion of a 'rational plan of life', which he employs to define citizens' conceptions of the good. In doing so, I will highlight the empirical background to Rawls's account of life-planning, identify the principles of reasoning that define the rationality of a life plan, and clarify both its temporal and hierarchical structure. I also highlight the roles that rational life plans play within justice as fairness, including Rawls's account of personhood and his grounding of equality. Ultimately, I will offer a detailed examination of Rawls's definition of the good and its place within justice as fairness.

Chapter 3

Hierarchies, Schedules, and Personhood: The Background to Rawls's Account of Rational Plans of Life

3. Introduction

So far, this thesis has undertaken two major steps. First, I demonstrated that Rawls motivates the adoption of the well-ordered society through what I called the two generative properties of justice as fairness; broadly, that persons assess their own situation from the constraints of justice and view institutions as independently valuable. Second, I provided an account of Rawls's argument from congruence and defended it from Barry's criticisms. From a Rawlsian standpoint, I argued that the congruence argument is a kind of stress-test for justice as fairness, which brings to the fore the scope of justice and considers its compatibility with certain justified stipulations about human nature. In doing so, I offered an account of one of Rawls's main responses (the congruence argument) to the practical problem of stability – both in terms of the issues that it resolves and in terms of the mechanisms that it aims towards promoting – as well as a defence against a relevant criticism within the literature. I will now investigate the central part of the congruence argument, i.e. Rawls's definition of 'the good' as a person's rational plan of life.

The notion of rational plans of life is crucial to both Rawls's argument from congruence and his framework of social justice more broadly. To recognise this, first consider Freeman's summary of the congruence argument, which can serve as a reminder of the previous chapter:

> What assurance do we have that it is realistically possible for people to affirm justice as fairness as part of their good, and that they will then consistently affirm and act upon their sense of justice and regularly observe requirements of justice? (Freeman, 2007, p.264).

Next, note that another way of construing this problem is through the reconciliation of two perspectives. On the one hand, there is the perspective of justice (or, if you'd prefer, the original position): an essentially shared viewpoint, freed from its contingencies, and defined by its commonality (Freeman, 2007, pp.265-266). On the other hand, there is the perspective of the individual: a viewpoint that is individuated, embedded within its contingencies, and defined by rational plans of life. As I will demonstrate for Rawls, a person is a human life lived according to a plan, and what is good for that person can be characterised as those things that it is rational for them to want given that plan. Thus, the notion of rational life-planning is a bedrock of Rawls's argument from congruence, standing as the conception of goodness, or the perspective of individuals, within this argument. Given this notion's significance for justice as fairness, this chapter is dedicated to investigating the background and theoretical implications of Rawls's account of rational life-planning.[18]

With these points in mind, this chapter proceeds as follows: in section 3.1, I offer a summary of RPLs by examining the psychological theses that underpin Rawls's definition of the good. In section 3.2, I provide a taxonomy of planning to clarify the level at which life-planning is meant to take place, as well as its primary function in conduct-organisation. In section 3.3, I provide an overview of Rawls's (1971, p.415) account of deliberative rationality, enumerating his so-called "counting principles"; in doing so, I define from a Rawlsian perspective how life plans are rendered *rational*. Finally, in section 3.4, I clarify the main roles that RPLs play within justice as fairness, as well as highlight what Rawls takes to be the advantages of his approach. Ultimately, these discussions situate rational life plans as a central concept in Rawls's argument from congruence, and specifically, as one of the main claims that he makes regarding the nature of persons. In taking this step, I provide a detailed investigation into the concept of rational plans of life, thus setting up the critical discourse of the ensuing chapters.

3.1. The Empirical Background to Rawlsian Life Plans

Broadly put, Rawls takes RPLs to consist of a hierarchical structuring of aims and sub-aims across a person's lifetime, organised according to the principles of deliberative rationality. Although Rawls (1971, p.452) states that his definition of life-planning is not an outrightly technical one, I will show now that the technical conception of RPLs contributes significantly to, and thus sheds light on, Rawls's own account. I will also use this technical analysis to further substantiate Rawls's position on life-planning and, therefore, his definition of the good. To this point, Rawls (1971, p.408, n.10) references Galanter, Miller, and Pribram (1960), Goldman (1970), and Galanter (1966) within *Theory*'s discussion of RPLs. Throughout this chapter, I

[18] From this point onwards, as a stylistic choice, I will sometimes refer to rational plans of life as RPLs, life plans, and rational life plans.

will focus solely on Galanter, Miller, and Pribram's work.[19] To this end, consider Galanter, Miller, and Pribram's summary of the main intuitions behind the importance of planning to the human condition:

> Consider how an ordinary day is put together. You awaken, and as you lie in bed… you think about what the day will be like—it will be hot, it will be cold; there is too much to do, there is nothing to fill the time…Whether it is crowded or empty, novel or routine, uniform or varied, your day has a structure of its own—*it fits into the texture of your life*. And as you think what your day will hold, you construct a plan to meet it. *What you expect to happen foreshadows what you expect to do* (Galanter, Miller, and Pribram, 1960, p.5, emphasis added).

To provide context for this quote, Galanter, Miller, and Pribram (1960, p.12) highlight the lack of interpretative work on the relation between intention and action within psychology. As they argue, standard accounts of action – for example, behaviourist and cognitivist theories – tend to focus on stimulus-response relationships in human behaviour, without fully explaining the operative interactions between these two notions. On the one hand, behaviourists often attempt to capture advanced human cognition through an essentially physiological reflex arc (Galanter, Miller, and Pribram, 1960, p.6). Whilst, on the other, cognitivist approaches insist that stimuli-response relationships "must be mediated by an organized representation of the environment" (Galanter, Miller, and Pribram, 1960, p.7). However, in both cases – whether it is between physiological stimuli and cognitive responses or environmental representations and active conduct – there is a theoretical gap that marks the movement from cognitive representation to overt action; a movement that, as Miller, Galanter, and Pribham point out, we usually experience as seamless:

> If a person forms a clear image of a particular action, that action tends to occur …The problem is to describe how actions are controlled by an organism's internal representation of its universe…What we must provide, therefore, is some way to map the cognitive representation into the appropriate pattern of activity. (Galanter, Miller, and Pribram, 1960, p.12)

Importantly, this problem is exacerbated by the fact that even seemingly straightforward intentions– for example, making a cup of tea – are mired by seemingly unending levels of complexity. To complete this single activity, different muscle groups have to be engaged (e.g. raise arm, clasp hand...), several tasks have to be executed (e.g. boil water, add tea bag…),

[19] My reason for this is that Galanter, Miller, and Pribram's work is more directly associated with Rawls's notion of planning. For example, Rawls's (1971, p.408, n.10) citation of Goldman regards the relation between planning and action, rather than planning as such.

timings must be accounted for (e.g. leave the bag in for *x* amount of time…) and so on. As a result, the characteristic response from psychology – spearheaded by Tolman (1932) – is to separate patterns of action into formal units, with larger units being termed "molar" and smaller units "molecular." In this way, molars encompass molecules at a higher-order level, whilst molecules provide further specificity to the structure and order of the original unit. As Miller, Galanter, and Pribram (1960, pp.13-16) clarify, the molar *X* can be subdivided into the molecules *A* and *B* (entailing that *X* = *AB*), whilst A and B can be further subdivided into ab and cde respectively (entailing that *X* = *AB* = *abcde*). Since this is a general construal, the ordering can represent many different standardising factors, e.g. temporality, conditionality, and so on. In this way, seemingly complex structures can be simplified through a process of subdivision and organisational layering.

With that being clear, I now want to take an important step in investigating the technical background of RPLs. First, keep in mind that Galanter, Miller, and Pribram are concerned with describing the move from mental representation to action. Second, recall that even the formulation and execution of basic tasks involves the coordination of many diverse multi-layered sub-tasks. In response, and as my third point, Galanter, Miller, and Pribram posit that a simplifying process is necessary for the successful execution of apparently complex patterns of action. Indeed, they criticise attempts to explain this process in terms of a *reflexive response*, for this still leaves indeterminate the supposed vacuum between the cognitive representation and the subsequent action (Galanter, Miller, and Pribram, 1960, p.19). As an alternative, Miller, Galanter, and Pribram's description of action begins from the internal representation of the organism's universe to a systematic organisation of putative acts or behaviours according to their relative valance.[20] Hence, no theoretical gap is posited between expectation and actions since this process encompasses all its constitutive elements under one organisational schema. Intentions and actions are explained through the simplifying process of a plan, which is defined in the following way: "A Plan is any hierarchical process in the organism that can control the order in which a sequence of operations is to be performed" (Galanter, Miller, and Pribram, 1960, p.16). Crucially, there are many parallels between this technical formulation and the definition of 'a plan' that Rawls puts forward in *Theory*:

> The aim of deliberation is to find that plan which best organizes our activities and influences the formation of our subsequent wants so that our aims and interests can be fruitfully combined into one scheme of conduct…A plan,

[20] Miller, Galanter, and Pribham (1960, p.64) do not use the notion of valance; however, since I am only exploring their work insofar as it pertains to Rawls, it seems to me an appropriate description.

> then, is made up of subplans suitably arranged in a hierarchy… (Rawls, 1871, pp.410-411).

In this way, both interpretations hold that planning is an introspective or deliberative practice which subdivides complex tasks and activities into a sequential, hierarchical, structure. I suggest that Rawls's approach to planning involves three forms of hierarchical ordering. First, there is hierarchical ordering as an *internal feature* of individual plans; for example, I can prioritise brewing a coffee before making breakfast within the context of my morning routine. Second, there is hierarchical ordering *between* plans; for instance, I can plan to make breakfast *or* to go for a walk. If this binary hierarchisation is correct, however, an additional plan is required to override tensions between, across, or within plans. Therefore, finally, the plans themselves are ordered according to a *unifying structure*, with a level of thematic consistency established over and above the individual subplans. Accordingly, even if a plan is constituted by a range of subplans, the hierarchy that a life-plan imposes ought to be considered more or less consistent, and more or less final, with "the whole plan…[possessing] a certain unity, a dominant theme" (Rawls, 1971, p.420).

Moreover, this unifying effect is one that we see in Miller, Galanter, and Pribham's technical formulation. Recall that larger-group molars can be subdivided into molecules and consider two possible explanations of their connection: a chain-link account (akin to the aforementioned behaviourist approaches) and a plan-account. According to the chain-link account, assuming some specific connection between molecules and molars, the execution of *X* leads to *B*, then *A*, and then ultimately to *e*. As a result, there needs to be no connection between these constitutive parts aside from this interaction; that is, outside of being a link within a casual chain. "When a chain is initiated with no internal representation of the complete course of action, the later parts of the chain are not intended" (Miller, Galanter, and Pribham, 1960, p.62). On the plan-account, however, we cannot make sense of these constitutive molecules *without invoking their role as part of a greater whole*. In this way, we start from *X* in order to explain *B*, justify *e*, exclude *c*, and so on. Therefore, instead of being a chain-link that connects sub-plans and sub-actions, such occurrences – in both the technical and philosophical notion of planning – are *embedded* within a higher-order plan. In this way, each sub-plan involves its own hierarchies, which are themselves situated within a higher-order structure. Indeed, sub-plans that are ultimately rejected (I want to have breakfast…) are rejected *because* of their place within this organisational hierarchy (…but I have to go to work).

In sum, then, both Rawls's philosophical construction and the 'technical' definition maintain that life plans, (i), require a process of deliberation or introspection, (ii), concern the evaluation and sequencing of actions and intentions, and (iii), organise subplans and establish a cross-plan, unifying, theme. In this light, though Rawls distances himself from these technical assumptions, it is evident that they contribute significantly to his understanding of RPLs. Indeed, this is in line with Rawls's (1971, p.571) broader position, which relies heavily on psychology when approaching the problem of stability: "participating in the life of a well-ordered society is a great good…this conclusion depends … [in part] upon the psychological features of our nature." In this section, I have drawn from Rawls's influences in psychology to provide a brief explanation of planning as a simplifying form of conduct-organisation. However, it should also be noted that a plan is not organised purely through the relative valance of its sub-plans. Instead, the higher-order structure of plans also concerns the 'sequence of operations' or 'scheme of conduct' that the agent then executes. As a result, there is an explicitly temporal element to the formulation and execution of plans. With this in mind, I will now investigate what Rawls means by a rational *life* plan.

3.2. A Taxonomy of Planning

At a gloss, a life plan is a singular coherent plan that extends across the entirety of the person's life. More specifically, as per the above, a life plan is a scheme for action that hierarchically and sequentially organises lower-order plans. This sequencing – or "scheduling", as Rawls (1971, p.410) calls it – must then account for the major aims that occur throughout a person's life, plotting the order in which those aims are realised. In that respect, life-planning can be understood as the *highest-order form of conduct-organisation*, providing structure to subplans according to the forms of hierarchisation mentioned in the previous section (i.e. between, within, and across plans). Subsequently, in order to fully understand RPLs, it is necessary to clarify its connection to other forms of planning; in particular, plans of a lower-order within the context of hierarchisation. To accomplish this, I will draw from Heyd and Miller's (2010, p.19) four-level taxonomy of planning. For the most part, I have used Heyd and Miller's proposal only as a touchstone, expanding on it in my own way and providing my own examples where appropriate:

> *First-Level Planning*: At the first level, or the lowest-order level of planning, a person sets out to achieve a particular pre-defined goal. Properly speaking, first-level planning is both localised and broadly self-sufficient. Most well-formed intentions – for instance, planning to watch the football at 7.45pm – are included within this class of plans.

Hence, the primary mark of first-level planning is a short-lived localised aim, which usually requires no further action once realised.

Second-Level Planning: At the second level, long-term aims arise with a deliberately imposed structure. For example, a financial plan which organises sub-plans and sub-aims according to relative urgency (e.g. paying the rent on time, and only then saving for retirement). Note that this form of planning is still limited: second-level planning depends on plans formulated at advanced levels of organisation, they are long-term but not global, and they do not cover all or even most of one's activities.

Third-Level Planning: At the third level, there are large-scale changes to one's life throughout a restricted period of time. For example, a person may be temporarily transferred onto a witness protection scheme. Understood in this way, these plans are comprehensive but not global. Though their evaluative scope will be greater than those at a lower-level of planning, their horizons for action often have a definitive temporal limit.

Fourth-Level Planning: At the fourth level, a plan is formulated to account for the effective totality of a person's purposes and interests across the entirety of their lifespan. Structure is a priority for RPLs. Rather than account for the actual totality of one's desires, a life-plan usually identifies large-scale guiding aims or values and then systemises them within a plan. Concomitantly, fourth-level planning involves both the linear and lateral organisation of lower-level plans: structuring them linearly in terms of when they will be completed and laterally in terms of their relative valence. As a result, this class of planning involves the prioritisation or exclusion of subplans: "desires that tend to interfere with other ends…are weeded out; whereas those that are enjoyable in themselves and support other aims as well are encouraged" (Rawls, 1971, p.411; Mabbot, 1953). Similarly, themes between lower-order plans begin to take shape. In this way, RPLs organise other forms of planning at a higher level of abstraction, thus providing structure throughout a person's lifespan.

In summary, RPLs are fourth-level plans that stand as the highest-order form of conduct-organisation. They dictate the range of activities that a person participates in as well as the ends that motivate their conduct. At the same time, RPLs simplify other plans by identifying, and providing structure to, their overarching aims or motivations. Whilst other plans are either temporally or evaluatively limited, life plans concern the effective totality of a person's major purposes and interests.

Crucially, then, RPLS are fourth-level, highest-order, plans because they have *global scope both temporally and evaluatively over one's system of ends*. Whilst, similarly, RPLs can be summarised as a *top-down* rather than bottom-up approach to the prioritisation of values and activities. "A plan will, to be sure, make some provision for even the most distant future and for our death" – whilst it is expected that RPLs take precedence over other forms of planning – "changes at the lower levels do not usually reverberate through the entire structure" (Rawls, 1971, p.410-411). Hence, Rawls (1971, p.409, emphasis added) thinks that it is appropriate to give RPLs finality when deciding both what is good for a person and who that person wants to be: "a rational plan of life establishes the basic point of view from which *all judgments of value relating to a particular person are to be made and finally rendered consistent*." I will expand on this thought in section 3.4. In the next section, however, I will complete my description of RPLs by explaining the notion of deliberative rationality, which is the method of reasoning that Rawls provides for constructing a rational life plan.

3.3. The Principles of Deliberative Rationality

For Rawls, a plan is only a plan if it is rational. To see this, contrast plans with hare-brained schemes. Broadly, a hare-brained scheme organises action through premises which – in light of any form of focused attention – would prove patently false. Deciding to steal the crown jewels by pretending to be the king, walking into the Tower of London, and demanding that the guards help me get dressed is a hare-brained scheme. If I stop to think for even a moment about doing this, I will realise that Charles III and I look nothing alike. In contrast to plans, a hare-brained scheme is carried out despite its otherwise obvious downfalls. Neither, it should be added, do we plan 'accidentally'. If I were to walk around Tower Hamlets and have the crown jewels fall miraculously into my lap, it would be incorrect to say that I had planned to take them (and I would certainly claim this if I were arrested). Instead, plans seem to presuppose a formal structure as well as relevant constraining conditions.

Here, the notion of deliberative rationality that Rawls employs fleshes out these constraining conditions and the broad structure that a life plan is supposed to take. Moreover, Rawls advances this refinement of life-planning by providing principles of reasoning, by which (sub-)plans and (sub-)actions are organised so that they constitute a unified schema. Hence, if the sections above can be taken as describing the activity of planning and the temporal scope of distinct orders of plans, this section describes the rationality of life plans. For the time being, although I discuss this again in sections 4.3 and 4.4, I will say that a Rawlsian life plan is rational once it has been formulated in light of the principles of rationality *via* a process of introspection. Subsequently, I intend to provide an overview of Rawls's counting principles to complete my unpacking of rational life-planning. In undertaking this task, I have used (when available) the names that Rawls provides for these principles and the term 'subprinciple' for those rules which Rawls employs but does not name; no other order is provided:

Principles for Hierarchisation:

> (1h) *Principle of Effective Means*: RPLs should take into the account the resources required to complete an objective. This means that given some end, "one is to achieve it with the least expenditure of means (whatever they are); or given the means, one is to fulfil the objective to the fullest possible extent" (1971, Rawls, p.412). Hence, when deliberating between plans with the same end, one should choose the plan according to the extent that its aim can be achieved, and the resources required for successful completion of the aim.
>
> (2h) *Principle of Inclusivity*: Inclusive plans ought to be prioritised over less or non-inclusive plans (Rawls, 1971, p.412). Given some set of desires, *xyz*, plan P^1, which achieves both x and y, should be chosen over plan P^2, which achieves only *z*. According to this principle, the greater proportion of realised aims contributes to the satisfaction that one has with one's RPL.
>
> (3h) *Principle of Greater Likelihood*: One should pursue the plan with the greatest chance of success (Rawls, 1971, pp.412-413). This can be taken in conjunction with the principle of inclusivity. Hence, if P^1 has a 70% chance of achieving *x* and *y*, P^2 has a 70% chance of achieving just z, whilst P^3 has a 62% chance of achieving x and y, one should still opt for P^1. According to the principle of inclusivity, one should prefer aims that occur further up the planning structure, since this will involve the realisation of more sub-aims. Whilst according to the principle of greater likelihood, one should prioritise between chains according to their chance of success.
>
> (4h) *Subprinciple*: One should be aware of the genesis of one's desires. The examples Rawls (1971, pp.419-420) presents are when desires are founded on "excessive generalization", are derived "from more or less accidental associations", or have "acquired their peculiar urgency as an overreaction to a prior period of severe

deprivation or anxiety." Broadly speaking, putting the origin of one's desires under critical scrutiny is an assurance mechanism that helps verify one aim as more important than another. Here, an interesting and plausible consequence is that planning involves both prospective and retrospective forms of deliberation. For example, not only does one consider what one's desires are now, but why one has those desires in the first place.

Principles for Scheduling:

(1s) *The Principle of Postponement*: One should postpone deciding when the best course of action seems indeterminate. When life-planning, there will be moments when two or more plans seem viable. Rawls's (1971, p.410; 420) solution to this dilemma is that one should put off deciding between these possibilities for as long as possible, until more information is available.

(2s) *The Principle of Continuity*: Continuity between plans is in itself desirable. As I identified in the previous section, planning is meant to be a unifying practice. An optimal RPL will structure second- and third-order plans so that their ends harmoniously interact and overlap. Not only will this enable more effective scheduling, ordering plans so as to more efficiently realise desires later on, but it also has the additional benefit of providing "a dominant theme" to one's life (Rawls, 1971, p.420).

(3s) *The Principle of Rising Expectations*: Plans with rising – or at least, not declining – expectations ought to be preferred (Rawls, 1971, p.421). This is the case even if the estimated utility of both plans is equal. The justification here is that individuals will take enjoyment out of the increasing success of their plans and be comforted by the avoidance of setbacks. Hence, plans should be scheduled in ways that prioritise incremental increases in their expectations and results.

Under the method of deliberative reasoning, then, one will attempt to maximise the resources available, make effective use of one's time-spent-planning, be aware of the origins of one's desires, and try to accomplish the greatest amount of aims with the greatest likelihood of success. These plans will then be ordered in terms of their relative expectations, the amount of effort required to complete and choose between subplans, and in a manner that establishes continuity between subplans. In accounting for these principles at the fourth-level of planning, one can establish a rational plan of life.

With the description above in mind, Rawls (1971, p.405) names his account of the good "goodness as rationality." That is, a person's good is given *via* their major purposes and aims in an appropriately formulated rational plan of life. Moreover, as I have shown throughout this chapter, this 'appropriate formulation' of a life plan depends on an introspective procedure whereby a person reasons from the principles of deliberative rationality. As one generally expects, an object might be good to the extent that it fulfils its *ergon* or purpose, but this description requires further elaboration when applied to persons. For example, goodness often necessitates contextualisation (e.g. a guitar can be good for jazz, but not for punk). Similarly, in the case of a person's good, there will be a background that elucidates their specific aims and purposes. For Rawls, *this background is a person's plan of life*, which itself has been rationally formulated, and from which the rationality of their interests is derived. Thus, Rawls (1971, p.399) states that "once we establish that an object has the properties that it is rational for someone with a rational plan of life to want, then we have shown that it is good for him." More extensively than this, and as I will continue to explore throughout this thesis, Rawls defines goodness (i.e. persons respective versions of a life well-lived) through this notion of rational life-planning: the introspective, unifying, process whereby one formulates and pursues one's major aims and purposes. Having investigated the ins-and-outs of rational life planning, I will close this chapter by expanding on the importance of RPLs to the overall project of justice as fairness.

3.4. The Importance of Rational Plans of Life to Justice as Fairness

As I have argued in chapters 1 and 2, Rawls's response to the problem of stability depends on RPLs as an accurate characterisation of persons. Here, I want to point towards some of the broader implications of this claim. Namely, that Rawls's description of persons *via* RPLs also supports his analysis regarding the claim of persons *qua* citizens. Here, a broad analogy can be brought in relation to human rights. For example, Gewirth (1978, 1982, 1996) attempts to ground the universal rights to well-being and freedom in the distinct capacity of persons for agency. According to Gewirth's broad view, all persons are required to accept the value of their own agency, as well as respect the agency of others as a logical corollary. By undertaking this approach, he attempts to support the universal application of particular rights (i.e. freedom and well-being) through accurately characterising the nature and values of persons, whilst drawing out further rights *via* a principle of logical consistency. Although Rawls's intention is perhaps less strongly normative than Gewirth's, given that the justificatory stage of justice as fairness has already been completed, it will become clear from this discussion that the notion of RPLs still plays a significant role in the broader political context of the well-ordered society. Specifically, I will examine the connection between RPLs and the equality and liberty of citizens, before investigating how Rawls uses RPLs to define happiness and, ultimately, unified selfhood.

3.4.1. The Basis of Equality

When justice as fairness is instantiated within a well-ordered society, Rawls expects that RPLs will play an important role in grounding equality when it comes to claims to social justice. More specifically, Rawls (1971, p.504, emphasis added). uses RPLs in conjunction with the sense of justice to respond to the higher-order question regarding the "sorts of beings" that are "*owed the guarantees of justice.*" To grasp this problem, consider the following intuition: whilst human beings can be treated justly or unjustly, different moral concepts are required to describe the treatment of animals, e.g. compassionately or cruelly (Rawls, 1971, p.505). In this way, it seems that there is *something about being human* that gives persons equal claim to the 'guarantees of justice.' Since all humans (or nearly all) are capable of developing an effective sense of justice, and since they will have their own conception of the good defined through RPLs, Rawls (1971, p.505) suggests that these capacities (i.e. 'moral personality') provide a suitable foundation for equal justice.

There are, however, three caveats to this point. First, in *Theory*, Rawls (1971, p.505) does not state that moral personality is necessary for equality, only sufficient for it. Second, all that is required for equal justice is the *capacity* for moral personality – meaning, for example, that one does not need to have a fully effective sense of justice to be able to claim fair treatment (Rawls, 1971, p.505). Finally, this grounding of equality is not directly argued for by Rawls (1971 p.510, emphasis added), but instead stands as an expected consequence of the fulfilment of justice as fairness: once the principles are publicly known and institutions are organised accordingly, citizens will come to recognise that "*those who can give justice are owed justice*". In other words, 'moral personality' – the possession of an RPL and the capacity for a sense of justice – is (at minimum) a sufficient condition for equal justice amongst persons.

3.4.2. Securing the Grounds for the Priority of Liberty

In the context of the problem of stability, part of Rawls's justification for the prioritisation of liberty in the ordering of the justice principles is the diminishing returns of social and economic goods. That is, as the "conditions of civilization improve", then the relative benefit from increased social/economic advantages becomes "marginal", whereas the worth of basic liberties "become stronger as the conditions for the exercise of the equal freedoms are more fully realized" (Rawls, 1971, p.542) Here, it might be responded that the idea of diminishing returns of economic advantages is misplaced, since citizens do not *only* retain the concrete reward of material goods, but also benefit gained from increased social status (the value of which will not diminish in the same way). However, much of Rawls's arguments – e.g. those regarding the persistence of envy, the promotion of self-respect, and the value of social unions – are dedicated to minimising the allure of such positional goods. As Rawls (1971, p.544) puts it, "in a well-ordered society the need for status is met by the public recognition of just institutions, together with the full and diverse internal life of the many free communities of interests that equal liberty allows." Assuming that basic wants and needs are met, the utility value of economic and social advantages is thought to lessen as the conditions for the realisation of liberties becomes greater.

Keeping the above point in mind, it is possible to see why RPLs justify the priority of liberty within the problem of stability. Given some threshold amount of economic and social resources distributed according to the difference principle is met within society (and given the other assurances that justice as fairness provides), the likely result is that citizen's plans will be more successful and diverse than they would have been otherwise, had the threshold amount not been met (more on this in section 3.4.4). Yet as this trend continues, citizens will not be able to *guarantee* the success of their plans – even though their availability, their diversity, and the likelihood of their success have increased by meeting or going beyond this threshold. The consequence of this is a tipping of the scales. Instead of securing further resources that will only have a limited effect on the expected success of their plans, citizens will want to solidify *their right to formulate such plans*, in order to enjoy the liberties that the first principle of justice offers them. As Rawls (1971, p.543) writes in this context, "under favorable circumstances the fundamental interest in determining our plan of life eventually assumes a prior place."[21] Subsequently, one can understand the prioritisation of liberty as the primary mechanism by which citizens protect their capacity to form and pursue their plans: as the expected success of plans increases, the value of being able to determine one's own RPL becomes overrulingly important.

[21] Another way of making this argument is this: citizens will begin to prioritise the principle of inclusivity and the security of their plans, rather than the principle of greater likelihood when expected returns begins to diminish. If this point is accepted, I can then put the role of RPLs in its strongest possible terms: the prioritisation of liberty *directly depends* on the notion of life-planning, since it seeks to consolidate the capacity to pursue plans which obtain at a higher-level within the organisational hierarchy.

3.4.3. The Definition of a Happy Life

For Rawls (1971, pp.548-554), a person is happy when they have achieved the successful ordering and execution of an RPL.[22] To grasp this claim, contrast it with a broadly hedonistic philosophy, which takes happiness to consist of a subjective feeling of pleasure. Taking a hedonistic view, what can be said to a person hoping to live a happy life? One can offer advice, tell them to discover pleasure and maximise its intensity/duration. Ultimately, however, a concession must be made. Dilemmas will arise that cannot be answered clearly by a hedonistic approach. For example, "Aristotle says that the good man… prefers a short period of intense pleasure to a long one of mild enjoyment, a twelvemonth of noble life to many years of humdrum existence" (Rawls, 1971, p.557). Yet such qualitative distinctions are unnavigable when the sole instruction is to increase some subjective feeling. Once the advice outlined above is exhausted, persons must reach their own decision about *how* to maximise pleasure; they "must make this decision, taking into account the full range of…[their] inclinations and desires, present and future" (Rawls, 1971, p.557). Thus, Rawls thinks that this broadly hedonistic standpoint is reducible to happiness defined through the successful formulation and execution of an RPL. After all, the hedonist must still reflect on the available facts and deliberate accordingly, only to then formulate a plan of action that best realises their desire for subjective pleasure. For Rawls, it is in the formulation and the fortuitous execution of this plan that the hedonist is happy, *over and above* their attainment of pleasure.

3.4.4. Respecting Diversity and the Route to Self-Realisation

The notion of RPLs reflects a positive diversity among citizens' ends. This point can be drawn out through a Rawlsian distinction between a 'concept' and a 'conception'. Put broadly, a concept is identified with its *intended role*, whilst a conception can be defined as an *interpretation of the given concept*. For example, 'numerical notation' can be our concept, whereas 'the Babylonian' and the 'Hindu-Arabic' numerical systems can stand as our conceptions. Whilst the intended role of numerical notation remains the same – namely, the ordering and the expression of numbers – the systems for notation can differ in their interpretation. In our example, the Hindu-Arabic system uses base-10 for notation, meaning it employs ten symbols from 0-10, whilst the Babylonian system uses base-60, using symbols for numbers 1-59 (with a placeholder for zero). Here, though the role of numerical notation remains largely the same, the systems differ meaningfully in terms of ordering, in the symbols

[22] To note, happiness is not identical to goodness. Not everyone's conception of the good will include the end of happiness, whilst others may only include it partially. However, Rawls thinks his life-planning approach to goodness can encompass the notion of happiness also.

they use to represent numbers, and even in the numbers that they employ (this latter point is deliberately contentious, with an example being the respective systems' use of 'zero'). Hence, 'the concept' (numerical notation) retains its unity – i.e. there may be many different conceptions that aim at capturing *a singular coherent role* – whilst 'conceptions' will remain naturally diverse (numerical systems). To this end, justice as fairness is not concerned with the good *per se*, but with *the flourishing of various conceptions of the good*.

Likewise, a citizen's RPL, *qua* conception of the good, will involve a particular rendering of what they take to be 'the good.' Just like in the example of numerical notation, this interpretation will be one among many. The upshot is that the well-ordered society will be marked by an *omnium gatherum* of different RPLs, some of which will maintain distinct but overlapping aims. Consequently, citizens will depend on other members of society both in the construction and in the maintenance of their RPL. Rawls, for his part, sees this as a major benefit. For example, RPLs will encourage individuals to accept the importance of social cooperation. Similarly, planning for our future can help us recognise that we *depend on others for our own self-realisation*: "it is as if others were bringing forth a part of ourselves that we have not been able to cultivate" (Rawls, 1971, p.448). This specific point is undeveloped by Rawls – who opts instead to focus on the social aspects of the sense of justice – but it remains an interesting advantage. Namely, that RPLs stand as a way of making the necessity of social cooperation explicit by forcing us to account for the ways in which we rely on others to realise our own ends, and ultimately, our sense of self.

3.4.5. Personhood and the Roycean Roots of Justice as Fairness

I have mentioned throughout the previous sections that RPLs are meant to provide a form of unity to one's life. In Galanter, Miller, and Pribram's technical sense, this might mean that life-planning is the 'molar' that encompasses all of the lower-order 'molecule' plans. Rawls, however, uses the notion of RPLs in a much stronger sense. More specifically, he follows Royce (1908) in claiming that "*a person may be regarded as a human life lived according to a plan*" (Rawls, 1971, p.408, emphasis added). As Royce (1908, p.107) puts it, "wherever a human selfhood gets practically and consciously unified" it does so through a plan of life, which is not "defined…in abstract terms." This means that "*your life as it is lived, your experiences, feelings, deeds*, these are the embodiment of your ideal [i.e. rational life-] plan" (Royce, 1908, p.176, emphasis added). Hence, life-planning is not an abstract procedure – intended purely to advance a philosophical framework – but a concept that captures real aspects of our experience. It is apparent, moreover, that Rawls adopts this position. Even a brief gloss of *Theory* supports this interpretation, with evidence consisting in Rawls's almost word-for-word Royce's (1908, p.106) summary – "a person…may be defined as a human life lived according to a plan" – and his similar continuation of Royce's (1908, p.175) definition of selfhood through "moral personality", i.e. a sense of justice and a capacity for life-planning.

Taking this further, for Rawls (1971, p.565), rational plans of life capture the autonomy and unity of the self: "the nature of the self as a free and equal moral person is the same for all, and the similarity in the basic form of rational plans expresses this fact."[23] In this way, Rawls (1971, p.493) defines an "intentional action" as one that "given our beliefs and the available alternatives… accords with our plan of life." Hence, the description of RPLs that I have provided in the sections above – as in the technical description that I began with – is intended to substantiate an account of persons *qua* practical selves. In this way, RPLs secure the grounds for liberty because they capture what it means to be an agent; they confirm the grounds for equality because we all formulate plans and are capable of a sense of justice; they provide the means for self-realisation because they constitute our basic sense of self; and they describe a happy life because they express a unity that takes precedence over momentary states of pleasure. For Rawls, to be a self that is situated within a practical world *just is* to formulate and pursue a rational plan of life. Similarly, a person's good – those values, purposes, aims, desires, interests, and associations that constitute their version of a life well-lived – is exhausted by their rational life plan.

For Rawls, rational life-planning is the highest-order form of conduct-organisation, which one undertakes through an act of introspection. As I have clarified, it is planning of the highest-order because it systematises lower-order plans at an advanced level of abstraction. Moreover, this structure consists of an evaluative and temporal hierarchy, formulated under the principles of deliberative rationality. As a result, RPLs provide their owns aims as well as structure sub-plans and sub-actions into a coherent schema. For Rawls, this schema is constitutive of the unified self. Accordingly, rational life-planning is a means to self-realisation, a foundation for equality and liberty, and a measurement for human flourishing. Importantly, this is not a hypothetical construction. For Rawls, *we are our plans*. To be a self, an agent, and to formulate a conception of the good all express the same thing for Rawls: one's capacity for rational life-planning.

[23] Here, I will remain general. Both of the commitments from Rawls, i.e. unity and agency, will get detailed treatment in chapters 4 and 5. My intention currently is to provide a general overview of Rawls's account of persons.

With this groundwork being completed, the upcoming chapters will be dedicated to critically engaging with Rawls's account of life-planning, and specifically, assessing whether it is able to fulfil the roles assigned to it here (i.e. goodness, agency, and unified selfhood). To begin with, I will defend Rawls from the claim that his account of life-planning involves a temporal abstraction. In turn, I will clear the way for my own critique, which draws out a tension between Rawls's account of deliberative rationality and his description of the self's unity. Following this, at the start of chapter 5, I consolidate the claims made at the end of this chapter by demonstrating that Rawls's account of the self's unity directly depends on his notion of persons *qua* agents. Again, the lynchpin to this overall view is *Theory*'s account of life-planning. As a result, I advance two Sartrean criticisms of Rawls; first, that his account of agency is untenable; second, that a person's good is not defined by their rational plan of life. Accordingly, throughout the following three chapters, I shall establish the main critical contributions of this thesis.

Chapter 4

Temporality, Prudence, and the Unity of the Self: Investigating the Rationality of Rawlsian Life Plans

4. Introduction

Throughout the upcoming chapters, my main aim is to investigate whether or not the Rawlsian notion of rational life-planning is suited for intended roles within justice as fairness; namely, as Rawls's definition of the good and as his account of persons. In this chapter, I take the first step in this investigation by assessing the main criticisms of Rawls on this matter. To begin, I consider the argument that persons do not and should not plan *globally*, in the way that is required for them to formulate a rational plan of life. Here, Mackie offers a useful summary of this argument:

> People differ radically about the kinds of life that they choose to pursue. Even this way of putting it is misleading: *in general people do not and cannot make an overall choice of a total plan of life*. They choose successively to pursue various activities from time to time, not once and for all (Mackie, 1984, pp.354-355, emphasis added). [24]

Importantly, this trend can be further specified into two main camps: on the one hand, scholars that criticise Rawlsian life-plans because they ignore the basic situatedness of decision-making (e.g. Williams, 1981, 2009, White, 2021, and Mackie, 1984), whilst on the other hand, scholars that criticise Rawls for overlooking goods that cannot be anticipated through a global life plan (e.g. Slote, 1983, and Larmore, 1999, 2010). With this in mind, I incorporate both camps under the umbrella of the 'temporality critique', as they each focus on the temporal dimensions of RPLs, either in terms of the decision-making processes that are required to formulate a conception of the good or in terms of the goods themselves, which are taken to be time-sensitive. Hence, Steinfath (2023, p.10) offers an accurate representation of the literature when he speaks of the "double claim": that rational life-planning "distorts the temporality of human life and…thereby clouds the prospects for a good life."

[24] To clarify, Mackie's (1984, p.354) criticism is not directly aimed at Rawls, but at Aristotle. However, the implication for Rawls (1971, pp.92-93) is clear, given that his theory of the good "is a familiar one going back to Aristotle…The main idea is that a person's good is determined by what is for him the most rational long-term plan of life given reasonably favorable circumstances."

Both horns of the temporality critique threaten Rawls's argument from congruence and thus undermine one of his primary responses to the problem of stability. In each instance, Rawls mischaracterises the good, which, as I have shown thus far, is necessary to show its compatibility with justice in the first place. For the first horn (section 4.1.1), Rawls mischaracterises goodness by committing to temporal abstraction, wherein distinct sets of values are supposedly made static and neutrally compared. For the second horn (section 4.1.2), Rawls overlooks the fact that some goods are temporally-sensitive, and instead carries the importance of prudence to a global level. In this way, *Theory* would not show that justice and goodness are congruent, for – in both instances – the Rawlsian account of the good is either incomplete or unrealistic. Subsequently, justice as fairness would fail to promote the attitudes necessary for stability within the well-ordered society. Understood in this light, *on Rawls's own terms*, to mischaracterise the good is to risk jeopardising the prospects of a just society altogether. This theme shall prove relevant to the ensuing critical chapters on Rawls.

Nevertheless, I will argue that the temporality critique fails to make headway against the Rawlsian notion of rational plans of life. In particular, I respond on behalf of Rawls to demonstrate that life-planning is a situated process, wherein one's plans go under periodic reassessment and reformulation (section 4.2.1). Additionally, I demonstrate that Rawlsian life-planning – as a process of formulation and revision – can include non-prudential virtues *via* periods of non-planning (section 4.2.2). I argue that a lack of life-planning can be explained through the same judgements that constitute a life plan in the first place, such that a sense of openness, childlike playfulness, uncertainty, etc., are all virtues and judgements that constitute one's life plan *ceteris paribus*. I finish my response to the temporality critique by emphasising the unifying effect of life-planning, precisely in its capacity to encompass judgements of all kinds into a person's major purposes and aims.

On the back of this discussion, I formulate a different but related argument against Rawls's account of deliberative rationality. To begin with, I clarify that the congruence argument is ambiguous between two markedly different forms of life plans: objectively rational life plans (section 4.3.1), and subjectively rational life plans, that aim towards an objective standard (section 4.3.2). I argue that both approaches are in tension with Rawls's account of the unity of the self. In particular, I show that objective rationality is an abstraction that ignores the way in which a person's life plan is partly constitutive of who they are, such that – by enacting this separation of the self – Rawls's approach fails to characterise that person's good. Similarly, I argue that using objective rationality as a guiding standard for otherwise subjectively formulated life plans has much the same effect; it implies that a person's good is settled in advance, fostering a mindset in which persons feel as if they must optimise their plans rather than self-authenticate them.

In light of these issues, I tentatively suggest that Rawls's account of RPLs can be formulated without the conception of deliberative rationality; that is, without supposing the existence of a person's objective good and without assuming that their current plans should be aimed towards it (section 4.4). Instead, as in my response to the temporality critique, I emphasise the processual nature of life-planning and suggest that life plans possess an internal structure that can be utilised to regulate one's commitments, aims, etc. Similarly, I emphasise that the aim of the congruence argument *is* to get at the difficulties in life; as I explained in chapter 2, it is a stress-test intended to explore the scope of justice. Ultimately, then, this chapter defends Rawls's account of rational life-planning by clarifying its temporal structure and revising its underlying account of rationality. In doing so, I aim to provide a relatively robust rendering of Rawls's account of the good, at least one that avoids the criticisms that are most immediately placed against it. Nevertheless, the result of this clarification and revision is necessary for this thesis as a whole. The congruence argument does not depend on the impossible standard of complete information and true beliefs. Instead, on the Rawlsian approach, a person's good consists of their major aims and purposes, identified through reflection, and systemised within their plan of life. By interrogating Rawls's account of rationality, this chapter takes the first step in dismantling that view of goodness.

4.1. The Timeless Perspective of Rational Life Plans

To unpack the temporality critique, I shall begin by considering empirical literature on this matter. After all, as I demonstrated in section 3.1, the Rawlsian account of RPLs also has part of its foundation in psychology, whilst such empirical observations serve to anticipate and bolster the philosophical claims made by Mackie and the like. Put differently, a cursory overview of the psychological literature against life-planning will offer a more balanced picture overall. Keeping this in mind, Waldon, Patrick, and Duggan (2011, p.486) observe in their analysis that "many studies have found little or no planning ahead during problem-solving", noting that "it has often been concluded that human problem solvers operate opportunistically." In particular, one example cited is Delaney and Ericsson's (2004, p.1232) four-fold experiment on two groups attempting to complete novel problem-solving tasks; in this study, they found that "people typically do not plan in unfamiliar tasks… because complete look-ahead planning is often unnecessary." In this way, more recent empirical evidence suggests – *contra* Galanter,

E. Miller, G.A. and Pribham, K.H. (1960) – that people, mainly when presented with a restricted amount of information, do not generally tend to plan their responses to problem-solving scenarios.

Moreover, at least *prima facie*, it is plausible to generalise these findings to the larger scale of life-planning. For example, one explanation for the absence of life-planning – which I have extrapolated from Benson's (1997, p.71) analysis on planning-development in children – is that the process for learning to plan is itself temporally structured, i.e. we move from familiar sequences of actions, *via* some given routine, to the internal representation of such sequences in relation to a particular end. Yet, given that this process permits of a specific temporal structure, then the absence of life-planning can be explained by the lack of precedence for coordinated action across a global scale. In short, there are suggestions in psychology that not only do people operate opportunistically, but that the comprehensiveness of life-planning precludes it from being a practical form of conduct-organisation. Both the issues regarding the development of planning-based skills, and the lack of precedent for fourth-level planning, mean that formulating a life plan can be an inaccessible task to undertake from the perspective of human psychology.

Bringing this critique to a more firmly philosophical context, Williams (1981) provides a thematically similar line of attack to that of Mackie and the critical psychological literature on this matter. For Williams, to plan one's life is to commit to a form of abstraction which, as opposed to seeing decisions being made in the 'here-and-now', consists in a kind of perspective *sub specie aeternitatis*. Here, I think that Larmore provides an accessible summary of this criticism:

> It is Bernard Williams's complaint about the timeless form into which Socrates cast the question "How should one live?" And Williams has brought the same charge against Rawls…Williams argues…[that] the fact that our deliberation takes place in the present is not so trivial as it may seem. The results of practical deliberation… cannot be more substantive than the premises from which it sets out. As a result, our thinking, even when directed toward our life as a whole, must draw its bearings from our present perspective. It always carries the mark of where we are at the time (Larmore, 1999, p.107).

More specifically, Williams (1981, p.34) criticises Rawls for presupposing a "fixed and relevant" perspective in his notion of rational life-planning. To unpack this, note that there are two main perspectives that one can take up in relation to one's life plan: a prospective and a retrospective position (Heyd, and Miller, 2010, pp.21-25). Prospectively, decisions relevant to the life plan are made, whereas retrospectively, a fixed position of greater knowledge is used to assess the original choices. For Williams (1981; 2009), these two perspectives are *partly incommensurable* because *they are inseparable*. After all, it is not only that time will pass, but also that my commitments, my values, my associations, my desires, and so on, will meaningfully change and act as the condition for my future assessments and commitments, etc. Hence, both positions are limited by the temporal perspective that they each inherit. Here, the upshot of William's argument is that it is only through an act of abstraction that both perspectives can be subject to supposed balanced scrutiny. In this way, Rawlsian life plans assume an impossible third perspective (i.e. a fixed and relevant perspective) – that involves a form of timeless neutrality – whereby these two, ultimately irreconcilable, viewpoints are consolidated:

> The recourse from this within the life-space model is to assume…that there is some currency of satisfactions, in terms of which it is possible to compare quite neutrally the value of one set of preferences together with their fulfilments, as against a quite different set of preferences together with their fulfilments. But there is no reason to suppose that there is any such currency, nor that the idea of practical rationality should implicitly presuppose it (Williams, 1981, pp.34-35).

Here, Williams constructs a critical image of rational life-planning (related to the main point above) by drawing on terminology associated with geometry and economics – thereby connecting the planners' intentions to a kind of mechanical, time-insensitive, perfectionism. However, an analogy to language is a better explanatory device to convey this critique, as it emphasises the *global* abstraction that life-planning involves. To see this, consider different sets of commitments (e.g. my past commitments in relation to my future expected commitments) as distinct sets of propositions in different languages – say, French and English. It is true that we can attempt to translate between these languages. However, if we adopted a 'global approach' to translation – for example, by translating every sentence in terms of its literal definition – we would be making a significant error. The French phrase, '*sauter du coq à l'âne*' literally translated would be 'to jump from the rooster to the donkey', rather than 'to change subjects erratically'. Different words and sentence constructions seem more appropriate depending on *which language one is using*, and to assume an abstract 'third language' where all the propositions can be standardised outside of this originary context is both unnecessary and misleading. In doing so, one ignores the basic fact that the sentences' original *meaning* can be, as it were, 'lost in translation' – just as the value of a prospective decision can be lost when analysed retrospectively, the value of both decisions is distorted by the viewpoint of life-planning.

In this way, according to the first horn of the temporality critique, one ought not (and cannot functionally do so) assume a globalised system – i.e. a rational plan of life – wherein distinct sets of commitments can be neutralised and compared. On this critical rendition, Rawlsian life-planners attempt to adopt a position of abstract neutrality, wherein their commitments can be assessed on a global level, and wherein such assessments are not subject to the ordinary drawbacks of context-sensitive decision-making. The fundamental error of this approach, however, is that the comprehensive organisation of one's major aims and desires conflicts with the contingent nature of the decision(s) required to formulate one's plan of life. That is, the structure of life plans – both in regard to their global scope and the decision-making procedure required to formulate one – is a basic abstraction, which treats practical deliberation as if it were not situated within temporally-embedded contexts. As a result, by being concerned with the total organisation of one's major commitments, the perspective of life-planning is situated outside of the localised context wherein the relevant decisions are actually being made.

4.1.1. Prudence and Unexpected Goods

At this point, I want to unpack the second horn of the temporality critique; that is, if life-planning operates under a distorted view of temporality, then it clouds the prospects of a good life. The main argumentative move of this critique – particularly as it is presented by its foremost proponent, Slote (1983, p.47, emphasis added), in *Goods and Virtues* – is to associate life-planning with the time-relative virtue of "*planfulness*" (which Slote takes to be broadly analogous with, but not identical to, prudence)[25]. The issue is that, *qua* the temporal structure of life plans, Rawls carries the value of planfulness, which is beneficial only in localised periods of one's life, to a global level – without the sensitivity to context or the openness to serendipity that other non-global virtues allow for. In this way, if there are particular goods and virtues for specific periods of one's life, then life-planning overlooks what it means to live a good life *ceteris paribus*. More practically, the act of life-planning leads agents to miss out on boons that might otherwise shape their conception of the good.

[25] Given this broad similarity, and in light of Larmore's (2010, p.192) arguments *vis-à-vis* the similarities between planfulness and prudence, I will use these terms. I will also adopt Slote's term of 'planfulness'. However, in his version of this critique, Slote also criticises Rawls for taking planfulness to be a virtue *tout court*. To this end, I am not convinced, like MacIntyre (1985, pp.204-207), that Slote does enough to define a 'virtue *tout court*' – nor am I convinced that such a distinction is required to carry the weight of this argument. Hence, I will not adopt this part of Slote's proposed vocabulary.

An informative starting point for appreciating this critique is by recognising, with Slote (1983, p.2), that life-planfulness is a virtue most strongly associated with adulthood. After all, long-term planning conjures the image of an adult amid grown-up responsibilities, who has the capacity and foresight to do the best with what they have or can reasonably expect to encounter. In other periods of one's life, however, such a virtue seems misplaced. Here, Slote (1983, pp.46-47) offers the example of a child – whom I will call Fatima – who has formulated a life plan around becoming a surgeon. In this case, Fatima has identified a particular medical school, with a certain qualification in mind, leading to a specific career path. *But what are we to make of such a plan*? As Slote (1983, p.47) observes, "such predetermination is likely to seem suspect – though somewhat awesome – to most of us." This 'suspicion', as Slote phrases it, is really a genuine *concern*: we may worry about parental pressure, about closing the door to other opportunities, and so on. For the sceptics that Slote is gesturing towards – of which I will include myself – it is not that we think Fatima's plan is unsound (after all, we could even see ourselves pursuing it); instead, we worry that the formulation and the pursuit of the plan will affect otherwise important aspects of Fatima's development. In a word, we do not think that now is the time for planfulness.

Again, Slote's position finds some vindication within the psychology literature. For example, Dalton and Spiller's (2012) study points towards a tension between implementation intentions (i.e. specifically formulated localised intentions) and planning in a broader sense (i.e. higher-order forms of planning, broadly similar to the account presented in the previous chapter). As they write, "the data shows that commitment can be undermined by planning, a point that has been neglected in the literature on goal pursuit and planning" (Dalton and Spiller, 2012, p.12). The interpretative move that I am proposing here is that one can draw a parallel between this data and Slote's argument regarding virtues: just as prudence conflicts with openness, so too does planning undermine implementation intentions. In this way, so long as implementation intentions contribute towards an agent's formulation of their good – through, say, the fulfilment of more immediate desires and aims – then the idea that planning *impedes* agents' access to specific goods may also find empirical justification.[26]

[26] In terms of Slote's claim that children do not generally plan, I would add the caveat that the empirical literature on planning in childhood traditionally focused on areas of expertise that children *lacked*, i.e. novel problem-solving tasks (Friedman and Scholnick, 1997, pp.3-25). Instead, children can prove to be very planful when they encounter tasks with which they are already familiar (Friedman and Scholnick, 1997, p.7). However, as I pointed out in the previous section *vis-à-vis* the temporal structure of life-plans, a child will arguably have no familiarity with life-planning.

In broad terms, then, the underlying claim of this critique is that we cannot – and should not – plan for everything. The virtues that we adopt are, more often than not, relative to the circumstances that we find ourselves in. For example, in terms of adulthood, we might worry about someone's overly prudential approach to finding love: they will miss out on those romance-movie-like moments when an unexpected series of events leads to meeting one's soul mate. Here, rational-life planning seems to preclude such happy endings precisely because they are *unexpected*. Taking up a prudential standpoint, one seeks to *minimise* happenstance; what is good for a person depends on its role within their plan, rather than a stroke of serendipity. Yet these context-sensitive virtues, and these goods that seem to evade expectation, seem just as important for characterising the good life as prudence and planfulness. In a slogan, rational life-planning overlooks what is so wonderful about life, precisely by closing us off to the wonder that we can only ever experience without planning.

4.2. Planning as a Situated Process

In mentioning the criticisms above, I have been highlighting the following theme: the idea that someone can plan for their entire life is misguided; at best, it is impractical or idealising; at worst, it is an abstraction. As I indicated in the introduction, my position on this discussion is that both horns of this critique against Rawls reduce to concerns regarding temporality. In short, for both Slote and Williams, the formulation and pursuit of a life plan requires planners to treat their commitments as *non-situated*. In Slote's argument, it means sticking to prudence when other virtues seem more appropriate; in William's critique, it means presupposing that one's preferences can be both fixed and relevant. In this way, Rawls is taken to overlook the basic proposition that "rational life-planfulness is a virtue with *a temporal aspect*" (Slote, 1983, p.47, emphasis added). For those critical of Rawls, there can be no definitive schema for approaching the major choices that emerge throughout one's life; if we assume such a schema exists, then we risk missing out on the plethora of experiences that life simply offers us by focusing only on those things that we can plan for.

That being said, I do not think that succeeds in undermining Rawls's position – who, I will soon demonstrate, has some ready-at-hand responses to boot. It is not that these arguments are without merit – for I agree that practical deliberation is a situated practice and that different attitudes allow us to remain open to different goods – but I think that there is enough evidence that *Theory* accounts for these concerns. Perhaps most significantly, Rawls can claim that both Slote and Williams (as well as others that follow this line of argument, like Mackie) fail to include the ways in which life-plans themselves are suspectable to change and reassessment, as he writes:

> *At any given time* rational persons decide between plans of action in view of their situation and beliefs, all in conjunction with their *present major desires* and the principles of rational choice… In time his choice will lead him to acquire a definite pattern of wants and aspirations (or the lack thereof), some aspects of which are peculiar to him while others are typical of his chosen occupation or way of life (Rawls, 1971, pp.415-416, emphasis added).

There are two crucial points in this quote alone, both of which respond to the criticisms developed in the previous two sections. First is that, at any given time, a person can reformulate their plan. Hence, it is simply untrue – as Williams contends – that Rawlsian life-plans presuppose a fixed perspective. Life-planning, for Rawls, is *processual rather than static*. Indeed, a person's life plan is global partly because they constantly refer back to their plan and to the activity of planning, not because they hold single-mindedly to a fixed set of preferences. Moreover, this is true both for the plan *in toto* and for individual subparts of the plan: whilst a person is generally expected to commit to their life-goals (say, becoming a surgeon) it is not ruled out that incremental changes will be made (e.g. studying in the U.K rather than the U.S.A) or that wholesale adaptions will take place (e.g. becoming a radiologist or focusing on physics rather than medicine). Rawls accepts, in this way, that the shape of one's 'life-space' is often indeterminate: there are no geometric axioms that provide a complete picture of one's life.

Second, and connected to this first point, a person's life plan is balanced against their present major desires.[27] In this way, whilst in the pursuit of my plan I will be acting within a pre-defined set of commitments, those commitments *will still be adjusted in light of my current goals and inclinations*. There is no 'currency' or 'third language' that adjudicates between these two weighty concerns; instead, "the details are filled in gradually as more information becomes available and our wants and needs are known with greater accuracy" (Rawls, 1971, p.410). Again, the mistake here is to think that life-plans are established somehow over and above the process of practical deliberation. Yet as I discussed in the previous chapter, higher-order plans are made up of evaluative hierarchisation and scheduling across sub-plans, meaning that the process of planning involves the temporally-embedded activity of relegating, rejecting, and affirming between different organisational substructures. In planning, I have not then set my destiny; instead, I have established "a hierarchy of plans, the more specific subplans being filled in at the appropriate time" (Rawls, 1971, p.410). For example, whilst a life-plan '*LP*'

[27] I will return to this issue, however, in section 4.3.

might be set at the earliest time T^1, this can involve deciding between subplans S^1 and S^2 at later time T^2, as well as a similar reassessment of *LP* at the later time T^3; including, but not limited to, the cessation of *LP* at T^3. As I understand it, an ongoing process like this is not at odds with Rawls's position. On the contrary, it seems to provide a broad overview of the procedure of life-planning.

Understood in this way, one of the main benefits of a life plan is that it serves to narrow down a person's horizons. In general terms, life is marked by a diverse range of possible activities, pursuits, and so on. Part of the function of life-planning is to organise this diversity into a more manageable hierarchical framework. Though a person's plan will not have a completely definitive structure, it will serve to condense and direct their other aims and purposes. This benefit of life-planning, and the more modest reading that it points towards (discussed in section 4.4), ought not to be overlooked. Rather, it suggests that planning incorporates a range of possible plans, and thereby allows the agent to make explicit to themselves their main interests and values. Having done so, potential courses for action are narrowed down, and the agent is capable of pursuing the things that are, in accordance with this plan, of the greatest importance to them. Conceived in this way, the charge of temporal abstraction not only seems unnecessary, but completely at odds with Rawls's initial intentions. With this in mind, I will continue my analysis to concentrate more directly on Slote's critique.

4.2.1. Including Non-Prudential Virtues

As I noted in section 4.1.1, the major concern of Slote's critique is that the benefit of planfulness is temporally restricted, whereas life-planning requires this virtue to be globalised across the entirety of a person's lifespan – leading it to mischaracterise the good life *ceteris paribus*. However, building on the above, it is clear that the absence of planfulness is also included within Rawls's (1971, p.413) notion of life-planning: "The limit decision to have no plan at all, to let things come as they may, is still theoretically a plan that may or may not be rational." The solution, then, should be relatively straightforward: deciding not to plan fits within the broader structure of a life-plan, and so there is plenty of room for practicing non-prudential virtues and receiving the unexpected goods that arise as a result. Whilst this accurately describes the position that I maintain, it is important to note that not all scholars are convinced by this immediate response. For example, Steinfath (2023, p.8) takes Rawls's position to be self-contradictory: "You know you cannot plan everything, but you then try to build this knowledge into your plans in such a way that the unexpected appears foreseeable all the same." Here, Steinfath focuses on planning as a "prospective" practice, i.e. the view that "planning my life…gives structure to my future and directs my activity in a rational manner that promotes my good" (Heyd and Miller, 2010, p.22). Precisely because the goods that this critique highlights are unexpected, Steinfath thinks that planning for them is tantamount to trying to *foresee the unforeseeable*.

However, Rawls's position is more plausible than Steinfath makes out. Rawls is not concerned with foreseeing the unforeseeable, but with the higher-order patterns (i.e. one's purposes and commitments over time) that emerge as one navigates the complexities of life. To begin unpacking this point, note that Steinfath's analysis overlooks – or, at a minimum, underplays – the fact that the decision to plan is itself contingent on a prudential assessment of the given situation.[28] For example, it might be that Fatima (example in 4.1.1) ought to abandon their plan to prioritise other aspects of their development; yet, even this seems to fit the process that Rawls describes, i.e. setting one's course of action through a deliberative assessment of one's circumstances, abilities, and desires. All of this, it should be emphasised, fits with Rawls's (1971, p.420) principle of postponement (recall, section 3.3), which he offers in connection to the account of deliberative rationality: "rational plans try to keep our hands free until we have a clear view of the relevant facts."

Hence, there is a plausible counterargument to Steinfath and Slote's criticisms: that the virtues that they emphasise – such as "a laid-back attitude of openness" – are often dependent on an inability to strongly identify with the future, e.g. I am not sure if *x* is best for me right now, so I will opt for *y*. Yet crucially, as Hershfield's (2011, pp.36-40) work suggests, this intertemporal identification is a *condition* for future planning – a condition which is in itself partly prudential – not antithetical to planning as such.[29] It is because Steinfath overlooks this point that he argues that 'planning for the unexpected' is paradoxical. However, all that is being suggested by Rawls is that the decision to plan now (for however long) can be put off; this is not the same as saying that 'unexpected goods' can be directly anticipated within the structure of the plan itself. To visualise this, recognise that a hierarchy which subdivides within and between organisational branches will have components that connect only transitively, *via* increasingly distant substructures, including those wherein prudence is absent. In this way, even if planning is context-dependent, one's life-plan will function as a regulatory system by

[28] This point is discussed again in section 5.2.

[29] As Hershfield's (2011, pp.30-43) study indicates, intertemporal decision-making relates to the agent's capacity to conceive of themselves in the future, i.e. it is dependent on *both* the person's current and future senses of self. In this way, given that once a person has a more vivid representation of their future existence, planning becomes both more likely and more effectively executed.

which one organises and pursues one's major aims. So long as this is a plausible function, one cannot reject life-planning because of the existence of localised forms of goods and virtues.

In this way, I am arguing against Steinfath's (2023, p.8) claim that life-planning "expresses a desire to control the future as far as possible." Instead, Rawls's position is closer to the view that, as our *capacity* and *desire* to control our circumstances increase, then we are more capable of formulating and sticking to a plan of life (additionally, not only are we more likely to take this course of action, but it is also prudential for us to do so). It might be that localised periods of one's life demand an increased sense of openness or that periods of great upheaval make it so there are meaningful epistemic boundaries to planning. Yet, Rawls accepts these possibilities because, at a minimum, his position is that prudence can be self-replacing, i.e. that it can be a prudential decision to no longer follow prudence. As Fried writes, who Rawls (1971, p.422, n.18) references directly for his account of life-planning and personal unity:

> The conception of a rational life plan has implications for the responses that persons having such a plan make to the facts of the human life cycle…A strategy of selective openness – keeping certain kinds of options open – implies a judgement about the various options at various stages in the life cycle. *But this is the kind of judgement of which a life plan consists* (Fried, 1935, pp.163-164, emphasis added).

In response to this part of the temporality critique, then, Rawls can emphasise the unifying function of RPLs as well as the fact that, more often than not, there is a sense in which a person's life plan remains open. After all, if unexpected goods are good because they are unexpected, then it matters little whether one has prospectively included them within one's life plan. They are goods that we encounter head-on in life, meaning that – either way – we can only ever include them within our conception of the good after the fact. As Fried puts it above, we come to different judgements about distinct parts of our lives, like molecules within a molar structure, such that even periods of openness can be expressed within the unity of a life plan. Hence, my position is that – contrary to the temporality critique – focusing exclusively on the global aspect of life-planning will not lead to much success when criticising Rawls. In the upcoming sections, however, I will put forward an alternative critique against Rawls's position on rationality, which has more success overall.

4.3. The Problem with Deliberative Rationality and its Objectivity

So far, I have considered one of the main lines of criticism against Rawls's account of life-planning: that it overlooks the basic situatedness of decision-making, leading Rawls to mischaracterise the good life by omitting non-planful virtues and unexpected goods. As I demonstrated above, however, this argument fails to make meaningful headway against Rawls; for one, it overlooks the many ways in which life plans are open to periodic reassessment, cessation, and adaption; for another, it mischaracterises the role of life-planning in determining a person's main purposes and narrowing down their horizons. With that being said, I now want to investigate whether Rawls's account of goodness commits to a form of evaluative abstraction; that is, whether the congruence argument assumes that a person's good is given by their life plan that *they would choose*, had they had access to all the relevant facts and beliefs. Throughout this and the following subsection, I will demonstrate how this issue generates significant problems for Rawls's approach to congruence. However, I will end, tentatively, on how *Theory* might fix them.[30]

To grasp this issue in Rawlsian terms, first recognise that there are two ways that *Theory*'s account of deliberative rationality might be cashed out. To this initial point, recall from section 3.3 that deliberative rationality specifies the principles of reason that persons should follow in formulating their respective plans of life. However, the kind and amount of information that is available to these persons will differ depending on the assumptions that are made about them, the result being two notably different versions of deliberative rationality. On the one hand, persons follow the principles of deliberative rationality in light of complete information and true beliefs; on the other, persons are limited to their current beliefs and the information immediately available to them. From this, note that – directly correlative to these distinct epistemological positions – there are two alternative characterisations of that person's good: their "real good", which is a life plan formulated with complete information/true beliefs, and their "apparent good", which is a life plan formulated with incomplete information/current beliefs (Rawls, 1971, p.406).

For the next step, recall – again, from chapter 3 – that Rawls utilises the notion of RPLs to substantiate his account of the self's unity. In the above section, I pointed towards this unifying effect when replying to Steinfath, wherein I posited – drawing on Fried – that judgements of all kinds can be included within a person's life plan when it is understood as the emergent unity of their major purposes and goals. For Rawls (1971, p.561, emphasis added), a

[30] It should be noted – as I explain in chapter 8 – that one of (later) Rawls's reasons for abandoning the congruence argument connects to this issue. However, that discussion pertains to the implicit metaphysical assumptions associated with *Theory*'s account of the good. In this thesis, I focus on the account of RPLs in its own right. It should also be noted that I take guidance from Freeman (2007) in formulating my argument here. Still, I offer my contribution by introducing Rawls's position on the self's unity into the discussion.

person is a unified self insofar as the intricacies of their life plan, at the highest level of conduct-organisation, string together with an overarching theme or sense of purposefulness: “Thus a moral person is a subject with ends he has chosen …*the unity of the person is manifest in the coherence of his plan*.” On the back of these observations, my point is this: Rawls cannot hold all three of these notions at once, i.e. subjectively rational life plans, objectively rational life plans, and his account of the unity of the self. As I will soon demonstrate, if a person is unified *via* the coherence of his life plan, then this unity is directly undermined by the Rawlsian notion of objective rationality.

Having grasped the tension here, recognise that rational life plans are, as Heyd and Miller (2010, p.30) point out, partly *constitutive* of who we are; “if the resources for defining the very criteria of life's value are themselves a matter of the individual's choice, then this choice becomes in a serious sense constitutive of the individual's identity.” This squares with the defence that I have been giving of Rawls so far: in formulating my life plan, I provide the grounds for the person that I will become – and as a result, I make a decision that partly defines who I am now. However, by including the notion of objectively rational life-plans, Rawls commits to the presupposition that the good can be settled *outside of its constitutive role* in comprising one’s sense of self and self-unity; here, Larmore makes an instructively similar point:

> His mistake is indeed a fundamental one, and he is far from being alone in falling prey to it. The error consists in supposing that the best way of living the life we have before us is determined in advance, before we have lived it…. [the fundamental mistake] does not concern the time-bound character of all deliberation so much as the time-bound and contingent character of our individual good (Larmore, 2010, pp.192-194).

As a result, the notion of self that the congruence argument is liable to depend upon is one that takes the good to be settled prior to – or, perhaps more strongly, in spite of – its realisation within a person’s life. In this way, so long as persons within the Rawlsian society are expected to identify with this idea of the good, then the account of the self is ultimately disjointed: it puts at odds a form of planning that is partly constitutive of who I am, with an account of planning that depends on who I could have been, had a rationalised set of conditions obtained.

This problem rules out the strongest version of the congruence argument. In this version, the Rawlsian sense of justice is *in fact* congruent with a person’s good, understood as the plan of life that they would pursue had they had true beliefs and complete information. As a result, this rendering would allow Rawls to exclude potentially illegitimate counter-responses against congruency by minimising the effect of agent fallibility. That is, if a person claims that justice is incompatible with their notion of a life well-lived, Rawls only needs to point them towards their 'real' rather than 'apparent' good. In doing so, he would assure this person that their sense of justice – when incorporated within a fully rational plan – would indeed be good for them; that is, part of their good *ceteris paribus*. However, according to the argument that I have just provided, this version of the congruence argument would come to undermine itself. In particular, since it conflicts with the role of life-planning in establishing a person's sense of self-unity, the objective-planning approach cannot fully account for that person's good. After all, injuries to self-esteem, institutions that lead to undermining the powers of the self, conceptions that require the disfigurement or bifurcation of the self, are all consistently ruled out by Rawls's (1971, p.440; 498; 554) approach more broadly. If the unity of the self is a good for persons, then objective rationality jeopardizes that fact by situating goodness outside of its originating context.

Understood in this way, Rawls's account of life-planning would indeed involve an act of abstraction. However, it would be more comprehensive than previously supposed by the temporality critique. In particular, the notion of rationality that Rawls imposes on life plans would abstract from the agent's current beliefs, frameworks of understanding, their immediate knowledge of the world, their sense of self-formulation, and the role that their life plan has within their sense of self-identity and self-unity. Hence, it is clear that what I have called the 'strongest' version of the congruence argument cannot get off the ground, particularly when contextualised in relation to my arguments in chapters 1 and 2 (e.g. generating attitudes, stress-test, and so on). With this in mind, I shall now consider an alternative response: that Rawls's account of objective rationality is meant to be a guiding standard for otherwise subjectively rational life plans. However, as I will demonstrate, this response also succumbs to many of the same pitfalls discussed here.

4.3.1. Optimisation and Self-Subversion

In the discussion above, I introduced a problem for Rawls's argument from congruence. In particular, I argued that *Theory* is ambiguous between two distinct accounts of the rationality of life-planning. On the one hand, there are objectively rational life plans, whilst, on the other, there are life plans that are subjectively rational. This ambiguity poses a significant problem for the congruence argument because *how we characterise the good* is half the battle in determining whether or not it is compatible with the pursuit of justice. I argued against the objective rendering of deliberative rationality as it brings disunity to the self – thus, paradoxically, furnishing an incomplete or inconsistent account of goodness. Yet this leaves a gap in Rawls's theory: how do we characterise the rationality of a person's life plan? The solution that I will consider now is that deliberative rationality expresses the weaker condition that one's life plan ought to be constructed using one's objective plan as a regulative standard; "When the question arises as to whether doing something accords with our good, the answer depends upon how well it fits the plan that would be chosen with deliberative rationality" (Rawls, 1971, p.421). Hence, a person's good is loosely defined through the life plan that they have constructed subjectively, having done the best that they can with the current information, characteristics, and so forth, that they possess.

My position is that this description of rationality still generates the problem of disunifying the self; only, in this case, it is understood on a more practical level. To see this, note that utilising the notion of objective rationality as a guiding standard maintains the implication that a person's good is set outside of its realisation within their lives, e.g. it is good *independent* of the agent's actual desires, beliefs, and so on. Specifically, the main change with this approach is that one should *aim* for standards that are disconnected from the constitutive role that life-planning has in one's life. Understood in this way, the demand of the congruence argument remains relatively the same. To accept congruency, persons still have to hold an ideal evaluative distance from themselves (i.e. from their actual desires, commitments, and so on). More broadly, my point is that if 'what is good for a person' is given through an objective standard, but their conception of the good is determined only subjectively, then it is difficult to see how justice as fairness is "suited to express…the nature of a single self" (Rawls, 1971, p.564). Instead, the Rawlsian account seems to imply a sense of self-conflict, wherein one's real good comes to stand like a judicious conscience over one's apparent good, seen here in Royce's description of life plans:

> So the plan or ideal of life comes to stand over against your actual life as a general authority by which each deed is to be tested, just as the judicial conscience of the judge on the bench tests each of his official acts by comparing it with his personal ideal of what a judge should be (Royce, 1908, p.176).[31]

The upshot of this is that the judicious mindset of objective-standard planning, rather than allowing one to formulate and self-authenticate one's good, can plant the seeds for "self-subversion" (Setiya, 2017, p.138). To see this, consider the implications of formulating one's

[31] Recall from chapter 3 that Royce is one of Rawls's main influences *vis-a-vis* the notion of RPLs.

good based on what one ought to achieve given idealised standards. In such instances, the unity provided by one's life plan is better explained as a process of *optimisation*, rather than of self-formation. One will be *living up to* a plan which most ideally expresses one's desires. Similarly, if a person's unity is given by the cogency of their life plan, then it is, in an important sense, exteriorised: a person is unified insofar as their ends hang together in a rational way rather than through their capacity to identify and pursue their purposes in life. Subsequently, the life-planner is perpetually disunified, always feeling like they are catching up to a life that they are yet to live. The practical consequence of this, moreover, is a sense of exhaustion, restless, and incompleteness that accompanies the life-planning approach to the good; in fact, Setiya's analysis of midlife provides a concrete example for this description:

> That is the problem with being consumed by plans, obsessed with getting things done…It is as if you are striving to eradicate meaning from your life, saved only by the fact that there is too much of it or that you keep on finding more…[As a result] my affliction is chronic, not acute, masked by the whirl of activity: more papers to grade, meetings to organize, books to read. It is not that I take no pleasure in going for a walk or spending time with friends…But the roots of meaning in my life are principally telic: they aim at terminal states…I am ruefully possessed by the telic mindset (Setiya, 2017, pp.134-138).

Here, my suggestion is that if a person is able to identify with their objectively rational life plan, then Rawls's approach does seem to usher in something like a 'telic mindset'. That is, persons may be required to pursue, affirm, or at least identify with a plan – given the standards imposed by deliberative rationality – that they do not feel like they have authored. Practically speaking, this may result in persons controverting their own good in the very attempt to optimise it, by comparing their immediate situation to an idealised version of their life. Accordingly, the worry is that it is not through the originating act of life-planning that a person is unified, but through the ends themselves – which, in an important sense, formulate a pre-defined hierarchy of long-term desires and commitments – that a person is meant to retain a meaningful and coherent sense of self-identity. Relatedly, Heyd and Miller (2010, pp.20; 30) – drawing from Nagel (1986, pp.214-223) – argue that this account of rational life-planning involves a form of self-enforced 'bootstrapping'. In particular, they highlight that the meaning of small-scale localised plans can always be provided by reference to broader intentions. However, since life plans are globalised structures, there is no appropriate reference point for verification of their significance or success. As a result, Rawls can be seen as quietly ushering objectivity into an ultimately subjective process:

> Thus, once the traditional conception of the good as anchored in the external order of things is abandoned in favor of an internally-based "life plan," the aim of giving it meaning involves some sort of bootstrapping, i.e., granting an objective meaning to a subjectively-based project. (Hyed and Miller, 2010, p.30)

Finally, note that the objective-standard approach does not, by itself, resolve conflicts regarding the congruence between justice and goodness. Recall from the previous discussion that this was an advantage of the 'strongest version' of the congruence argument: that there is an abstract version of a person's good that is, in fact, congruent with their sense of justice. However, on this objective-standard reading, the abstraction is self-imposed, pulled up by its own bootstraps. There is no way to determine a person's objectively rational life plan; disagreements are not resolved but *pushed back*. Similarly, it is not whether the agent's plan is rational, but whether it correctly characterises what they would have done, *had they been* fully rational. However, not only is it unclear as to whether this would be accepted by the person involved, but it is also unclear as to how this process would be executed. If my life plan is partly constitutive of who I am, then I have to identify with a set of desires, commitments, responsibilities, aims, beliefs, of which I am *not*. The rationality of my disagreement, as a result, would turn on whether or not I could ever identify with this idealised version of my life. This is wrongheaded: the issue at play is whether justice can be part of *my* good, not an abstracted version of my life plan. In the next section, I will offer a possible Rawlsian solution to this problem.

4.4. The Priority of the Self and the Practicalities of Life-Planning

To take stock of this chapter so far, I have considered and dismissed what I called the 'temporality critique' of Rawlsian life plans. Within *Theory*, there is a plausible account of the diachronic unity of life plans, which allows for the existence and use of non-prudential goods and virtues, as well as accepts the basic fact that our decisions are necessarily situated across time. In light of this, I put forward a different – but, in a sense, related – critique of Rawls, which calls into question the relationship between life plans and the Rawlsian account of the unity of the self. This critique highlighted two unpalatable consequences for the Rawlsian account of rationality. First, the objective-standard approach to life-planning sows the seeds for self-subversion and undermines the agent's sense of self-authorship. Second, the notion of objective rationality is at odds with the constitutive role of life-planning in partly defining the agent's sense of self-identity and self-unity. I also clarified what is at stake: accurately characterising the account of the good that will be employed within the congruence argument, thus determining whether or not justice as fairness will be a sufficiently stable conception.

A worry that emerges from these discussions is that a person's life plan – and thereby their conception of the good used within the congruence argument – can be formless or even irrational. I doubt Rawlsians would accept this formulation, as the entire account of congruency would break down. To see this, recall the analysis of section 3.3, where I highlighted the distinction between a plan (which seems to be constituted by a degree of rationality) and a hare-brained scheme. As a result, if a person's plan is rendered entirely structureless, then there is also no means for determining whether – on this Rawlsian account – something is actually good for that person. Similarly, it would seem that the good would have no consistent definition across ordinary language use, seen here through the close connection Rawls draws between rationality and goodness:

> Thus, for example, when we are asked for advice someone wishes to have our opinion as to which course of action, say, is best for him. He wants to know what we think is rational for him to do…The meaning of "good" and of related expressions does not change in those statements that are counted as advisory (Rawls, 1971, p.406).

That being said, I do not think that Rawlsian life plans have to be irrational just because the account of objective rationality is at odds with Rawls's broader project. One can conceive of life plans as rational, without maintaining that they must aim at some objectified or 'best' plan. Instead, there is a degree of rationality that is given in the plan's very formation, by which *structures internal to the hierarchy* can be used to evaluate, offset, and justify one another. Similarly, the process of formulating a life plan is also, in a sense, procedurally rational; that is, one life-plans *via* an act of deliberation, taking the time to consider and scrutinise one's major aims and purposes. Consistent with my reply to the temporality critique, then, rational life-planning is an ongoing, deliberative, process, whereby – within temporally embedded contexts – one seeks to balance one's commitments against each other, to get to grips with one's purposes and narrow down one's horizons for activity.

To unpack this as a plausible rendering of Rawls's original account, the first thing to note is that the congruence argument will not succeed for everyone; "it can even happen that there are many who do not find a sense of justice for their good" (Rawls, 1971, p.576). As I have highlighted throughout this discussion, Rawls does not require the strongest version of the congruence argument. Instead, as per chapter 2, congruency stands as a form of stress-test; with the crucial point being that, given the scope of justice, its relationship with the good is rendered relatively consistent, in a way that does not bring the other into ruin. Moreover, the issue of congruence can still be used to compare different conceptions of justice without this strongest version (Rawls, 1971, p.576). To see this, recall from chapter 1 the comparison with the Hobbesian conception of justice, wherein stability in the well-ordered society is ensured – not just by a sovereign-backed system of punitive measures (e.g. Hobbes) – but also by the attitudes that it fosters amongst its citizens. Understood in this light, a person's good does not need to be perfectly rational; in fact, the shortcomings and nuances of day-to-day life seem to be a necessary factor to the congruence argument overall.

The second thing to note is that, in my response to the temporality critique, I have been emphasising the processual nature of RPLs. A person's life plan does not need to be settled all at once and in advance. Instead, one's plan considers all of life's temporal dimensions, e.g. past commitments, future aims, present capabilities, and so on. In this way, emphasising the process of life-planning may reduce Hyed and Miller's concern that, since one's life plan is a globalised structure, then the task of verifying that plan and giving it meaning is exteriorised. Whilst Rawls's account of objective rationality commits to this approach, it is also clear that this is not an inescapable component of life-planning. Instead, there are definitive structures that one can refer to in order to scrutinise and give meaning to one's plan in its current and ongoing state. Put plainly, it consists of the rest of one's plan – that does not constitute an infrangible whole, but a hierarchy that becomes entrenched during a process of deliberation and enaction.

The final thing to note is that Rawls's account of the unity of the self, as opposed to encouraging a telic mindset, is *meant* to give self-authorship and self-authentication priority over a person's systems of ends. Indeed, Rawls (1971, p.560) expresses this by claiming that "*the self is prior to the ends which are affirmed by it*." Understood in this way, it is perhaps unsurprising that there is a tension between Rawls's (1971, p.561) account of the self's unity and the notion of deliberative rationality; the latter is a prescription for formulating one's 'real good', whilst the former consists in the fact that each person is free to "fashion [their]… own unity." In this way, the capacity to formulate one's life plan becomes crucial to the Rawlsian viewpoint overall. Above all, a person must feel as if they have constructed and self-authenticated their own conception of the good; otherwise, there is no guarantee that the pursuit of justice is not – as discussed in chapter 2 – the result of nefarious forces, incompatible with the powers of the self, and so on. Hence, the connection that I began to draw between self-unity, agency, and goodness in the previous chapter, and that I will continue to develop throughout this thesis.

In this way, I do not think that Raws's approach requires any substantive notion of 'best' to be able to determine the rationality of a person's life plan – in fact, I do not think this sits well with the Rawlsian account of life-planning that I have presented so far. Instead, the agent's current desires and aims function well enough to provide structure to one's plans. Specifically, there are normative implications to the commitments that we have made in the past and are currently maintaining, commitments which allow us to 'narrow down' potential courses of action. Moreover, whilst the metaphor of a perfectly moral arbiter may help the judge execute their duty (i.e. in Royce's interpretation mentioned above), it is only on the basis of their current commitments that this image takes shape and proves informative. Hence, in terms of practical rationality, we ought to *reverse* the Roycean approach: it is on the foundation of our current beliefs, subplans, desires, and so on, that the judicial conscience is rendered remotely intelligible – and it is only as a process of deliberation and execution that it remains this way.

My solution is that the process of deliberation itself ensures the rationality of a person's life plan. Accordingly, disputes between commitments/subplans can be resolved by analysing how one's life plan fits together, e.g. whether it is successful in establishing either a dominant theme or a definitive set of purposes. Rationality is an internal feature of the life plan itself. This does not mean that deliberative rationality – i.e. the account of rational choice that Rawls puts forward – has no role whatsoever. The crucial point is that it is significantly minimised: it is a heuristic that a person can follow and implement in their own decision-making. The principles that Rawls offers are applicable to a person's life plan without the notion of objective rationality (e.g. the principle of consistency, where plans should be scheduled in harmonious relations to one another). Instead, the operative part of Rawls's argument from congruence is that persons formulate their conception of the good within the process of life-planning, wherein this process consists of deliberating – across time and in a way amenable to revision – over one's major aims and purposes. In chapter 5, I shall reinforce this view by clarifying the connection between agency and self-unity from a Rawlsian perspective. Again, however, I shall argue that it fails on its own terms. In doing so, I demonstrate that Rawls ultimately mislocates the source of our agency.

Chapter 5

Agency, Self-unity, and Self-reproach: Sartrean Phenomenology and the Error of Rationalistic Voluntarism in Justice as Fairness

5. Introduction

In chapter 4, I dismissed the temporality critique of Rawlsian life-planning and instead advanced my own criticism: that objective rationality leads to the disunification of the self within justice as fairness. While I offered a tentative solution to this issue by providing a more modest reading of RPLs, in this chapter, I will specify exactly how life-planning connects to Rawls's account of the unity of the self. Specifically, I will demonstrate that the Rawlsian self is unified as an agent, and that its agential capacity consists in its ability to formulate and pursue a plan of life. For Rawls, this is a major advantage of his approach. Given the connection between self-unity, life-planning, and agency, the congruence argument demonstrates that one is not compromising one's status as a free and equal self by acting in accordance with justice. Instead, by giving priority to 'the right' (i.e. the Rawlsian sense of justice) one secures the conditions for free expression, *via* the serial ranking of the justice principles, whilst framing one's life plan within the more manageable confines of just conduct (Rawls, 1971, p.563). In other words, by affirming justice as part of one's good, persons are capable of unifying their own lives whilst accepting normative standards that specify the boundaries of social existence. The upshot of this, in line with chapter 1, is that justice as fairness successfully appeals to the status of persons as agents.

In this chapter, my strategy is to undercut the congruence argument by criticising its working parts, i.e. the identification of agency and self-unity with rational life-planning. In section 5.1, I highlight that Rawls assigns special significance to deliberative error because it can lead to the experience of self-reproach. Here, one should keep in mind the importance of rationally formulating one's life plan. As I showed in chapter 4, rationality is an internal feature of life-planning, whereby persons take the time to consider their main purposes and interests. Importantly, Rawls's description of self-reproach can be understood as the mechanism that ensures this process: if persons do not plan appropriately, then they are liable to self-reproach, and thus liable to misconceive their good. In fact, as I will spell out in section 5.1, one of the main functions of deliberative rationality (Rawls's principles from section 3.3) is that it secures against the undesirable consequence of self-reproach.

Building on this, in sections 5.2 and 5.3, I then argue that *Theory*'s treatment of self-reproach, and its strong dependence on RPLs, is indicative of Rawls's broader commitment to rationalistic voluntarism. In section 5.2, I anticipate a potential Rawlsian response: that *Theory* attributes no special value to deliberation. I argue that this response is illegitimate since it falsely exploits an ambiguity regarding Rawls's use of the term 'value'. As I demonstrate, Rawls views deliberation as valuable precisely because it constitutes one's agentic capacities. In section 5.3, I consolidate this reading and apply these observations directly to the example of self-reproach. In doing so, I highlight the precise ways that Rawls grounds agency within localised moments of deliberation, so as to tease out his commitment to rationalistic voluntarism.

In section 5.4, I argue that self-reproach cannot be explained through a localised conception of agency. Instead, understanding self-reproach means involving a range of normative phenomena that prove incompatible with a voluntaristic depiction of the agent. Here, I draw on Sartre's (2018) concept of a 'project' and the pre-reflective structures that form across them. In section 5.5, I develop this critique further by highlighting Rawls's view that those who have planned successfully, but nevertheless experience self-reproach, are being irrational. In response, I argue that this attribution of irrationality misrepresents both the function and the phenomenology of self-reproach. More specifically, by drawing on Bagnoli (2000), I suggest that Rawls's position only makes sense if we abstract self-reproach to a third-person perspective. However, given what Rawls is trying to achieve within his argument from stability (chapters 1 and 2), such an abstraction is untenable. Finally, I bring this chapter to a close by anticipating a Rawlsian counter: that Rawls invokes the notion of agency for limited theoretical purposes. Against this, I clarify that it is on the basis of *Theory*'s broader framework that Rawls's rationalistic voluntarism proves ill-founded. Using notions internal to justice as fairness, I will demonstrate throughout this discussion that mislocating the source of agency proves a crucial error within the congruence argument.

5.1. What it Means to be a Unified Self in Justice as Fairness

For Rawls, the primary function of RPLs is their capacity to capture each person's conception of the good. As I highlighted in section 3.4, Rawls then uses this point to contend that happiness can be defined through the successful formulation and execution of a plan of life. With this in mind, one can ask: are we meant to be *unhappy* when our plans do not go well or when we do not plan at all? Although not explicitly framed in this way, Rawls responds to this question by providing an alternative justification for defining goodness through rational life plans. In particular, it is not only that life plans define what it means to live a happy life, but that those who plan do not come to "self-reproach"; either in blaming themselves for their inadequacies, whatever they may be, or in resenting their plans for going awry (Rawls, 1971, p.422). In this way, Rawls contends that whilst planning does not *guarantee* happiness (since many uncontrollable factors contribute to whether our plans are successful or not), it does protect against unhappiness by avoiding feelings of regret:

> Acting with deliberative rationality can only insure that our conduct is above reproach, and that we are responsible to ourselves as one person over time… One who rejects equally the claims of his future self and the interests of others is not only irresponsible with respect to them but in regard to his own person as well. He does not see himself as one enduring individual (Rawls 1971, pp.442-423).

There is an interesting implication to Rawls's claim here, which I will tease out by two connected hypotheticals – both concern Ben, who wants to become a lawyer and so has to pass the bar. In the first hypothetical, Ben studies haphazardly without engaging in planning. Although he knows that he needs to pass the bar, he pursues this aim (if at all) without forethought or deliberation. In this case, Ben fails to pass the bar. In the second hypothetical, Ben studies in a thought-out, structured, manner. He schedules his time, he follows principles of reasoning, and he organises his commitments hierarchically. However, in this second case, Ben also fails to pass the bar. Part of Rawls's aim with self-reproach is to establish this: Ben has no reason to reproach himself for his actions in the second hypothetical, since has taken the time to set and deliberate over his options and aims, but he *does* have grounds to reproach himself in the first hypothetical. Thus, by drawing on the concept of self-reproach, justice as fairness allots special significance to the failure to appropriately plan one's life; thereby bolstering Rawls's argument from congruence, given that justice will be harmonious with *deliberatively formulated* life plans. As Williams puts it, although in a critical light:

> This model [life-plans structured in light of deliberative rationality] is presented not only as embodying the ideal fulfilment of a rational urge to harmonize all one's projects. It is also supposed to provide a special grounding for the idea *that a more fundamental form of regret is directed to deliberative error than to mere mistake* (Williams, 1981, pp.33-34, emphasis added).

Here, I accept Williams's claim that Rawls is concerned with a 'more fundamental form of regret' when it comes to life-planning. However, Williams fails to identify *why* the failure to plan appropriately results in particularly unpleasant feelings of regret - leading him to give less credit to Rawls's position overall. His interpretation is that negligence in planning is analogous to letting down a friend or loved one: "the grounding relies on an analogy with the responsibility to other persons" (Williams, 1981, p.34). Indeed, Williams (1981, p.34, n.8) extracts this understanding from one particular quote from Rawls (1971, p.423), which serves to motivate the importance of deliberative rationality: "we should indeed be surprised if someone said that he did not care about how he will view his present actions later any more than he cares about the affairs of other people (which is not much, let us suppose)." Whilst Rawls's phrasing here does support Williams's interpretation, I do not think that sticking exclusively to this explanation does service to the concept of self-reproach as it is presented in *Theory* overall. Instead, for Rawls, the regret associated with the failure to plan is not grounded on mere analogy, but the very real sense of *letting oneself down*.

Broadly, the Williams-esque interpretation of 'letting oneself down' is that we each have a duty to our future selves. In life-planning, I should do all I can to ensure success for the subsequent versions of myself and to honour the claims of the past version of myself. However, 'responsibility to my past self', for example, does not fully explain why I should feel regret *now* on behalf of who I was *then*. In fact, from a practical perspective, this would lead to a weaker sense (or even absence) of self-reproach: it easier to explain my lack of responsibility if my sense of self is, so to speak, *disjointed* between temporal states. More directly, my responsibility to myself cannot be analogous to my responsibility to others since the latter implies some form of separation; yet to see only *iterations* of myself is the exact mistake that Rawls (1971, pp.442-423) wants to avoid. Even though I will change from who I am now (whilst I plan) to who I am later (after the plan is executed), there must be something more fundamental that grounds this sense of self-reproach, despite the differences that obtain over time. Here, I am not only making a point against Williams. Instead, this analysis suggests a positive thesis: that self-reproach connects to Rawls's account of the unity of the self, since without a unified self-conception, persons would not have reason to reproach themselves in the first place – this is a point I shall return to throughout this discussion.

In contrast to Williams, I argue that Rawls's use of self-reproach is grounded in the following way: I will not want to resent my past decisions, since this could lead me to doubt my very *capacity to effectively formulate a conception of the good.* For Rawls (1971, p.422, emphasis added), nothing less than my ability to choose is at stake; "a rational person may regret his pursuing a subjectively rational plan, but *not because he thinks his choice is in any way open to criticism*." From this perspective, the distinction between 'deliberative error' and 'mere mistake', as Williams puts it, concerns the degree to which my sense of agency is at play. That is, having planned to the best of my ability – but having come up short anyway – I will say that my *plan* could have improved, that *the circumstances* for its execution could have been more favourable, but I will not feel the same sense of regret and responsibility that I would have if my plan (or lack thereof) were demonstrably irrational. Put differently, I will have reason doubt the product of my choice, but not my capacity to choose.

At the same time, in planning, I have attempted to *treat myself as one unified person*; "viewing himself as one continuing being over time, he can say that at each moment of his life he has done *what the balance of reasons required*, or at least permitted" (Rawls, 1971, p.422, emphasis added). Accordingly, Rawls's main focus is on a kind of *normative unity*: that I am unified in virtue of my agency, which sustains my capacity to value, to be responsible to myself, and to give reasons for my actions. Crucially, it is these features of my experience that are at stake when I fail to plan appropriately, the failure being that I set my horizons and purposes as if I were a different person. Hence, if 'deliberative error' is associated with a more fundamental form of regret than mere accident or mistake, it is because such errors indicate a disunity i*n regard to the agent's sense of self-authorship*; they have not taken seriously the reasons and desires that they have set themselves, or that they would have set themselves, had they formulated an appropriately consistent plan of life. Indeed, this interpretation fits with the broader understanding of *Theory* that I have been moving towards in this thesis; for Rawls, my life plan correlates to my levels of happiness, whether others will respect me, what kind of occupation I hold, what kind of person that I am, what activities that I enjoy, whether I feel ashamed of my actions, and likely more than this.[32]

Thus, the correct analysis of Rawls's position on self-reproach reverses Williams's original interpretation. It is when I think of myself as another person – or more precisely, when I fail to see myself as *the source of my actions and commitments across time* – that will lead me into feelings of self-reproach. Rawls is concerned, above all, with a kind of normative unity, which in turn depends on the person's capacities as an agent. Given the previous chapter, I take

[32] Evidence for all of these points in be found here in Rawls's (1971) *Theory*: Happiness (pp.548-554); Respect from others (p.441); Occupation (p.415); The kind of person that I am (p.416); What activities a person enjoys (p.426-427; 441); Shame (p.444). This is in conjunction with chapter 3's discussion.

it for granted that this normative unity is temporally extended – but note that this does not mean that a person acts in light of their future self's 'recriminations', only that they are motivated to act *as a unified agent*, rather a person whose commitments are susceptible to the whims of the moment or to a structureless life. Ultimately, Rawls maintains that when I take the time to deliberate over my major desires and interests, I have protected myself from the most undesirable of consequences: bringing into question my agential capacities and undermining the unity that I ought to be able to fashion from my own life. In this light, taking time to reflect on my options is crucial to characterising my good overall.

With this in mind, two general points come to the fore. First, the capacity to plan is tied directly to Rawls's account of practical agency. It is through life-planning that Rawls conceives of persons as purposive, self-forming, beings. Second, this capacity for agency similarly underpins *Theory*'s account of the unity of the self. When Rawls speaks of a unified person, he is concerned not only with temporal cohesion, but with a form of *reason-giving* unity. A person's plan of life is the schematisation of their major desires and aims (e.g. their reasons for their future endeavours), and so its formulation becomes a matter of treating oneself as one coherent person – or to put in Rawlsian terms, as the self-authenticating author of one's own life. Here, Rawls does draw direct inspiration from Royce (1908):

> Royce uses the notion of a plan to characterize the coherent, systematic purposes of the individual… [that which] makes him a conscious, unified moral person…And I shall do the same (Rawls, 1971, p.408, n.10).

In the next section, I will double down on my interpretation and argue that the concept of self-reproach is tied to a determinate, plan-related, decision. Moreover, I will defend my current analysis against a potential objection – that, contrary to my description here, Rawls does not allot special value to deliberation. These argumentative steps will then allow me to establish my main interpretative conclusion: that Rawls commits to a form of rationalistic voluntarism. Having established this interpretation, I will then criticise it throughout the remainder of this chapter.

5.2. Rawls's Account of Agency

To state my position now, when I say that Rawls endorses rationalistic voluntarism, I mean that justice as fairness takes persons' capacity for agency to require a conscious act of will, involving the primarily intellectual activity of positing, organising, and assessing their major aims and purposes. More technically, for Rawls, agency consists in higher-order, localised forms of reflection. At the same time, I hold that this description entails a very practical upshot:

whenever he makes an appeal to agency in his argument from stability (e.g. chapters 1 and 2), he does so on the assumption that persons are voluntaristic agents. Reassessing Rawls's account of agency means reassessing these arguments from stability, and *vice-versa*. Here, my claim is that Rawls locates the source of agency within our deliberative capacity to life-plan.[33]

Whilst I have supported this interpretation so far by responding to Williams, some Rawlsians may resist the charge of rationalistic voluntarism. To do so, defenders of *Theory* would make use of the following clarifications from Rawls:

> Goodness as rationality does not attribute any special value to the process of deciding. (Rawls, 1971, p.418)
>
> The efforts we should expend making decisions will depend like so much else on circumstances. Goodness as rationality leaves this question to the person and the contingencies of his situation (Rawls, 1971, p.424).

On this reading, *Theory*'s account of agency cannot privilege higher-order reflective acts (e.g. deliberation or planning), since Rawls states that such acts have no special value within his description of goodness. As these quotes indicate, active life-planning is but one option among many, where non-deliberative courses of action may prove just as apposite as deliberative ones (recall, to this end, my response to the temporality critique). As a result, my charge of voluntarism would fall flat on its heels, given that I am focusing on just one activity type (i.e. deliberative life-planning) among many – regardless of it being Rawls's primary focus. Consequently, without a plan-related emphasis on reflection, there would be no localised will that could stand as the source of Rawlsian voluntarism.

On a more holistic view of *Theory*, however, this response does not hold weight. To see this, note that the first quote above depends on an ambiguity concerning the term 'value', which can be resolved by further exegetical analysis. To be clear, Rawls's actual position is this: *deliberation has no special value in terms of 'goodness-for-the-person'* – meaning that a person can live an enjoyable life without planning for it (say, by living spontaneously) – nevertheless, *deliberation is especially valuable for other reasons,* e.g. avoiding self-reproach or expressing the self's unity across time. More precisely, when Rawls claims that his account of goodness 'does not attribute any special value to deciding', he is affirming the former proposition, i.e.

[33] To note, as Lindsay (1918, pp.433-455) and Dewey (1916, p.245) recount, this particular view of agency (i.e. voluntarism) has Roycean roots. Indirectly, given that I have emphasised the Roycean background to Rawls's position throughout the previous two chapters, this should support my interpretation here. Similarly, in chapter 3, I also explored the psychological underpinnings of Rawls's account and its emphasis on higher-order acts of consciousness – this too should be kept in mind.

that a person can be happy without planning. The crucial caveat, however, is that the spontaneous person realises their conception of the good through serendipity, or “good fortune”, without necessarily expressing their nature as a self-authenticating agent. As Rawls (1971, p.409; 549; 552) puts it, “by good luck he is not cast out of his fool’s paradise.” Since most of us are not so fortunate or can appreciate the errors involved in living ‘accidentally’, deliberative life-planning must still be employed when characterising the good.

Additionally, it is when the spontaneous person *fails* to live a good life that Rawls’s endorsement of the second proposition (i.e. that deliberation is valuable for other reasons) becomes clear. Here, Rawls (1971, pp.421-422, emphasis added) posits that “the value of the activity of deciding *is itself subject to rational appraisal*.” As a result, the decision not to plan is assessed in essentially rationalistic terms: “a person is being irrational if his unwillingness to think about what is the best (or a satisfactory) thing to do leads him into misadventures that on consideration he…would avoid.” (Rawls, 1971, p.424). Crucially, even if a person’s life can go well without planning, for Rawls, *the choice not to plan is still one made deliberately*, entailing that the agent is irrational if they fail to treat themselves as the source of their commitments *via* life-planning.[34] Hence, though a person’s conception of the good may not give any ‘value’ to deliberation, Rawls ultimately views *the value of that decision* – or that refusal, if you like – in rationalistic and voluntaristic terms. Understood in this way, even if a person does not give value to deliberation, their good nevertheless *depends* on it.

Establishing that Rawls views deliberation as valuable for reasons independent of a person’s particularised conception of the good does not *ipso facto* demonstrate that he is committed to rationalistic voluntarism. However, I have indicated that a form of traceability exists within Rawls’s framework, where acts of agency are drawn back to localised moments of deliberation (i.e. assessing the value of not planning in rationalistic terms and understanding that ‘choice’ through the lens of deliberation). This suggestion, moreover, is not without preliminary justification. By diminishing the achievements of the spontaneous person, Rawls also ties the notion of practical responsibility and continuity directly to life-planning. Consequently, the only decisions that a person needs to regret *qua* agent – that is, the ones that they are responsible for – are those relevant to their life plan. This connection comes out clearly in Rawls’s (1971, pp.421-422, n18) direct citation of Fried (1970, pp.158-169), who writes:

[34] This is all consistent with my reply to Slote. In my reading, non-planful virtues can be included within a life plan. However, I am highlighting the broader issue that this inclusion is contingent on a deliberative act of agency.

> This entails the possibly surprising conclusion that *there is no such thing as failure in a life plan*. For just as a life plan succeeds if it properly orders circumstances as they are, so also it succeeds if it properly… makes provision for uncertainty. To be sure, a person will regret that… uncertain outcome eventuates, but this regret is not in principle different from the regret that the life plan must be formulated under known and inevitable external constraints… True failure comes when one must recognise either that he has not been faithful to life plan, or that he has chosen what now seems the wrong life plan…[Here], the concept of responsibility for past and future are significant… (Fried, 1970, p.165, emphasis added).

Here, the implication is that it is *only through life-planning that a person's agency is put at stake*; "there is…no commitment to a life plan, but rather the life plan itself is a commitment to a certain course of action" (Fried, 1970, p.162). In the indented quote, this is suggested through an operative binary between contingency-related regret and agent-relative regret, or so-called 'true failure'. If this is to be taken seriously, then getting to grips with Rawls's account of agency means expounding the link between agency and self-reproach, i.e. the form of regret that a person experiences when they feel as if their capacity for agency is impugned. With this in mind, the final interpretative development in my argument is this: if Rawlsian self-reproach can be shown to depend on a determinate decision, and if this indicates a general theme of isolating and privileging agency to localised moments of deliberation within *Theory*, then the charge of rationalistic voluntarism is both well-founded and plausible; this is what I shall argue in the next section.

5.3. Examining the Rawlsian Account of Agency

To focus on the notion of self-reproach, I will use an example from Dillon's (2001) work, which will provide a consistent point of reference throughout this discussion. To clarify, my current aim is to apply the analysis of the previous two sections to a concrete description, thereby demonstrating that the Rawlsian account of self-reproach is indicative of a broader commitment to rationalistic voluntarism. An additional benefit of this example is that it can operate as a touchstone for critical engagement with Rawls's view on both agency and self-reproach, which shall be the focus of later sections. Hence, my overarching intention here is to provide a compelling description of Rawlsian self-reproach, which sheds light on its broader commitments and can stand as a point for critical discourse. Dillon's example reads as follows:

> Thirty years later, Alison still recalls an episode in her teens… always with something akin to self-loathing. There was this girl, Dana… [who] was nice and smart and funny, and she was deformed… That hadn't mattered to Alison… but it was a big deal to her high school friends. They… made fun of Dana in the halls. Alison never [joined in] …and she told her friends to

> stop it when they ridiculed Dana. But… she knows she was too cowardly and too needy of acceptance to stand up for Dana…and she knows that she did laugh. After all these years, Alison can't forgive herself for Dana… Recurrent self-reproach reminds Alison of things she wants not to forget (Dillon, 2001, p.53) [35]

It is important to recognise that *prima facie* this example can be captured, in a relatively convincing and articulate manner, by the Rawlsian account that I have described throughout the discussions above. Based on Dillon's description alone, we are pointed towards the essentially *episodic* nature of Alison's regret (e.g. momentary recollection; individuated events; recurrent, but not ongoing, experiences of self-reproach). It is only a small step to incorporate this into the full picture that Rawls advances: had Alison planned more effectively (or at all) – i.e. had she deliberated over her major desires, her abilities, and her future aims – she would have acted differently; electing instead to stand up for Dana, rather than be complicit in her abuse. As Care (1996, p.7; 14, emphasis added) writes, in a way reminiscent of this Rawlsian viewpoint, "problematic parts of one's past might be episodic", meaning that those who experience self-reproach "are stuck with something in their past that is, or appears…to be, a manifestation of 'self' that does not really fit *who or what they think they are*." [36]

With this contribution from Care, all of the core features of the Rawlsian account are present within this description of self-reproach. There is an episode, or a collection of episodes, that persists in Alison's memory as something regrettable. Descriptively, Alison's memory-retention is predicated on feelings of irresponsibility, e.g. feeling like she abandoned normative standards or values that she would otherwise uphold. More deeply, Alison's past actions are inconsistent with her sense of self-unity, and so the practice of 'not forgetting' such actions affirms – among other things – that she is capable of straying from her purposes and commitments. In this way, the experience of self-reproach is as much of a *reminder* for Alison as it is an affective response – a reminder to be more careful, to think of others more, and so on (Bagnoli, p.186). Hence, this description is indicative of Rawls's main point: that Alison has come to doubt her status as one unified self, *or as the author of her own commitments*.

Here, I am not suggesting that a lack of life-planning always leads to self-reproach. As I clarified in the previous section, things could have gone well for Alison whether or not she had planned (though she would have missed out on its other benefits). Instead, I am bringing

[35] As I have mentioned earlier, self-reproach within Rawls's framework is not principally a moral feeling; instead, it is a practical feeling orientated towards the achievement of one's ends. Hence, the moral nature of this example ought to be set aside.

[36] Source found in Dillon (2001, p.62, n.23).

to attention Rawls's underlying commitment that only when we fail to plan, do we have reason to doubt our agential capacities. As Rawls (1971, p.421) puts it, in a way reminiscent of Fried's main sentiment, "a rational person… does not regret… following the plan he has adopted." Correspondingly, this analysis of self-reproach exemplifies Rawls's position that deliberation is the privileged source of agency. From the Rawlsian viewpoint, Alison's failure to contemplate her aims and values through life-planning manifested in a disconnect between her actions and her sense of self-unity. The upshot of this is that Alison's failure is that she did not deliberate, or at a minimum, she did not stick to a life plan that is derived from deliberative acts. As Rawls writes:

> Thus we should say that *given our plan of life*, we tend to be ashamed of those defects in our person and failures in our actions that indicate a loss or absence of the excellences essential to our carrying out our more important associative aims (Rawls, 1971, p.444, emphasis added)

Rawls, then, endorses a form of reducibility wherein the exercise of agency boils down to the person's capacity to reflect on their aims and deliberate over their intentions. This commitment comes to the forefront when we accept that Rawls ties the experience of self-reproach – a feeling that directly impugns one's status as an agent and a unified self– to *a determinate, plan-related, decision*. Hence, Rawls's endorsement of a localised, higher-order, and rationalistic account of agency.

Given the preceding sections, not only have I given reason to attribute to Rawls a form of rationalistic voluntarism through exegetical analysis, but I have also provided a description of how this account operates *via* the notion of self-reproach. In particular, I have determined that Rawls considers decision-making processes in isolated terms, by reducing the agent's normative commitments to their decisions relative to a given life plan. Similarly, I have shown that *the execution of agency itself must stand up to a criterion of deliberative scrutiny*. Indeed, Rawls (1971, pp.414-415) construes decision-making as a function of "the higher-order desire to act upon…the principles of rational choice", wherein the advanced order of such commitments serve as "regulative ends that moves us to engage in rational deliberation and to follow its outcome." Even understood as a process that I described in the previous chapter, the non-static elements of Rawlsian life-planning are still viewed in terms of *periodic reassessment* – where the agent adapts their life plan *via* a deliberative form of retrospective-prospective analysis. As a result, whilst a person's life can go well in the absence of planning, it is only through life-planning that they can express their status as a responsible and unified agent, capable of forming and authenticating a conception of the good. It is this picture that I shall criticise throughout the remainder of this chapter.

5.4. Before and Beyond a Voluntaristic Conception of Agency

In this section, my main contention is that whilst self-reproach often appears to be directly associated with localised decision-making, further analysis reveals that it implicates our capacity to choose in a far broader sense, *in ways that cannot be explained by a voluntaristic approach to agency*. Drawing on the relevant phenomenology, I argue that the object of self-reproach is not so much a specific event or decision, but a constellation of normative phenomenon, all of which can be better understood by assuming that the subject experiences their sense of agency non-locally. Put differently, it is not clear that self-reproach has any definitive source when it comes to the experience of regret. However, *the existence of a definitive source is precisely what Rawls's approach commits to*, given that he ties our sense of agency (and the cause of self-reproach) directly to the decision to plan. Hence, whilst the temporality critique argued against life-planning on the grounds that our normative commitments cannot be globalised (recall, chapter 4), my argument is that agency cannot be localised. By analysing the notion of self-reproach more deeply, I aim to expose the flaws in the Rawlsian account of agenthood.[37]

One way to begin to grasp my argument is by recognising that it is not necessarily that Alison (in the example provided by Dillon) regrets her actions understood *episodically*, but that the culmination of these events reflects *undesirable patterns in Alison's character*. Here, I want to distance myself from Steinfath's (2023) work to minimise the potential for misinterpretation. In his paper, Steinfath (2023, pp.1-2, n.1) argues that Rawls's account of goodness (*vis-à-vis* rational life plans) ought to be replaced by the different "ways of being" that a person can adopt. To briefly explain, a person's 'way of being' is intended to be a form of *praxis*, which a person undertakes to express their conception of the good, e.g. "character-traits like virtues and vices" and the fulfilment of "social roles like being a mother or being a teacher" (Steinfath, 2023, pp.10-11). Broadly speaking, a person adopts a 'way of being' by habituating some established norm or role. Yet crucially, whilst this description appears similar to my statement regarding patterns of character, it is through the emphasis on social roles that Steinfath's account goes awry. Put directly, since 'ways of being' are defined by culturally-

[37] In this discussion, I draw on Sartre's account of pre-reflective agency, which I defend in chapter 7. For this current investigation, I remain general to make my claims intelligible without this background knowledge. Here, my only aim is to apply these observations to Rawls.

accepted norms and values, Steinfath omits an essential facet of Rawls's approach overall: grounding a person's good within their agential capacities. In doing so, Steinfath overlooks the fact that agency enables us to speak about character at all; *agency is a condition for character*, and this is a point that requires close attention.

In contrast to Steinfath*,* my argument is that we cannot meaningfully individuate specific decisions that comprise Alison's self-reproach, entailing that her experience ultimately resists a voluntaristic explanation. Instead, we can only make sense of this form of regret by considering a range of normative phenomena, including (but not necessarily limited to) the projects that Alison pursued before and alongside the actions directly implicated in her self-reproach. Even if we accept that Alison's self-reproach refers to undesirable patterns in her character, we ought to be especially careful about how we understand this claim. It is not that individuated sets of acts constitute the realisation of a particular character-trait, or that agency plays no role in this process. Quite the contrary, to understand how Alison experiences a break in her sense of self-unity, we must also understand her agentic capacities – and by doing so, it becomes clear that they are not reducible to localised reflective acts. Thus, I agree with Webber's summary of Sartre:

> An individual's character… consists in the projects that individual has freely chosen to pursue. *This does not mean that each character trait is itself a project*, as if someone could only be jealous or cowardly by aiming to be so, but rather that the distinctive patterns in the ways in which that individual sees, thinks about, feels about, and behaves in the world *are due to the total set of projects that they are pursuing* and need not be pursuing, whether or not they acknowledge these projects to themselves (Webber, 2009, p.4, emphasis added).

As Webber explains further, this point can be grasped by drawing a distinction between "determinables" and "determinations" correspondent to an agent's projects:

> Just as colour is a determinable of which red is one among many possible determinations, and red is in turn a determinable of which scarlet is one among many determinations, so particular projects are determinables whose determinations can be either actions or other projects (Webber, 2009, p.52).

According to this analysis, whilst Alison may reproach herself for not taking the time to contemplate her regulative aims, this is situated (for example) within the broader project of wanting to be popular at school, which itself is a determinable of an even broader determination. Moreover, regarding the agentic aspects of this description, Alison could have found different 'determinations' to realise her project of school-wide popularity; being mean to Dana was just one possibility among a structured hierarchy of available projects. The upshot of this is that to understand Alison's self-reproach, we must consider the pattern that persists across the totality of her projects – not identify reflective structures that are isolated, episodic, and de-contextualised. [38]

In recognising this, the cracks in Rawls's account of agency and self-reproach begin to show. The issue is that Rawls's treatment of self-reproach is incomplete, dependent on conditions that *Theory* does not have the tools to explicate properly. In terms of the origin of self-reproach, we cannot chalk this up to a singular decision (as Rawls's account suggests that we can) since to do so would be to misrepresent the proper order of things: it is to explain the whole through its parts, and *it is to view such parts as independently significant*. Yet it is only on the background of the agent's other projects that the lack of a plan is rendered remotely intelligible. Alison's failure to deliberate is not the cause of her self-reproach, any more than her actual actions in light of Dana's abuse. Instead, there is a synthetic unity that a description of these choices presupposes; as Sartre (2018, p.604, emphasis added) writes, regarding his account of agency, "we are not at all dealing here with a deliberate choice. And that is not because it is less conscious or less explicit than a process of deliberation but on the contrary because *it is the foundation of any deliberation*." To claim, as Fried and Rawls do, that 'there is no such thing as failure in a life plan' is to abandon this presupposition; it is to attempt to cleave apart choices made through reflection from the rest of the world.

At this point, a Rawlsian might reply that the description I have provided is strikingly similar to the idea of a life plan: there is an organised hierarchy of increasingly broad structures or projects, which possess overarching themes across the entirety of a person's life. However, I have two replies to this point, both of which develop my analysis more deeply. First, the Sartrean approach includes a form of agency that is incompatible with rationalistic voluntarism. In particular, it is an account of agency that is fundamentally non-localised, meaning *there is no determinate decision with which we ought to associate the agent's capacity for free choice*. In this way, to associate Sartre's position with life-planning is to misconstrue the issue entirely: it is not only that Sartre's analysis 'cannot fit' Rawls's account of deliberative rationality, but that it is practically and conceptually *distinct* from the higher-order processes that make up a persons' life plan in the first place. Second, and in the same vein as the first point, the structures that Sartre describes are always 'in question' – in our current discussion, this means that we

[38] A defence of this Sartrean account is given in chapter 7. For now, I apply these observations to Rawls.

cannot explain Alison's self-reproach through a bilateral reference between her life plan and her past actions. Instead, Alison's experience can only be 'explained' when we take her choice to be *ongoing* in a sense that precedes her capacity to reflect on her rational purposes. As Sartre writes:

> I alone, in fact, am able to decide at each moment on the impact of the past, not by debating, deliberating and evaluating in each case the importance of such and such an earlier event, but, by projecting myself towards my goals, I rescue the past as well as myself and I decide on its meaning through my action (Sartre, 2018, p.649).

In the next section, I shall say more about this second point (thereby revealing the general thrust of the first) to conclude against Rawls's commitment to rationalistic voluntarism.

5.5. Rationality, Practicality, and Self-Reproach

To see that the experience of self-reproach depends on a sense of agency that is non-localised and ongoing, recognise that even Alison's experience of self-reproach *now* – that is, her affective state when she undertakes a reflexive position on her past actions – *retains its significance on the basis of her current projects.* As Sartre (2018, p.717, original emphasis) puts it, "we ought…to note that this permanence of the past… and of one's character are not qualities that are *given*; they show up on things only in correlation with the continuity of my project." Moreover, this process of continuous renewal is behind Sartre's (2018, p.716) claim that, for the agent, "there is no character – there is only a project of oneself." Here, Sartre is not suggesting that there are no 'patterns of behaviour' (patterns which, from the outside, one might associate with particular virtues or vices), only that there is no 'character' over and above the projects that the agent pursues. Hence, the crucial point: whilst the source of Alison's regret appears to be situated within a decision taken in the past, the truth is that it consists in a decision that is – so to speak – still being made.

In light of Sartre's analysis, my argument is that identifying self-reproach with a voluntaristic picture of agency overlooks a vital aspect of the normative experience; descriptively, it is to fail to see that, just as Alison can continue to regret her actions, she can also practice self-forgiveness. In broadly Sartrean terms, it is Alison's way of *committing herself within the world* that enables her to view her past actions as reprehensible. There is no singular decision that captures this but an ongoing one; which requires constant renewal through further choices and precedes the agent's capacity to momentarily reflect on their purposes. Alison's experience of self-reproach depends on a form of agency that extends beyond the sporadic and transient nature of her deliberations and recollections; even though such acts may prove pivotal to the agent's reconstruction of the event. As Bagnoli (2000, p.178, emphasis added) writes, "reference to the agent's decision and deliberative process does not confine the notion of agency to voluntary agency."[39] In this way, we ought not conflate the act of making self-reproach explicit, with the experience of self-reproach *tout court*.

Another way of appreciating this point is by recognising that, on the Rawlsian account, a person who experiences self-reproach, despite having planned sufficiently, is irrational. As I have mentioned, Rawls (1971, p.422) – following Fried's insistence that 'true failure' can only appear correlative to an unsuccessful plan – maintains that the act of life-planning ensures that "there is no cause for self-reproach." However, I am not convinced that this is the case, as my example below – which takes inspiration from Bagnoli (2018, p.170) – serves to demonstrate:

> Consider Maeve, who must choose between a career as a musician or a literary critic. From her perspective, both options are on completely equal footing, though they are incompatible. Since it is ultimately up to Maeve to decide between the two, she elects to pursue a career in literary criticism (that is, after due consideration and forethought). After some years, Maeve develops a well-founded and well-recognised reputation for insightful literary analysis. One night, however, she sits down to write another paper. As usual, the words spill out with a natural cadence, only this time they feel empty. The feeling of self-reproach that Maeve experiences is both spontaneous and forceful, its apparent concern is that she chose not to pursue a career in music.

In this example, it is clear that Maeve regrets her choice of career deeply, to the point where her sense of agency is impugned, and she experiences self-reproach. However, it is also clear that Maeve could not have done any better, given her life plan. After all, she faced two equally good options, taking ample time to deliberate between them. I have shown, moreover, that this deliberative process is the main requirement for the rationality of a life plan (i.e. section 4.4). For Rawls (1971, p.422), then, there is no grounds for self-reproach, and more strongly, Maeve is confused in thinking that there is: "there was no way of knowing which was the best or even a better plan… [the planner] does not regret…[their] choice."

Here, I argue that Maeve's response is not irrational because the source of her regret is not exhausted by any deliberative process. I mean this in two interconnected ways. First, the

[39] To note, Williams's (1981, pp.27-30) use of "non-voluntary agency" is just as misleading, given that it conjures the imagine of a reflexive response rather than a pre-reflective commitment – "the agent does not 'decide' to regret." In contrast, Sartre maintains that the occurrence of regret is still an act of agency, without invoking the voluntary and rationalistic picture that Rawls otherwise endorses.

experience of regret can be spontaneous and evaluative, without bearing the mark of rational contemplation. For example, I must choose between two ice cream flavours: strawberry or chocolate chip. Say I chose the strawberry option. In this instance, I realise my regret through the very act of eating the ice cream; it will taste slightly bitter, less sweet, and less satiating. Importantly, this sense of regret was not the outcome of deliberation, nor does it need to be made explicit by way of reflection; instead, it is part of a more basic orientation, i.e. my way of responding to salient aspects of my situation. The same reasoning applies *mutatis mutandis* to Maeve's case. It is through her writing that she spontaneously realises an evaluative, but not deliberative, attitude of regret. The sentences that she formulates appear analytic rather than lyrical, constraining rather than creative. It is through a practical, "pre-judicative", comprehension of her situation that Maeve commits to her affective state (Sartre, 2018, p79; 186). The upshot is that the experience of self-reproach does not need to be located in, or even disclosed through, a deliberative act of consciousness.

This leads me to the second point: cashing out Maeve's response as irrational misrepresents the function and the phenomenology of self-reproach. On the Rawlsian account, the normative force of self-reproach depends on the person's life plan and arises because they failed to take seriously their status as an agent through rational deliberation. However, in Maeve's case, both of these factors have been neutralised; her life plan is going well, and there were no apparent flaws in her reasoning. Hence, *from a third-person perspective*, she either has nothing to regret or the object of her regret, properly construed, is the contingencies of her situation. Yet imposing an outside perspective ultimately distorts Maeve's experience of self-reproach and its significance *for her*; as Bagnoli points out, in relation Tolstoy's Anna Karenina:

> My point is that the exclusive attention to the question whether Anna Karenina's *life plan turns out to be successful or not* has the undesirable consequence of making her life regrettable for *anybody acquainted with its outcome*. If 'the situation' is to be blamed, how does Anna Karenina's regret differ from the regret… expressed by a sympathetic writer…? What distinguishes her attitude when she reconsiders her life as her own from the moment when she contemplates it as a spectator? (Bagnoli, 2000, pp.176-177, emphasis added)

My claim is that Rawls's life-planning account of self-reproach is equivalent to imposing a third-person perspective on an essentially first-person experience. By founding self-reproach on the person's plan of life, *it is discrepancy between the plan and the event* that determines whether the agent's response is apposite. Yet, in light of my first point, this interpretation of self-reproach completely overlooks its significance "for-the-agent" (Bagnoli, 2000, p.175). The Rawlsian view conceives of Maeve's regret as a kind of deliberative recalibration, rather than a "*practical capacity*", which is *as much of an extension of her current projects* as it is an evaluation of her past (Bagnoli, 2000, p.181, original emphasis).

Moreover, given that I have established Rawls's commitment to rationalistic voluntarism, this connection between life-planning and third-person abstraction should not be surprising. Once we conceive of the deliberative will as the privileged expression of agency, then we encounter an unpalatable binary: "human-reality appears…as a free power that is besieged by a collection of determined processes" (Sartre, 2018, p.580). Indeed, this is why Rawls's use of self-reproach makes it an appropriate case study, given its implicit assumption that "we can distinguish between acts that are entirely free, determined processes over which the free will has power, and processes that necessarily escape the human-will" – or, in the way that I have been exploring this issue, regret *tout court* and true failure (Sartre, 2018, p.580). In this vein, Rawls misinterprets the experience and function of self-reproach by grounding agency in an essentially localised deliberative act, meaning that once this act has been committed, the *significance* of self-reproach is determined by factors necessarily exterior to the agent.

As a final point on this matter, Rawlsians may put forward a more general response: Rawls is *not* attempting to provide a complete account of agency since he requires a description of agency only insofar as it comprises a functional part of *Theory*'s argument from stability. As I have established so far, this 'functional part' consists of a collection of interconnected roles, e.g. accounting for the unity of the self, characterising persons' conceptions of the good, describing a happy life, and so on. Here, I have two main points in response. First, I am not arguing that Rawls operates under an incomplete account of agency *as such*, but that *it is incomplete relative to its intended role(s) within justice as fairness*. The notion of self-reproach demonstrates this directly, given that it stands as one of the main motivations for citizens within the well-ordered society to formulate their life plans, and thereby to recognise and accept the role that the principles of justice have within their conception of the good. However, if my analysis in this chapter is correct, self-reproach is not a function of the person's life plan, but instead arises as a part of a more fundamental sense of agency. Practically speaking, we have reason to question whether life-planning really does ensure that our conduct is beyond reproach or that a conception of the good without the existence of self-reproach is even desirable. I do not aim to settle such questions here, only demonstrate that – *given the conceptual apparatus* *of justice as fair*ness – Rawls's understanding of life-planning and agency proves importantly incomplete and inadequate (a point that I extend to the next chapter).

Second, as I argued in the first chapter, Rawls's overarching argument from stability makes a direct appeal to agency. Rawlsians cannot dismiss the concerns I have raised here on the grounds that justice as fairness invokes the concept of agency in a limited sense. Instead, it is an urgent task to say *how* Rawls defines persons as self-forming beings. At this point, it should be clear that Rawls mislocates the source of agency in the capacity to formulate a life plan. I have supported this through several broad points; first, by identifying the connection between life-planning, agency, and the self's unity in justice as fairness; second, by demonstrating that self-reproach depends on a determinate decision and that life-planning is the primary expression of personal responsibility in the Rawlsian framework; third, by attributing to Rawls a commitment to rationalistic voluntarism. In this chapter, I have demonstrated why – again, given *Theory*'s conceptual framework – this background of voluntarism is both unpalatable and ill-founded. Self-reproach has been a relevant case study for this point, given that it resists explanation through a localised account of agency, even though that is precisely what the Rawlsian account takes for granted. In the next chapter, I shall develop these points further and complete the critical portion of this thesis. In doing so, I will argue that Rawls's emphasis on life-planning leads him to mischaracterise the good.

Chapter 6

Reflection, Self-Conceptions, and Normativity: The Illusion of Goodness in the Process of Life-Planning

6. Introduction

In this thesis, I have taken two critical steps against Rawls's philosophy. First, I argued against *Theory*'s use of objective rationality by demonstrating that it leads to a disunified account of the self. Second, I argued that Rawls's endorsement of life-planning ultimately mislocates the source of our agency by committing to a form of rationalistic voluntarism. In this chapter, I argue against Rawls's claim that a person's good is determined by their rational plan of life. By drawing on the work of Sartre, I provide a structural analysis of pre-reflective and reflective forms of awareness in order to assess the normative and epistemic connotations of life-planning. In doing so, I argue that Rawls mischaracterises goodness, establishing two main points to support this conclusion. First, life-planning is *an extension of what the agent already takes to be good* and so cannot characterise it. Second, this extension takes the form of *a reflective self-conception*, which is both inexhaustive and liable to misrepresent the subject it aims to encapsulate. The upshot of this, and the previous chapters, is that *Theory*'s account of life-planning fails to fulfil the three major roles that Rawls assigns to it, i.e. bringing unity to the self, capturing the source of our agency, and characterising a person's good. Insofar as the congruence argument depends on these features, and I have shown throughout this thesis that it does, it fails to secure the stability of the well-ordered society.

This chapter is built around a premise-conclusion style of argument. In section 6.1, I identify two basic assumptions that constitute the bedrock of my argument (i.e. premise one). In particular, that pre-reflectivity is a condition for reflective awareness and involves an essential normative element. Whilst I shall make these claims more specific as I go on, and whilst I provide a comprehensive account of pre-reflectivity in the following chapter, section 6.1 gives a general overview of these notions and lays out the theoretical claims that my argument builds upon. In section 6.2, I establish the second premise of my argument: pre-reflectivity anticipates and frames the content of reflection. Through Sartre's work, I show that without a pre-reflective form of awareness, reflective forms of consciousness lapse into an infinite regress. More directly, by considering the phenomenology of reflection, I argue that we cannot take reflection to introduce self-constitutively novel content into consciousness. Instead, the structure of reflective awareness is essentially comparative, meaning that it assumes – at least as a condition of its occurrence – a pre-reflective framework of commitments.

In section 6.3, I anticipate a potential Rawlsian response to my argument thus far: that life-planning *just is* a form of representation, meaning that a focus on life-planning is still capable of accurately characterising a person's conception of the good. In response, I establish my third premise, which holds that reflection non-trivially adapts the normative content of the pre-reflective experience. I support this premise through two sub-arguments. First, I highlight the fact that the agent is always pre-reflectively aware, and that consciousness always maintains a normative element. The upshot of this is that reflective acts are not neutral, not a way of making the innerworkings of the agent explicit – but are themselves extensions of the agent's projects. Hence, when one reflects, one does so *with a purpose*. Insofar as this is the case, reflection adapts the content it claims to represent by continuing the projects the agent is already pursuing. Second, I argue that reflection enables three errors of introspection: epistemic uncertainty, normative-downplaying, and the objectification of the self. Here, I examine Sartre's claim that reflection posits links between states/events without explicating the processes that connect them. As I demonstrate, this omission arises because reflective acts misrepresent what Sartre calls the 'datum of freedom' – broadly, the possibilities that shape our awareness of the world – primarily by misrepresenting the nature of our subjectivity.

In section 6.4, I summarise my argument and identify two concluding points that arise as a corollary. First, that goodness cannot be defined in terms of a life plan, since the activity of life-planning (and the purposes that are identified thereof) is an extension of what the agent already takes to be good. Second, that Rawls mischaracterises goodness and brings disunity to the self, since active life-planning distorts commitments that otherwise comprise the agent's conception of the good. In this section, I focus on the first concluding point. I suggest that Rawls attempts to account for the interconnectedness of our projects by generalising the process of planning across an entire life. However, the argument I provide cannot be spelled out in temporal terms but instead reveals something novel about the nature of the good: it presupposes a pre-reflective form of agency. I draw on the work of Larmore (1999; 2010) to argue that life-planning involves an extension to the agent's normative commitments *via* the subject's self-conception. Hence, Rawlsian life-planning is itself a localised and particularised phenomenon, meaning it cannot define a person's good overall.

In section 6.5, I complete my argument by demonstrating that life-planning not only misrepresents the agent's commitments but also the normative nature of the agent themselves. I argue that the self-conception that life-planning takes for granted encourages objectification, meaning that it ascribes concrete characteristics to a subject that necessarily alludes such attributes. Subsequently, I take issue with Rawls's account of the unity of the self. Here, Rawls's position is embedded in the idea of 'becoming one with oneself' – which, as I demonstrate with Larmore (2010), is an intention *disrupted* by the contemplative stance that life-planning imposes. Hence, Rawls's congruence argument – central to his account of stability and generating attitudes of support for his conception of justice – fails to get of the ground. Ultimately, I conclude that a person's good cannot be defined by their rational plan of life, for such an account is limited by the structures of experience it depends upon.

6.1. Primacy, Normativity, and Forms of Self-Awareness

So far in this thesis, I have relied on a general distinction between reflective awareness (which, in chapter 5, I associated with rationalistic voluntarism) and pre-reflective awareness (which I associated with a Sartrean account of agency). I must now provide a more direct overview of these ideas to establish two assumptions that will feature significantly in my argument overall.[40] As I will maintain, reflection denotes *an act of consciousness consisting in an explicit form of self-awareness*, whereas pre-reflective consciousness delineates an *immediate, implicit, and ongoing form of self-awareness*.[41] This alone is not very concrete. To resolve this issue, consider some of the main explanatory devices available in the relevant literature. For example, Larmore (2010, p.24) puts forward a definition of reflection as "thought turning back upon itself", which gives an intuitive brush over this notion. It is common, moreover, to find a distinction between first-personal (pre-reflective) and quasi-third-personal (reflective) forms of consciousness drawn alongside these lines – e.g. Eshelman (2016, pp.176-208). More technically, Rowlands (2016, pp.101-120) explains this distinction in terms of *the order of conscious awareness*. In his reading, pre-reflective consciousness consists in an "adverbial modification" of first-order awareness, whereas reflection consists of a second-order positional act of consciousness. Whilst the distinction between pre-reflective and reflective consciousness has also been explained by a juxtaposition between subject and object: pre-reflectivity is

[40] In the next chapter, drawing on Sartre and Ratcliffe, I will argue that a pre-reflective sense of agency is a structural feature of consciousness, further contributing to the ideas that I present here.

[41] Sartre (2018, p.12) uses parentheses to emphasise the non-positionality of pre-reflective modes of consciousness and italics to emphasise positional forms of awareness, e.g. consciousness (of) self and consciousness *of* self. I have not followed Sartre's use here, as sufficient context should be provided to distinguish reflective and pre-reflective forms of awareness.

identifiable with the subject of consciousness, whilst reflection involves the subject's attempt to turn themselves (or some part of themselves) into an intentional object for consciousness – e.g. Legrand (2007).

Regardless of any putative differences between these readings, a shared theme persists throughout. That is, *pre-reflectivity is an ongoing form of awareness that is a condition for reflective consciousness*. Whether this is explained in terms of the perspective one takes (e.g. a first-personal vs quasi-third-personal perspective), the order of conscious acts (e.g. adverbial modification vs second-order act), or the layering of mental processes (e.g. a thought turning back on itself vs consciousness taken transparently), is beyond the scope of this present discussion. The point, nevertheless, remains the same. As such, I shall assume a distinction between pre-reflective and reflective awareness, where the latter involves a deliberate act of consciousness that is secondary to the immediate awareness involved in pre-reflective consciousness. Accordingly, the distinction that I am assuming is between a form of awareness that is pre-reflectively and unproblematically *lived through*, and a form of awareness wherein the experience is characterised by *deliberate self-referential* acts of consciousness. As Eshleman (2010, p.45) puts it, "our reflective deliberations form a kind of epiphenomenal surface over our lived projects." Indeed, this is a pivotal part of Sartre's (2018, p.12) view overall: "reflection lacks any kind of primacy in relation to reflected consciousness… on the contrary, non-reflective consciousness is what makes reflection possible."

This leads me to the second core assumption of my argument. Namely, *that pre-reflectivity involves an essential normative element.*[42] My point here is that all pre-reflective mental states are, as Thomas (2010, p.164) puts it, "reason-sensitive" – a claim that can be grasped by appreciating two points. First, as I demonstrated in chapter 5, pre-reflective states occur through the organised structures of one's projects. The upshot of this is that pre-reflectivity is characterised by a kind of *practical* rationality, wherein hierarchical value-structures form across the agent's projects. Second, and connectedly, pre-reflective states establish normative standards for future/current projects and endeavours. Indeed, given that I do not act out of neutrality and given that I act within a pre-reflectively organised hierarchy, the question arises for each state as to whether I "affirm" or merely "acquiesce" it within the context of the given hierarchy (Thomas; 2010, p.163; Moran, 2001). In this way, Sartre (2018,

[42] I employ the notion of 'normativity' because I believe it accurately reflects the picture that I am presenting – i.e. that pre-reflectivity is agential, centred around values and commitments, and so on. However, I accept that this is an idiosyncratic use of the term. Hence, for an account that supports this view, see Larmore's (2010, p.16) *The Practices of the Self.*

p.646, emphasis added) frames pre-reflectivity through the lens of commitment: "at any moment whatsoever, I will apprehend myself *as committed within the world*…But this commitment is precisely what gives my contingent place its meaning, and what my freedom is."

To support this assumption further, note that the persistence of 'reasons' at the pre-reflective level does not require a deliberate act of consciousness, nor does it need to appear thematically by means of rational appraisal. Instead, reasons occur *because the agent is orientated within the world through an interrogative or practical mode of being*, as Sartre writes:

> In a nutshell, the world can only give its advice if we interrogate it, and we can only interrogate it in relation to a clearly defined end. Far from determining the action, therefore, a reason only appears in and through the project of an action (Sartre, 2018, p.588).

Understood in this way, pre-reflectivity allows us to capture some of the intricacies of unproblematically lived-through experiences. For example, my experience of my immediate surroundings does not normally appear neutrally valanced – as Webber (2009, p.31) puts it, objects and bodies in my environment do not appear to me as "brute existence" – but rather they appear "as a set of entities varying in degrees of salience and kinds of significance in relation to …[my] aims". In the same sense, one does not ordinarily live through one's experience passively, anonymously, in a formless world; rather, one's immediate self-awareness is an awareness of a world of values, possibilities, and activity. Pre-reflective agency, then, is not a matter of how our choices appear to us somehow internalised and isolated, but how the world appears to us when we are acting within it. As Webber writes:

> The central theme of the phenomenological movement is that the world we experience reflects the structures of experience itself and here we see part of the Sartrean version of this idea: action responds to aspects of the world that reflect the structures of our awareness of our surroundings…While it is true that Sartre understands us to have freedom over the way we experience the world, and hence over the way the world appears, *he does not locate this freedom in voluntary decision* (Webber, 2012, pp.327-328, emphasis added).

Ultimately, then, pre-reflectivity involves a practical form of agency that takes the shape of the projects that the agent pursues, that structures the way the world appears to us, and which persists without the requirement of a reflective act of consciousness. Indeed, this general description has been supported by two assumptions, which shall stand as my first premise and which I shall continue to build on throughout this chapter:

> Premise One: Pre-reflective consciousness is a normative form of self-awareness that is ongoing and a condition for reflective acts of consciousness.

I shall introduce my first non-assumptive premise (premise two) in the next section, where I argue that this first premise requires a strong connection to be drawn between the normative content of reflection and the agent's pre-reflective experience.

6.2. Deliberation, Normative-Framing, and Pre-reflectivity

In light of the assumption discussed in the previous section, I shall now formulate the second premise of my argument against the Rawlsian approach, which again draws directly from Sartre:

> Premise Two: Pre-reflectivity, in being essentially normative, ongoing, and a condition for reflection, determines whether reflection will take place and frames the content of rational deliberation.

On the face of it, this seems to be a fairly radical interpretation of the relationship between reflective and pre-reflective consciousness. In part, I think this is down to the unfortunate connotations associated with 'determines'. However, in using this word, I do not mean to suggest that there is a deterministic connection between pre-reflectivity and reflection – for, if any mental state has a normative element in the way I have just described, then this ought to apply to reflectivity also (that it too is 'in question', as Sartre might put it). Instead, the argument here is that one's projects establish whether the agent is *predisposed* to reflect on their commitments, rather than strictly 'determines' that course of action. My claim is that for any reflective act, there is a system of pre-reflective commitments that is a precursor to its occurrence.

This is not, however, the main point that I am advancing. Indeed, to say that pre-reflectivity is a precursor to reflection merely suggests that it is a condition for its eventuation. Yet this is a purely technical description, which I have already taken for granted in premise one, and which would bear little on a critical assessment of life-planning. Instead, the assumptions I have set out require a stronger relationship between these two forms of awareness. In particular, since pre-reflectivity is essentially normative – in a way that is traditionally (but wrongly) ascribed exclusively to a voluntaristic will – and since it is ongoing, then pre-reflective consciousness ultimately *anticipates the normative content of reflection*.[43] As Sartre writes:

> When I deliberate, the die is already cast. And if I must come to deliberate, it is simply because it is part of my original project to take account of my motives by *means of deliberation* rather than through this or that other mode of discovery… When the will intervenes, the decision is taken and it has no value apart from that of an announcement (Sartre, p.591, original emphasis).

In general terms, the argument for this position includes two main points, which support the second premise of my argument. First, assume for a moment that reflection *is* meaningfully separable from one's practical orientation and that a deliberate act of consciousness is required for every instance of agency. If this is the case, then an infinite regress would take place, owing to the fact that deliberation would be required to make the deliberative consciousness voluntary in the first place, and then deliberation would be necessary for that act also, *ad infinitum* (Webber, 2009, p.33). In order to avoid such a regress, voluntarists would have to maintain that reflection does draw on pre-reflective experience in some fashion. However, as I showed earlier, this point only provides the technical requirement that pre-reflectivity is a condition for reflection *to take place*. Hence, the second claim: as a self-referential act*, the structure of reflection* involves weighing up competing desires (interests, actions, or what have you) through comparative evaluation (Webber, 2009, p.33). Indeed, this structural description of reflection is something that Rawls (1971, pp.30-31) implicitly endorses – "the satisfaction of these desires must be weighed in our deliberations according to their intensity, or whatever, along with other desires" – and which Sartre highlights critically:

> The illusion here is caused by the attempt to regard reasons and motives as entirely transcendent things that I might weigh as if they were weights, and which could possess a weight as a permanent property even while, on the other hand, we want to regard them as contents of consciousness – which is contradictory (Sartre, 2018, p.591).

Here, Sartre highlights that the content of deliberation is not abstractable from the 'determination-determinables' structure that I described in the previous chapter, i.e. that such 'reasons and motives' cannot be seen as possessing normative weight outside of their place within a given project. Rather, what appears as 'ends' or 'reasons' within deliberation are, in many ways, *transformations of world-directed norms and possibilities*. In this way, whilst

[43] Here, there may be some difficulties with using the term 'content' for both pre-reflective and reflective forms of awareness. It should be noted that I am primarily concerned with the 'normative', and not the 'cognitive', side of this discussion. Accordingly, by using the word 'content', I am not advancing any particular thesis within the philosophy of mind.

reflective awareness involves a disruption to the spontaneity of conscious experience, it does *not* do so in a way that is "constitutively self-fulfilling" (Zahavi, 2015, p.184). Instead, the structure of reflective consciousness is, in a broad sense, relational; it involves a form of widened "fissure" wherein consciousness adopts a positional awareness of itself as its intentional correlate (Sartre, 2018, p.592). Whilst reflective awareness may differ from pre-reflectivity in terms of its "intensity, articulation, and differentiation", such qualitative changes are best thought of as amplified or adapted modes of presentation, rather than as means of introducing self-constitutively novel content to consciousness (Zahavi, 2015, p.195; Eshelman, 2016, pp.176-208).

The upshot of this description is that reflection does not manifest a disjunction from one's sense of practical agency, but instead involves *a continuation of it by means of retrieval*. Yet *qua* retrieval, the content of reflection must be anticipated by the normative structure of the pre-reflective experience. Indeed, a description of how this process establishes the 'illusion' of comparative analysis is even offered by Sartre, who draws an analogy to the Husserlian method of phenomenological reduction:

> …to take up Husserl's famous formulation, the mere act of voluntary reflection, through its reflective structure, operates the ἐποχή [epoché]; in relation to the reason; it suspends it, placing it between brackets. In this way *a semblance of evaluative deliberation can be introduced*, owing to the fact that a deeper nihilation separates the reflective consciousness from the unreflected consciousness (or motive), and owing to the suspension of the motive (Sartre, 2018, p.592, emphasis added)

With this in mind, if deliberation involves the 'weighing-up' of reasons and ends (as Rawls thinks it does), then Sartre's account provides an explanation for this function: reflection involves a self-referential act that draws on the essentially normative elements of pre-reflective consciousness. What is involved, in this way, is a form of suspension: when we deliberate, we single out the *desiderata* of deliberation in order to create the 'semblance' of evaluation within reflective consciousness – but this 'singling out' *does not separate* reflection from its foundation in pre-reflective awareness. Instead, the pre-reflective experience is a precursor to both the occurrence (premise one) and the content of reflective consciousness (premise two). Following Sartre's analysis, I have demonstrated this by deference to the structure of reflection, as well as by briefly suggesting (as in the quote above, and in my observations regarding the mode of presentation) how reflection might *appear* as a unique focal point for purposive agency.

6.3. Epistemic Implications, Introspection, and Agency

Thus far, I have argued that a pre-reflective sense of agency is a precursor to life-planning in that it anticipates its normative content – and in doing so, I substantiated the first two premises of my argument:

> Premise One: Pre-reflective consciousness is a normative form of self-awareness that is ongoing and a condition for reflective acts of consciousness.
>
> Premise Two: Pre-reflectivity, in being essentially normative, ongoing, and a condition for reflection, determines whether reflection will take place and frames the content of rational deliberation.

As it stands, however, this line of argument is susceptible to an important response. Specifically, Rawlsians could reply with something like the following: *even if* pre-reflective experience anticipates voluntary deliberation, it might be that the content in both instances *remains broadly identical*. After all, when we speak of the structural features of deliberation, we ought not to omit the general idea that reflection is associated with a form of *privileged access*. It is true, so the Rawlsian will continue, that we reflect in order to 'retrieve' pre-existing beliefs/ends/motivations/etc. – but this is precisely what life-planning is supposed to do! We plan in order to make explicit to ourselves our major aims and desires, so that we can pursue them more effectively and shape our lives as we see fit. Although planning might not be the direct focus of self-reproach (the Rawlsian will explain), reflection is simply the best way to make sense of a diffuse and often nebulous set of pre-reflective experiences. Hence, it is no wonder that Rawls has given life-planning special place when it comes to the issue of agency and goodness: it is simply the best expression of these notions that we have, and by emphasising its importance, *we lose nothing along the way*.

If this description is correct, then the Rawlsian has a genuine response to my argument so far, i.e. demonstrating that, whilst the distinction between pre-reflectivity and reflectivity admits of some interesting phenomenological nuances, it has no meaningful impact when it comes to characterising persons and their respective conceptions of the good. This is a significant challenge for two reasons. First, it provides a direct counter to my argument whilst employing identical assumptions. Second, it aims at the main issue at hand: whether the distinction between pre-reflectivity and reflectivity has any bearing on Rawls's account of goodness. Jointly, this counter holds that reflection does not involve any meaningful changes to the normative content that it inherits from the pre-reflective experience – i.e. that reflection *just is* a form of presentation – meaning that it accurately characterises a person's conception of the good. In recognition of this reply, the third premise of my argument states the following:

> Premise Three: Reflection involves non-trivial adaptions to the agent's normative commitments.

Here, recall that my main target is the Rawlsian account of subjectively rational life plans. With this in mind, I can define 'non-trivial adaptions to normative commitments' more narrowly. Specifically, my argument is that there are *epistemic and normative implications to the reflective act* and not just to *the information* that reflection draws upon. This clarification is important, if only because Rawls's account of subjective rationality shows that he resolves the epistemic shortcomings of reflection with, as it were, *more reflection*:

> In this account of deliberative rationality, I have assumed a certain competence on the part of the person deciding: he knows the general features of his wants and ends… (Rawls, 1971, p.418).
>
> Awareness of the genesis of our wants can often make it perfectly clear to us that we really do desire certain things more than others. *As some aims seem less important in the face of critical scrutiny*, or even lose their appeal entirely, others may assume an assured prominence that provides sufficient grounds for choice (Rawls, 1971, p.420, emphasis added).

I will not labour this point much further, as Rawls's commitment to reflection and its benefits has been made clear throughout this thesis – for example, in chapter 3's discussion on the psychological background of life-planning and the principles of deliberative rationality. Nevertheless, the upshot of this approach is that Rawls overlooks the ways in which reflection is genuinely adaptive: by altering the mode of presentation of normative content, reflecting on one's major commitments only tells us one side of the story. As a result, whilst Rawls accepts that reflection is epistemically limited by the information that it receives, he has no account of the errors or normative implications associated with *the structure of introspection and planning itself*. I shall supply such an account in the next two sections, thus supporting my third premise and dismissing the Rawlsian reply I identified in this discussion.

6.3.1. Agency, Misrepresentation, and the Structure of Reflection

To begin with a general summary, my argument for the normative salience of reflection is that reflective acts involve a form of *reconstruction*, not only does it draw from pre-reflectively undertaken projects, but it also reconstructs them in the process of changing their mode of presentation. More specifically, I agree with Sartre (2018, pp.592-593) that reflection is not how an agent chooses their ends, but instead concerns the way the end is presented. In many ways, reflection is both interpretive and anticipatory: it represents the agent's attempt to come to terms with their projects, and in doing so, sets expectations regarding how their ends will be achieved. However, there is no guarantee that such a reconstruction will be accurate. In fact, talk of accuracy is misleading: to deliberate is itself a choice, it is *not solely a function of consciousness to make its 'innerworkings' somehow explicit and immediately available*. The movement from pre-reflective to reflective awareness must be normatively inflected since to change how the content of consciousness is presented *is* to make a choice that bears upon the agent's sense of their ends, beliefs, and personal responsibility.

Two main points support my claim in premise three. First, *the act of deliberation is itself a choice*, meaning that the transition from pre-reflective to reflective awareness is essentially normative (in the terms that I have set so far, i.e. reason-giving/reason-sensitive). Second, epistemically-relevant errors occur because of the structure of introspection, meaning that *the very act of deliberating can misrepresent the normative stance of the agent*. As a result, I challenge the idea that reflection is merely a means of making explicit one's major aims and desires (Rawls, 1971, pp.414-415). In deliberating, one is not implementing a formal procedure to understand one's desires more clearly. Instead, one is undertaking an act of consciousness that is reconstructive and that is itself a part of the normative stance of the agent. Larmore makes a point similar to my own:

> In self-reflection, so I conclude, we are getting a grip on ourselves so as to be able to *reappropriate ourselves*…Every self-conception we frame not only refers to what we are insofar as we are committed to being it, but also constitutes in its own right a commitment, a belief governing what we should think about ourselves (Larmore, 2010, pp.86-87)

In this case, talk of 'reappropriation' – or perhaps better, 'appropriation', since one's immediate experience is first given pre-reflectively and only then represented through reflection – conveys both of my main points: to appropriate something is to act on it, to present it *as if it were one's own,* and in doing so one is liable to misrepresent core aspects of whatever has been appropriated (importantly, I will say more about this latter claim – i.e. my second point – in the next section). Indeed, I do not have a thematic sense of 'ownership' when I experience pre-reflectively; instead, I am actively engaged in the world, navigating my environment without my self-awareness being rendered explicit. In deliberation, however, this sense of ownership comes to the fore as a constitutive part of the rational appraisal, which Sartre highlights by repeating the use of first-personal pronouns: "in the act of reflection I bring judgements to bear on my reflected consciousness; I am ashamed of it, I am proud of it, I want it, I reject it, etc."

(Sartre, 2018, p.11). Yet, in order to ground such judgements in the first instance, the agent must determine themselves towards a judicative stance; that is, they must choose to reflect on themselves, and they must attribute some value to this act of reflection.

My main point is that reflection is not a neutral form of representation but *another way for the agent to commit themselves within the world*. When I speak of the primacy of pre-reflectivity (as I did in the previous section) one might wrongly think that freedom only persists *prior* to reflection; but this is a mistake equivalent to thinking that agency consists in explicit acts of volition – by which I mean, it provides an equally fragmentary account. Instead, as Sartre (p.583, 2018, emphasis added) puts it, pre-reflective agency is "a foundation that is *strictly contemporary* with the will or the passion and that each of these latter manifests in its own way." Pre-reflectivity does not 'precede' reflection in anything more than a formal sense; instead, it is that deliberation appears as a spontaneous choice *alongside* the organisation of one's ends – just as I have to choose my ends, I also have to choose to make such ends explicit (Sartre, 2018, p.591). As such, deliberation must involve non-trivial adaptions to the agent's commitments since it stands as *a commitment in extension to their pre-existing projects* – "reflection does not therefore capriciously arise within being's pure indifference, but is produced within the perspective of a '*for the purpose of*'" (Sartre, 2018, p.229, original emphasis).

All of this is to substantiate my first claim in support of premise three, that deliberation is itself a choice – and not purely a way of representing the innerworkings or innermost desires of the agent. A further upshot of this is that, when this claim is taken independently, it also brings into question the *function* of reflection. Indeed, as I have suggested, Rawls (2007, p.310) describes the core aspects of life-planning as a form of representation, a way of consolidating pre-existing commitments, i.e. "making our desires our own…bringing our desires and impulses into balance…[having] appropriated it in our thought…we affirm our way of life after due reflection."[44] My point in this section, put in the language that Rawls employs here, is that this form of 'appropriation' cannot be understood purely in terms of 'affirmation'. Rather, we must understand reflection as a continuation or an adaption to the agent's normative commitments, rather than just a means of making such commitments explicit. As I will demonstrate in sections 6.4 and 6.5, this will have important implications for Rawls's account

[44] Here, Rawls is actually defending Mill's account of life-planning from the claim that it imposes an eccentric ideal of individuality on the subject. However, the relationship between Rawls's (1971, p.426, n.20) account and the one provided by Mill (1987, ch.2) is relatively clear, as he also cites *Utilitarianism* when spelling out the relationship between life-planning and the Aristotelian principle in *Theory*.

of goodness. In the next section, however, I shall complete my defence of premise three by highlighting the main structural features of reflection that enable its normative connotations – beginning with a clarification, that reflection need not be considered a dysfunctional aspect of consciousness.

6.3.2. Possibilities, Normative-Downplaying, and Epistemic Uncertainty

The discussion above brings me directly to my second point in support of premise three: the structure of reflection permits normatively-relevant epistemic errors. Indeed, at this point I have only established that reflection is not a neutral form of representation; what the second point serves to demonstrate is *why* this is the case. As such, I need to pinpoint the parts of reflection that distinguish it from a purely representational act of consciousness beyond locating it as an act alongside the agent's pre-reflective awareness. By doing so, I shall criticise the Rawlsian position more directly, as this point shall serve to anticipate my conclusion overall, that life-planning does not suitably express persons' conceptions of the good. To begin, I want to note the extreme version of this analysis, summarised here by Larmore:

> …Sartre… insisted that reflective consciousness is fated to produce in its object—the image of the self—*a deformation so profound that the idea of its leading to self-knowledge constitutes an illusion*. Reflective consciousness is incapable… [Sartre claimed] of accounting satisfactorily for the cohesion of the self, since lived reality necessarily eludes the distanced point of view that is its essence (Larmore, 2010, p.90, emphasis added).

Here, the extreme view is that reflection is inherently distortionary, and that self-knowledge is doomed to failure – and indeed, this is the view that Larmore attributes to Sartre. Whether Sartre held this precise viewpoint, however, is itself questionable and I would certainly resist it. As Webber (2009, p.34) notes, Sartre's description of reflection as a "*truquer*" (a 'deception') refers to the fact that *its rigged*, or as I have put it, that its content is anticipated by a pre-reflective structure of the agent's projects. Yet the fact that reflection is rigged does not mean that it is distortionary *through and through*; a card trick is rigged but not illusionary as such, nor is it deceptive *proper* once one knows what the game involves. As I will suggest in section 8.1, reflection is not always or entirely doomed to misrepresentation – nor should it be conceived of as a 'dysfunctional' feature of consciousness.

The above does not, however, undermine the argument that the structure of reflection can non-trivially adapt the agent's commitments by introducing a myriad of introspective errors. Even though reflection can fulfil some role (discussed in chapter 8), this does not mean that one should not take heed of its shortcomings – especially when reflection is used to characterise something as comprehensive as a person's good, and if such shortcomings are part and parcel of the act of reflecting. In this way, Larmore does capture the broad strokes of Sartre's view: that the structure of reflection enables the agent to misrepresent the object of awareness and ultimately distort the foundation for self-knowledge. More concretely, Sartre observes that reflective consciousness establishes connections between distinct mental states/events *without being able to explicate the processes that link them together*:[45]

> In the first case, reflective consciousness apprehends as a single object two psychological objects that were at first given separately... The consequence is a total action by one on the other, occurring across a distance and by means of a magical influence. For example, my humiliation yesterday is what motivates entirely my mood this morning, etc. (Sartre, 2018, p.240).

Here, Sartre observes that in reflection distinct actions are linked causally, without their link being explained, and without any attempt to make intelligible the role that the agent has in shaping how such instances come about, e.g. 'I am sad this morning because I was humiliated yesterday'. The upshot is two-fold: there is both an *epistemic uncertainty*, in that the agent has deferred to a 'magical' explanation for the relation between two distinct events or states, and a form of *normative-downplaying*, where the agent's participation in these events is minimised or even overlooked. Thus, reflective consciousness enables the agent to view the relationship between events through an almost epiphenomenal lens, with each state taken as a product of the last without an intermediary, and with the agent conceiving of themselves as a kind of passive recipient of such states.

A similar description, I should add now, can apply to life-planning. After all, *the misrepresentation is a function of the reflective act*, and not simply of the temporal relation between the two original events (e.g. chapter 4). Indeed, one can see this by slightly altering Sartre's original example, e.g. 'I will be sad *tomorrow* because I was humiliated today.' Much more directly than this, Rawls (1971, p.550) draws a connection between happiness and the successful formulation and execution of one's life plan; "a rational plan…makes a life fully worthy of choice and demands nothing further in addition." Yet, at most, the activity of planning can only achieve a broad outline of the relationship between a person's major aims. Hence, the idea that the life-planner requires 'nothing further in addition' seems precisely the

[45] Here, I focus on one example of the epistemic limitations of reflection that Sartre highlights. However, Sartre (2018) has a more expansive account of reflection's shortcomings in *Being and Nothingness*. For an account of such errors, which provides a useful background to my current argument, see Eshelman (2016, pp.176-208).

form of misrepresentation that Sartre highlights in his work. Such an outlook leaves a normatively and epistemically relevant gap in that the agent does not just become happy because they have set and realised their plan, there is something more – they must also, as it were, *live* happily. Although I will say more about this in section 6.4, Stendhal's (1982, p.39) warning against reflective analysis gives expression to this point: "I feared deflowering the happy moments I have met by describing them, anatomizing them. Well, that is what I will not do; I will skip over the happiness."[46]

My main point here is that these observations point towards a kind of *transmutation of possibilities* within reflective consciousness. It is not that reflection only posits connections between objects without explicating their underlying processes, but it also misrepresents the way in which such processes present themselves as intelligible within the lived experience. Hence, deliberation is not purely 'methodological' either – in the sense that we might expect when one sets one's horizons for action – since there is a sense of openness and possibility to the pre-reflective experience that reflection is liable to distort; "what we are accustomed to call 'the revelation of inner sense' …is a process that is already constructed, with the clear intention of concealing…the genuine 'immediate datum' of our freedom, from ourselves" (Sartre, 2018, p.84; 660). So that I can explain this further, consider Sartre's brief analysis of fear, which directly references the role that planning can often play:

> My reaction will be reflective in type…*I plan ahead for myself a number of future courses of action*, whose purpose is to distance myself from the world's threats…I escape fear by the very fact of placing myself in a framework where my own possibilities are substituted for the transcendent probabilities in which human activity had no place (Sartre, 2018, p.68, emphasis added).

Here, whilst one's pre-reflective experience is necessarily shaped by a world-directed sense of possibility, reflection distances the agent from this basic sense of engagement – meaning that the agent is able to view their situation *deterministically*. To be specific, reflection allows what I have called 'normative-downplaying' and 'epistemic uncertainty' because of two structural features of the reflecting consciousness. First, reflection discovers consciousness as 'consciousness-reflected-on', and not 'consciousness-in-action' (Thomas, 2010, pp.161-163; Sartre, 2018, p.232). Such an awareness is "always too late" – both literally, in that reflection *retrieves* previous states along the temporal flow, and descriptively, in that reflection presents

[46] This quote is also found in Larmore's (2010, p.21) *The Practice of the Self*.

itself as an originary act of will, despite being (as we have seen) *already* implicated within a more foundational sense of agency (Thomas, 2010, p.162).

Second, reflection *qua* positional act of consciousness posits as its correlate as an object of intention. As I mentioned earlier regarding perspectival ownership, this form of objectification is absent from the pre-reflective self-experience – as Thomas (2010, p.164, original emphasis) puts it, "you do not occupy your mental life or view things *from* it: you *are* it." However, in reflection, the focus *is* thetically explicit and self-referential. Since pre-reflectivity is not prior to reflection but contemporaneous with it, then consciousness enacts this self-reference *by taking itself as an object of intention*; "its motivation lies within itself, in a twofold movement… of internalization and objectification…" (Sartre, 2018, p.230). Since I am pre-reflectively aware whilst reflecting, and since the former is constitutive of my sense of subjectivity, I cannot also be reflectively aware of myself as the subject of phenomenal consciousness (for this would bifurcate or double-up one's subjectivity); instead, I must be aware of my 'self' within a subject-object relation. That is, I must 'invent' a self to be aware *of*.

Accordingly, not only does reflection permit the agent to view the world deterministically, thereby distorting the 'datum' of agency (i.e. possibilities), but it does so by a process of 'Othering' where the agent views themselves in a quasi-objectified light; "I may hide from myself that these possibles are *myself*… as if [they] were generated by an already constituted object which is nothing other than…[the] Self" (Sartre, 2018, pp.82-84, original emphasis). In this way, if consciousness is typically diaphanous – by which I mean, that it is world-directed without explicit self-awareness – then reflection subtends this quality by treating one's self as an object of intention. Yet this is precisely why reflection maintains an illusionary quality: one attempts to 'bracket' the pre-reflective experience, which cannot really be bracketed, and one objectifies the self, which consists originally in subjectivity (Sartre, 2018, p.19; 81; 591). In turn, this shift in terms of awareness – from, as one might put it, an unproblematically first-person perspective to a quasi-third-person perspective – enables epistemically and normatively relevant adaptions to the content of consciousness. A summary of this point is provided by Eshelman, who aptly draws out its similarities to many proposed approaches to ethical reasoning (e.g. Hume's 'general point of view' or Smith's 'impartial spectator'):

> If philosophers… masqueraded about as if they could adopt a view from nowhere, when, as a matter of fact, they inevitably take a view from

> somewhere, then the problem of introspection inverts this game of charades…. We find ourselves stranded in between two impossible positions: taking a view from nowhere and trying to adopt a view from somewhere other than our own. We try to adopt a third-person perspective on an essentially first-person phenomenon and in so doing we vitiate our inhabited perspectivity…we cannot take a first-person perspective upon ourselves. (Eshelman, 2016, p.196).

Throughout this section, I have argued in favour of premise three: reflection involves non-trivial adaptions to the agent's normative commitments. This is no small topic, but I have attempted to focus this discussion by narrowing it down, i.e. it is the structure of reflection that adapts the agent's commitments, and not merely the information that the reflective act processes. I have supported this premise by arguing in favour of two more specific claims. On the one hand, the act of deliberating is itself a choice. On the other, deliberation leaves the agent open to several errors of introspection. The former point dismissed the understanding of reflection as a purely neutral form of representation – which was key to the Rawlsian response detailed at the start of this section – whilst the latter point built on this and demonstrated that reflection also enables normatively-salient forms of misrepresentation. Taken together, it is apparent that reflection allows for the transformation of possibilities, permits the agent to recede from their agential role, and involves an adaption to first-person perspectivity (meaning that the agent views themselves as the object, rather than the subject, of experience). Although more could be said here, and indeed I will develop this line of reasoning further when drawing my concluding points, this suffices to support premise three of my argument.

6.4. An Extension of Commitments, Not the Definition of Goodness

So far in this chapter, I have established two basic assumptions, which allowed me to develop three main premises. These premises will enable me to draw two concluding points against Rawls's position on life-planning, which I shall put forward in this section. To summarise the discussion so far, my argument is this:[47]

> Premise One: Pre-reflective consciousness is, (i), a normative form of self-awareness that, (ii), is ongoing and a condition for reflective acts of consciousness.
>
> > Subpart (i): pre-reflectivity is reason-sensitive or reason-giving, meaning – among other things – that an agent's normative commitments (i.e. their projects) can be 'pursued' or 'made' without reflective acknowledgement.

[47] Here, the claims given in Roman numerals are subpart premises.

Subpart (ii): Whilst reflective and pre-reflective forms of awareness are contemporaneous with one another, primacy is given by the fact that pre-reflective awareness is a condition for reflective acts to take place.

Premise Two: Pre-reflectivity, in being essentially normative, ongoing, and a condition for reflection, (iii) determines whether reflection will take place and, (iv) frames the content of rational deliberation.

Subpart (iii): Here, 'determines' refers to the fact that a hierarchy of pre-reflective commitments stands as a precursor to reflective acts. Indeed this claim builds on, instead of undermines, the thesis present in subpart (i), i.e. that all mental states are reason-sensitive.

Subpart (iv): Since the structure of reflection is comparative, and since the content of reflection is not self-constitutively novel, reflective acts retrieve 'reasons' or normative commitments from projects that the agent pursues pre-reflectively.

Premise Three: Reflection involves non-trivial adaptions to the agent's normative commitments.

Subpart (v): Reflection is itself a normative activity (though the originary decision to reflect is not one that requires a deliberative act). As a result, reflection is not a neutral means of representation; it is a commitment alongside the agent's other projects.

Subpart (vi): Reflection permits key introspective errors; namely, epistemic uncertainty, normative-downplaying, and the objectification of the self. Whilst reflection need not be characterised as a dysfunctional feature of consciousness, such adaptions always take place by virtue of the structural dynamics of reflective acts.

Whilst I have applied parts of this argument to *Theory* throughout this discussion, the culmination of these points supports a more comprehensive dismantling of Rawls's position, which bears directly on the argument from congruence. As a reminder, Rawls's approach to congruence can be described as the reconciliation of two perspectives; on the one hand, the shared perspective of justice, which is abstracted from contingent matters; and on the other hand, the individual perspective, which is embedded within its contingency (Freeman, 2007, pp.265-266). Rawls advances an account of life-planning through which he defines the perspective of the individual, and concurrently, their conception of the good. In sum, Rawls defines the good as a life lived according to a plan and uses this account to argue that goodness and justice are congruent.

The identification of goodness with life-planning is precisely the target of my present argument. My issue is with Rawls's (1971, p.395) claim that a person's good is determined by their rational plan of life. More concretely, I disagree with Rawls's position that a person's conception of the good can be given in the formulation and pursuit of a subjectively rational life plan, even when such a plan is understood in processual terms, where each person sets and follows their main purposes or interests. Crucially, my reason for this lies in two concluding points, which result from the premises that I have laid out and defended above. I will unpack conclusion one in this section and discuss conclusion two in section 6.5:

> Conclusion One: Goodness cannot be defined in terms of a life plan since the choice to engage in life-planning (and the purposes that are identified thereof) *is an extension of what the agent already takes to be good.*
>
> Conclusion Two: By focusing on life-planning, Rawls mischaracterises goodness and brings disunity to the self. In particular, since the structure of reflection (i.e. active life-planning) is normatively and epistemically adaptive, *it distorts commitments that otherwise comprise the agent's conception of the good*.

So far, I have been speaking about the connection between reflection and pre-reflective awareness through the lens of normativity. Similarly to previous chapters, I have maintained that when an agent undertakes a reflective act, they are already committed within the world *via* a hierarchy of pre-reflective projects and values. The step that I take in conclusion one is to connect this normative hierarchy with the agent's conception of the good, where it becomes clear that a life-planning approach to goodness is, at best, inexhaustive. To demonstrate this, I will again engage with Larmore's work, which advances a similar line of critique to my own:

> …though our self-conception revolves around our purposes, it does not follow that our good is determined by the purposes we pursue. Our good may instead involve the sort of bafflement of our aims which provokes us to change our idea of who we are (Larmore, 1999, p.103).

In the current literature, Larmore's critique is often considered as being broadly identical to the one advanced by Slote – i.e. that Rawls's focus on life-planning excludes certain localised goods – which I dismissed earlier in chapter 4.[48] For the most part, this misattribution arises because of Larmore's *use of the same concept*, i.e. unforeseen goods:

> …the unforeseen good he now pursues does not fit a purpose he had at the time, but instead is the source of the one he has subsequently devised. In this case, the good is not being judged by reference to his purposes; rather, his purposes are justified by appeal to something he understands to have been good for him (Larmore, 1999, p.110)

On closer inspection, however, it is clear that Larmore (1999, p.110) is moving towards a far deeper point than the argument that life-planning omits unforeseen goods – as he clarifies, "let me remark that I am not arguing that some goods by their very nature elude the art of planning." Instead, it is the latter part of Larmore's claim (in the second indented quote above) that carries most of the argumentative weight: goodness cannot be defined in terms of the agent's rational purposes (i.e. their life plan), *since such purposes are derived from something that the agent already takes to be good*.

Whilst I do not think that he recognises this directly, my position is that a distinction between reflective and pre-reflective awareness is a necessary component of Larmore's claim here – and correlatively, of the first concluding point that I draw from my argument. To put this in critical terms, Larmore's contention is that the agent's purposes are 'justified by appeal to something he understands to have been good for him' – yet this description reads as *importantly incomplete*, owing to the indeterminate use of the term 'something'. To make up for this, Larmore (1999, p.111) posits that, "…it is a good—a good which befalls us—that acts as the criterion of rational purpose, and not the other way round." Whilst I agree with the implications of this point, I take issue with its formulation. It suggests that our good is determined exclusively by the 'things' or 'events' that we encounter. Yet this account is as unpalatable as the one Rawls provides, given that it frames the agent's role in essentially passive terms, and still fails to explain *why* the agent accepts the unforeseen good *as a good*. As a result, Larmore's argument – though it correctly identifies that goodness is presupposed by one's life plan – fails to explicate the underlying process that makes this relation possible.[49]

To clarify, there are two overlapping points that come out of my first conclusion and from my engagement with Larmore. First, a pre-reflective sense of agency is a necessary

[48] Examples of this misinterpretation are found in White (2021, p.303) and Heyd and Miller (2010, p.32).

[49] I do agree with Larmore, however, that the good is something dynamic and continuous. In criticising his position, I am emphasising the need to carry this reasoning to a deeper level, as it were. The good is dynamic in the ways Larmore points towards, the essential question is *why*.

condition for the process of formulating and pursuing a conception of good. Unforeseen goods do have the benefit of demonstrating this point more clearly (e.g. I encounter a good that I did not plan for, and *only then* incorporate it into my major purposes); however, my contention is that this applies universally. Even the planned-for-good is only good within the context of the agent's projects. For example, my desire to complete a PhD is one of my goods because of my aim to become a professional philosopher, which itself has some even more general setting; say, in the enjoyment I get from studying philosophy. Crucially, however, we should not explain the increasing generality of these projects in purely temporal terms – as Rawls does, by carrying the act of planning over to the entirety of a person's life – *but instead regard it as revealing something genuinely novel about the nature of goodness*. The fact that constitutive parts of my plans defer to something else that I find valuable does not suggest that a globalised plan can provide a catch-all solution. Instead, it demonstrates that a person's aims and interests – that is, what they understand to be good – take for granted commitments that they have already made without reflecting on them.

Second, Rawls provides a fragmentary account of goodness by defining the good life through a process founded principally on reflection. As I demonstrated in chapters 3 and 4, Rawls defines a person, and their version of the good, as a life lived according to a plan. However, as I have indicated above, plan-formulation primarily involves the attribution of purposes and interests to a particular understanding of the self. Hence, Rawls makes a crucial error in tying a person's conception of the good as a whole *to an explicit self-conception* founded on a description of their 'purposes' and 'causes'. Against this, I argue that a person does not formulate a self-conception and then pursue their good. Instead, this self-conception is as much a part of their good as the actions required to fulfil explicit intentions – indeed, this is implied from the fact that reflection is itself a normative activity (premise three, subpart v). As such, fragmentation comes into Rawls's account of goodness since it conflates the agent's good *overall* with an idea of the good derived from a *localised form of self-understanding*. In light of my argument, the process that Rawls describes can be explained in terms of reassessing a particular understanding of oneself, rather than exploring the nature of one's good.

With these two points in mind, which are supported by the argumentation I have provided throughout this chapter, I can summarise my first concluding point. Goodness cannot be defined in terms of a life plan since life-planning itself is dependent on what the agent already takes to be good. Crucially, I have made this claim far more specific. I have argued that 'what the agent already takes to be good' should be cashed out in terms of a pre-reflective normative framework, i.e. the commitments and values that the agent prioritises, without the requirement of reflection. Additionally, I have argued that the act of life-planning depends on the agent's self-conception, and that this conception is an essentially localised phenomenon – meaning that it is a part of their good, rather than a neutral way of rendering their aims, interests, and projects determinate. Accordingly, the Rawlsian life-planner can be understood as reassessing their self-understanding, rather than giving expression to their conception of the good. In the next section, I shall develop these points further by arguing that Rawls's account of life-planning not only mislocates the good but also mischaracterises it – thus establishing my final concluding point of this chapter.

6.5 A Form of Self-Suspension, Not of Self-Unity

As I have just argued, an agent's life plan depends on a localised self-conception, entailing that goodness cannot be defined in terms of life-planning. Combined with the analysis of premise three, subpart (v), it is relatively straightforward to see how this leads to my second concluding point, which, as a reminder, states the following:

> By focusing on life-planning, Rawls mischaracterises goodness and brings disunity to the self. In particular, since the structure of reflection (i.e. active life-planning) is normatively and epistemically adaptive, it can distort commitments that otherwise comprise the agent's conception of the good.

To recall the process of life-planning, in line with my analysis in chapters four and five, the agent sets their horizon for action by reflecting on their major aims and interests, thereby formulating a conception of themselves and their good. The issue here is that this process involves the same errors that I described in section 6.3, i.e. epistemic uncertainty, normative-downplaying, and the objectification of the self. Though all three of these errors apply to life-planning (for example, epistemic uncertainty, wherein one sets one's intentions but does so in a way that misrepresents the 'datum of freedom'), the objectification of the self applies most forcefully to Rawls's (1971, p.561) position, given his claim that a person, and the self's unity, is defined by a life plan. Hence, the strongest way to put my argument is this: Rawls's account of the self is disunified and the account of the good that he provides is distortionary, since *Theory*'s focus on life-planning entails a prioritisation of a *specific form* of self-conception; in particular, an idea of the self that is ultimately objectified. [50]

[50] To stress, my issue is with Rawls's classification of goodness and unified selfhood with reflective forms of consciousness, it is in that totality that the self is objectified. Thus, Sartre's (1960, p.20) claim in the

To appreciate this point, recall that pre-reflective awareness – a form of awareness that I have associated with the subjectivity of consciousness – is essentially normative, i.e. subjectivity exists *through* its commitments, not prior to them. In contrast, reflection is a deliberative process that *suspends* such commitments. Hence, the process of reflecting is liable to attribute objective characteristics to that same self. Importantly, this aspect of life-planning is directly foreseen by Sartre:

> In examining myself, my goal is, in effect, to determine exactly what I am, in order to resolve that I will be it, without deviation — even if that sets me searching, thereafter, for the means by which I might be changed. But what does this aim amount to, *other than that of constituting myself as a thing*? (Sartre, 2018 p.107, emphasis added)

The upshot of this is that one's life plan ascribes concrete qualities, tendencies, or desires *to a subject that necessarily eludes such attributes*. Moreover, the cause of this mismatch is that life-planning ultimately depends on acts of reflection, meaning that it takes the self as it appears through a deliberative consciousness to embody the self *simpliciter*. However, as I demonstrated previously, such a consciousness always misses its target; judgements on one's character often concern the past and not the present; plans that anticipate the future assume a deterministic outlook or reshape the self's relation to its possibilities; patterns of conduct are presented as if they are typical of, and not perpetuated by, the agent (Sartre, 2018, p.107). Contrary to Royce, we do not say 'who we are' when we formulate our life plans, we only say what we think we are; and in this process of deliberation, we make ourselves into an object, thus abstracting our good from its foundation in subjectivity.

In this way, to define goodness through life-planning depends on a process of self-conception that is disunified, and which distorts the very subject that it claims to represent. As Larmore (2010, p.15) writes, "…we cannot become one with some form of existence without ruining the oneness we seek in the very act of contemplating it." In more general terms, then, Rawls *overdetermines* the notions that he attempts to define: the good is given as a concrete list of purposes, rather than as a constitutive part of the agent's experience, whilst the self is understood through an objectified lens, rather than a subject committed within the world. As Alford (1991, p.151) puts it, "this is not to say…that the self may not be more or less unified, more or less complete…it is not an all-or-nothing affair. But this is precisely what Rawls makes

Transcendence of the Ego proves relevant: "reflection 'poisons' desire… my state is suddenly transformed into a reflected state, there I am watching myself act, in the sense in which one says of someone that he listens to himself talk."

it in his account of the unity of the self." In the same way, I do not want to claim that the self cannot be unified, only that its unity cannot be dependent on a Rawlsian notion of life-planning. Instead, a definition of goodness and selfhood needs to strive towards the complexities of life, not simplify them through rationalistic decision-making procedures.

To bring this discussion to an end, Taylor's (2015) comments during a presentation at Duke University prove appropriate, given that he explores the general failure to define goodness with modern philosophy (implicating, I should add, the life-planning account directly):

> I think all these forms…fail…So where could you possibly turn? This must be the end to the intellectual universe; but I don't think it is… Again, we need to do a bit more phenomenology and look at how we actually operate…for instance, if you have some idea of the good life, you have some ideas of what the motivations human beings have [or of the motivations that you have] … and with this palette of motivations is part of how you understand the world around (Taylor, 2015)

I agree with Taylor's general sentiment. After critical inspection, there is a *lacuna* left when it comes to defining the good life. In this particular context, Rawls's argument from congruence fails, for without a workable account of the good, he has no way of demonstrating that justice and goodness are congruent. As a result, a positive alternative must be provided. As I show in chapter 8, *one's choice of the good ought to be considered as identical with one's choice of oneself within the world.*[51] To hold a conception of good is not only a lens through which one sees the world – it is a way of existing within it. As one might expect, the solution to this *lacuna* requires paying more attention to pre-reflective experience, whilst remaining sensitive to the structures of reflective awareness. In my positive contribution, I will still situate reflection as pivotal to the congruence argument as a whole. Yet, this picture must be kept balanced. As Sartre's work shows, there are normative implications to different forms of (self-)awareness, meaning that a person's values and purposes cannot be abstracted from that context. *It is in light of this complexity* that the congruence argument must be advanced, not despite it. Whilst the final chapter of this work aims to achieve that goal, chapter 7 will develop the notion of pre-reflectivity in more detail, to support the main phenomenological claims of this thesis.

[51] Here, I adapt Sartre's (2018, p.79) phrasing of a "project…which is akin to my choice of myself within the world." I shall use 'choice of oneself' throughout.

Chapter 7

A Pre-Reflective, Non-Positional, and Non-Localised Kind of Freedom: Developing and Defending Sartre's Phenomenological Account of Agency

7. Introduction

In the previous two chapters, I employed Sartre's notion of pre-reflectivity to criticise Rawls's account of agency and goodness. In doing so, I took up a Sartrean position on normativity and self-awareness to argue against Rawls's commitment to rationalistic voluntarism and his characterisation of the good *via* life-planning. In regard to the first point, I argued that – even in the context of justice as fairness – the experience of agency resists a voluntaristic explanation, whilst in regard to the second, I argued that life-planning is a transformative extension of what the agent already takes to be good. In more general terms, I criticised the working parts of Rawls's argument from congruence, i.e. the claim that one can incorporate one's sense of justice within one's conception of the good. Here, the implication is that, since Rawls mischaracterises the good and the notion of persons *qua* agents, the congruence argument fails as a whole. Hence, the final chapter of this thesis – chapter 8 – reformulates Rawls's congruence argument from a Sartrean perspective.

In this chapter, however, I look to tie up some loose ends regarding the Sartrean account of pre-reflectivity and anticipate some concerns that Rawlsians may have in concerning the arguments that I have advanced so far. For example, throughout this thesis I have largely assumed and applied Sartre's notion of pre-reflectivity to justice as fairness, rather than argued for it directly. This chapter remedies that omission by defending this notion of pre-reflective consciousness independently through the work of Ratcliffe (2017; 2024). Additionally, whilst I criticised Rawls's rationalistic voluntarism in chapter 5, the same charge has also been levelled against Sartre. Here, I demonstrate that this charge is misplaced and identify some of the misinterpretations of *Being and Nothingness* that have led scholars to wrongly associate Sartre with voluntarism. In sum, I investigate the core features of pre-reflectivity – namely, that it is non-localised, world-directed, and involves the norm of responsibility – to shed light on some of the relevant phenomenology and dispel concerns that Rawlsians may have regarding the Sartrean account of agency.

In section 7.1, I anticipate a counterargument to my approach: the Sartrean account of agency utilises metaphysical claims that are incompatible with Rawls's liberalism. In considering this response, I draw a Rawlsian distinction between metaphysical claims, ruled out by justice as fairness, and general facts about persons and the world, which are required to respond to the problem of stability. I demonstrate that a phenomenological thesis about agency can be teased apart from a metaphysical commitment to freedom. Having done so, this section consolidates the relevancy of Sartre's account of agency to the Rawlsian argument from congruence, thus securing (as in the ensuing sections) the possibility of a positive interchange between the two.

In section 7.2, I demonstrate – primarily using the work of Ratcliffe (2017;2024) – that a pre-reflective sense of possibility is a structural feature of non-positional consciousness. Specifically, I establish that – without frameworks of agent-relative possibilities, given primarily through the 'I can' – the otherwise immediate and unproblematic sense of being in a particular intentional modality breaks down. Concurrently, I submit that a Sartrean form of agency is presupposed by an intentional or reflective act of consciousness. The upshot here is that by providing independent support for the Sartrean account of agency, I look to secure it as a (phenomenological) thesis about persons that Rawlsians may draw from in their account of stability.

In section 7.3, I clarify how pre-reflective consciousness is an active, world-orientated form of awareness. To achieve this, I demonstrate *via* Sartre that (in line with section 7.2) one's possibilities cannot be a quality of a purely subjective consciousness, nor can they be a property solely belonging to the world. Instead, both consciousness and the world arise as a basic, unitary, phenomenon. In showing this, I clarify a misinterpretation of Sartre's account, which takes one's sense of possibility to either be an abstraction from, or an imposition onto, the world. At the same time, I anticipate my positive contribution in chapter 8, where I establish various links between the Rawlsian organisation of society's basic institutions and normative forms of self-awareness.

In section 7.4, I answer a prevalent criticism of the Sartrean approach: that it is committed to a form of presentist-voluntarism. Against this understanding, I show that our possibilities and commitments are dynamic and open and can even elude more direct, explicit forms of awareness. Accordingly, by responding to this charge of voluntarism, I emphasise the aspects of Sartre's work that are distinct from the assumptions traditionally associated with Rawlsian liberalism, but that are nevertheless essential to understanding the subject's normative experience.

Finally, in section 7.5, I examine the Sartrean notion of responsibility, responding to concerns from Detmer (1986) and Warnock (1966) as I do so. Again, this serves to clarify how different forms of self-awareness maintain a normative nature, whilst also dispelling the more extreme views that are often wrongly associated with Sartrean philosophy. Throughout this chapter, I show that Sartre's account of a non-localised and pre-reflective form of agency is phenomenologically plausible. Additionally, given my arguments here and the criticisms of the previous chapters, I situate this account of pre-reflective agency as directly relevant to justice as fairness.

7.1. Phenomenology, Metaphysics, and Agency

In the previous two chapters, I employed Sartre's account of pre-reflective agency to criticise *Theory*'s notion of rational life-planning. Here, I will anticipate a possible counterargument against my approach: the Sartrean account of agency requires a metaphysical foundation incompatible with Rawls's liberalism. As I will explain in a moment, from this Rawlsian viewpoint, political and moral discourse is given priority over, and independence from, metaphysical doctrines (Rawls, 1971, pp.213-214; 453-454). The upshot of this is that even if it is possible to identify true metaphysical propositions *in principle* – and so, presumably, even if my criticisms are well-founded –political agreement must take place *in lieu* of substantive metaphysical premises. As a result, justice as fairness achieves greater levels of parsimony by appealing only to general facts/beliefs about persons and the world. Here, the question is whether Sartre's account of agency fits with the suitably general category of knowledge that Rawls already draws from, or whether it involves commitments incompatible with *Theory*'s approach overall. As Rawls writes:

> We should… note that since principles are consented to in the light of true general beliefs about men and their place in society, the conception of justice adopted is acceptable on the basis of these facts. There is no necessity to invoke theological or metaphysical doctrines to support its principles, nor to imagine another world that compensates for and corrects the inequalities which the two principles permit in this one. Conceptions of justice must be justified by *the conditions of our life as we know it or not at all* (Rawls, 1971, pp.453-454, emphasis added).

To see the reasoning behind Rawls's position here, consider a modified version of Pascal's wager, where society is organised in service to God because it might be the case that such a being exists. Here, the background metaphysical assumption – say, an afterlife of infinite pleasure – is doing all of the heavy-lifting for the conception of justice, acting both as its justification and as the motivation for persons to uphold its institutions. As a result, the Pascalian conception is not likely to generate the kinds of support that justice as fairness strives for. Neither, for that matter, is it likely to fit with Rawls's construction of the original position, or cohere with other general facts about persons and the world that prove relevant to social justice. And yet all of these requirements, from encouraging the appropriate attitudes to fitting with the other facts underpinning Rawls's approach, are crucial to *Theory*'s account of stability:

> The task… is to explain how justice as fairness generates its own support and to show that it is likely to have greater stability than the traditional alternatives, since it is more in line with the principles of moral psychology... Inevitably we shall have to take up some rather speculative psychological questions; but all along I have assumed that general facts about the world, including basic psychological principles, are known to the persons in the original position… (Rawls, 1971, p.456).

Recall from chapter 1 that Rawls's account of stability aims towards the two generative properties of justice as fairness: institutions will be seen as independently valuable and that citizens will want to do their part in maintaining them. According to the analysis above, a conception of justice that achieves these properties indirectly, *via* a metaphysical doctrine, is ruled out; the attitudes of support must be a result of the conception itself, otherwise it is not the agreement, but a dominant end, that regulates the political structure of society. Whilst a conception of justice cannot depend on wishful thinking – recall that stability is a practical problem, which requires certain knowledge about persons and the world to be made available – it also cannot rely on substantive metaphysical viewpoints that may undermine political and moral discourse.

With the above in mind, I will show that a phenomenological approach to agency can be meaningfully teased apart from any particular metaphysical standpoint on freedom. As such, it is possible to discuss the experience of agency without committing to the existence (or, for that matter, non-existence) of free will as a metaphysical fact. This is not to say that Sartre's (2018, pp.798-808) account of agency precludes metaphysics entirely – one *could* argue that Sartre's account entails the existence of a metaphysical form of free will – only that it is possible, both theoretically and in practice, to separate the two in order to focus solely on the relevant phenomenology. Here, I have adapted a hypothetical scenario initially provided by Gerassi (2009, pp.72-73) – and endorsed by Sartre in their one-on-one interview –to reinforce and clarify this distinction between a metaphysical and phenomenological thesis:[52]

> Imagine Matt – through the miracle of science, we know everything about him, and we can predict exactly what he will do next. Suppose he goes to the cinema and is given the option of seeing a film noir or a superhero movie. He chooses the film noir and, after it is finished, he complains: 'that was a bit boring, I should have seen the superhero movie instead'. We knew, by pouring over the data and putting it through our algorithms, that Matt was going to choose the film noir – he was also aware of our research. However, when all is said and done, Matt reproaches himself for not acting differently, not the data or our scientific curiosity; in being bored by avant-garde cinema, he feels as if he should have gone to see the superhero flick.

In this scenario, a broad form of metaphysical determinism is assumed to be correct. However, the Sartrean conception of agency can be teased apart from this metaphysical thesis by concentrating on the phenomenological viewpoint: although Matt is situated within an objective world of facts and inevitabilities, his experience takes for granted frameworks of possibilities, which are made descript by various modal verbs, e.g. I 'can', 'may', and so on. Note also that Matt's reflexive assessment of his situation – i.e. 'I should have seen the superhero move instead' – is not the 'origin' of this sense of agency but an implicit recognition of it. Descriptively, Matt reflects on his past and holds himself responsible; the organisation of his situation through the modality of "*I can*" is retrospectively affirmed by Matt's assertion that he *should have* seen the superhero movie instead. On the face of it, then, a pre-reflective sense of agency is required to provide a complete phenomenological profile of Matt's experience, which is the case whether or not determinism (or compatibilism, fatalism, etc.) obtains as a metaphysical fact.[53]

Subsequently, Rawlsians can include phenomenological observations within justice as fairness by treating them not as additional metaphysical commitments, but as analogous to the psychological notions that are already deeply embedded within Rawls's framework. To appreciate this point, recall that Rawls attempts to respond to the problem of stability by

[52] I have offered my own version of this example to take it out of its conversational setting and establish more of a focus on the present topic.

[53] I will return to this claim – that a pre-reflective sense of agency is crucial to the subject's overall phenomenology – in section 7.2.

demonstrating that persons can affirm justice as part of their conception of the good. Yet to accomplish this task, *Theory* must define persons, their good, their status as agents, how they organise their aims, and so on; importantly, none of this happens in a vacuum, rather Rawls draws on a diverse range of thinkers in psychology, sociology, and philosophy, to support his conclusions (chapter 3, which looks at the background of rational life-planning, demonstrates this directly). As Rawls (1971, p.51) himself puts it, "the analysis of moral concepts and the *a priori*…is too slender a basis [for a conception of justice] …moral philosophy must be free to use contingent assumptions and general facts as it pleases." Crucially, it is on these terms, and not in light of some broader metaphysical doctrine, that Sartrean phenomenology is relevant to justice as fairness. Like Rawls, my approach is concerned with the 'conditions of life as we know it', rather than substantive *a priori* assumptions.

Throughout this thesis, I have not advanced (and will not advance) a particular metaphysical doctrine; instead, I have utilised a phenomenological analysis of pre-reflectivity to interrogate the pre-existing notions of agency and goodness within justice as fairness. Moreover, I have done so in a way equivalent to Rawls's use of psychology and other relevant disciplines. Instead of extending justice as fairness towards metaphysics, I have deepened the fountain of resources from which it already draws. In the next section, I shall demonstrate that without a pre-reflective sense of agency experience as a whole lacks structure. Hence, I will situate Sartre's account of agency as a plausible phenomenological thesis, thereby consolidating the claim that Rawls ought to utilise it within the argument from congruence. Thus, I will complete my analysis of pre-reflectivity in the remainder of this chapter, before establishing a Sartrean-Rawlsian approach to congruence in chapter 8.

7.2. Pre-reflectivity, Possibility, and the Structures of Consciousness

In the previous section, I suggested that a pre-reflective sense of agency is necessary to provide a complete phenomenological profile of the subject's (e.g. Matt's) experience. I will now support that claim by demonstrating that the subject's pre-reflective sense of possibilities is crucial to the overall structure of their experience. To achieve this, I shall employ Ratcliffe's (2017; 2024) argument from 'We are our Possibilities: From Sartre to Beauvoir to Løgstrup', which is developed more fully in *Real Hallucinations*. Here, the first step is to recognise that there are a range of different ways in which I can be intentionally aware *of* something; I can imagine a lion, perceive a table, remember a position in chess, and so on. In each of these standard cases, I am unproblematically and implicitly aware of my intentional state under a particular modality. My perception of a lion, for example, seems to me importantly different from imagining a lion, whilst my recollection will seem different still. I take it, then, that a person can be intentionally aware in a number of different ways, and more specifically, that they have a pre-reflective sense of being aware in this, rather than that, intentional state.

As an intermediary step, consider examples where the subject's mundane experience is no longer organised *via* agent-relative possibilities. Here, again, I will borrow Sartre's examples. My intention is to anticipate my concluding point by highlighting how such experiences appear structureless throughout, encouraging the analogy to fantastical or dream-like states:[54]

First, from the *War Diaries*:

> Let a genie give me power to realize my desires there and then, and at once I fall asleep – being unable to hold them off, to *prevent* them from being realized (Sartre, 1999, p.38, original emphasis).

Second, as a reoccurring idea in *Being and Nothingness*:

> …if in fact the ends that I am pursuing could be attained through a purely arbitrary wish, if it were enough to wish for something in order to obtain it…I would never be able to distinguish within me a desire from a volition, or a dream from an act, or the possible from the real…since in order to actualize something it would suffice to conceive of it (Sartre, 2018, pp.437-438).

And again, in *Being and Nothingness*:

> If it is sufficient to conceive of something for it to be conceived, I will find myself suddenly plunged into a world resembling the dream-world, where the possible is no longer in any way distinct from the real (Sartre, 2018, p.629).

Throughout these quotes, Sartre is discussing what I will call '*intentional blurring*'. Descriptively, intentional blurring occurs when one's sense of being in a particular intentional state breaks down, and it becomes unclear as to whether one is perceiving *x*, or remembering *x*, etc. Accordingly, these examples show that without agent-relative possibilities – think of those modal verbs mentioned earlier, like 'can' and 'may' – the subject no longer maintains the intimate and unproblematic sense of being in a particular intentional state. In the example from *War Diaries*, the genie can grant my desires without the intervening processes (i.e. a course of action which appears to me as *only possible*) that ordinarily lead to their realisation;

[54] In addition, these examples provide a useful touchstone for some of the relevant phenomenology.

as a result, this loss of agentic possibility culminates in the blurring between ‘desires’, ‘volitions’, ‘the real’ and so on. This is the position that I will consolidate.

The next step is to note that one normally experiences the world as soliciting various demands or requirements without the intervention of a deliberative will. “We do not generally act on the basis of preceding mental states that are experienced as internal to ourselves” (Ratcliffe, 2024, p.3). Instead, as I will unpack in section 7.3, one is actively responding to salient features of one’s environment. This is a complicated, multifaceted, process. As I demonstrated in chapter 5, the projects that I am pursuing now are not isolated. Rather, they depend on intricate hierarchical structures of values and commitments (Ratcliffe, 2024, p.3). Likewise, just as specific objects in my environment are synthetically organised (say, my keyboard and computer are organised as I write), my projects are embedded within other projects, and call for some wider situation. In this sense, acting in one’s environment involves reaching out in a way that can sometimes feel groundless. We act by responding to what is “lacking” in our situation, and this process continues without the intervention of a deliberative consciousness (Ratcliffe, 2024, p.3; Sartre, 2018, p.570).

From this step, now recognise that there are various *kinds* of possibilities: possibilities that belong to me, that concern objects or the relation between them, and so on. This includes possibilities that are “phenomenologically irreducible” (Ratcliffe, 2024, p.4). A raincloud, for example, does not first include the perception of a cloud and then imagined rain (Ratcliffe, 2024, p.4; Sartre, 2018, p.153). Instead, I experience a unitary phenomenon, wherein the object surpasses itself towards its possibilities. Hence, Sartre (2018, p.153) describes possibility as “a concrete property of realities which already exist.” Here, the point is that my experience involves a multitude of different kinds of possibilities, some of which are given irreducibly with the objects of experience. More forcefully, and in line with the previous step in this argument, “all of our experiences are imbued with a sense of the possible” (Ratcliffe, 2024, p.4). These possibilities are not all agency-related; rather, some possibilities concern others, objects in our environment, and so on.

Finally, note that the experience of possibility is “organised in specific, intricate, and interdependent ways” such that, without a sense of agent-relative possibility (i.e. modal verbs, or the ‘I can’), “experience as a whole would lack structure” (Ratcliffe, 2024, pp.4-5). As with the above points, consciousness involves a complex and intricate structure. Not only do I have a sense of being in a particular intentional state, but I also have an awareness of distinct possibilities within my environment. This includes, as anticipated by my brief description of the active orientation of pre-reflective awareness, the experience of possibilities as connected to my own agency. Crucially, when I experience a possibility through the medium of ‘I can’, I always experience it “*as a possibility*, rather than as something inevitable” (Ratcliffe, 2024, p.4). There is not only a sense of lacking, but also a *lacuna* between “what is now the case and how I act” (Ratcliffe, 2024, p.5).

In a similar sense, whilst I ‘know’ that if a rock loses its perching on the side of a cliff, it will fall to the ground, the ‘knowledge’ (in a non-judicative and pre-thematic sense) of my own possibilities appears markedly different. Such possibilities *depend on me* – as I will explain in section 7.3, they are requirements that call to be acted upon. And yet, other possibilities remain available. Faced with the possibility of kicking the rock off the cliff, I feel as if I could stop, even if such a possibility is not immediately salient (Ratcliffe, 2024, p.4). At the same time, in kicking the rock off the cliff, I might think it is possible that it falls by the roadside rather than into the river. Yet this epistemic uncertainty is distinct from possibilities associated with my agency; “we experience our own possibilities in the guise of ‘I can’ rather than ‘it might happen’” (Ratcliffe, 2024, p.4). This is all to say that the world is organised according to a range of different possibilities, which includes a pre-reflective sense of ‘I can’. Moreover, if this sense of ‘I can’ “could not be distinguished phenomenologically from…other possibilities” then experience would be rendered structureless (Ratcliffe, 2024, p.5). There are meaningful phenomenological differences between the forms of possibilities that make up one’s experience, and a sense of *my possibilities* is required to complete that profile.

Another way of cashing out this argument is by recognising that part of what makes an intentional state consist in a particular modality *is the kinds of possibilities that it affords*. To borrow from Sartre’s example above, if a genie gave me the powers to realise all my desires there and then, I would simply ‘fall asleep’. The reason for this is that I have gotten what I wanted, without the possibility of not wanting it; that is, *without it being a possibility*. Intentional blurring occurs when experience lacks a sense of ‘I can’. Phenomenologically, a perceptual act is not just distinct from an act of recollection because of its correlative object – roughly, a perceived object rather than an imagined or memorised one. Rather, this distinction also concerns the kinds of possibilities presupposed by each act. The upshot of these observations, when taken together, is that agency plays an important role not only in constituting what my experience is like, but also *in restricting the kinds of experiences that I can have*. Intricate and unthematized frameworks of possibilities are given part and parcel with one's intentional awareness. Without these complex frameworks, the cohesion of experience erodes.

As a final point, recognise that this pre-reflective sense of agency must consist in a *non-positional* consciousness. To clarify, this point is worth emphasising because – as I will discuss in section 7.4 – Sartre's account of agency is often wrongly associated with discrete, positional, acts of consciousness. Even otherwise faithful renditions of Sartre's work – for example, Eshleman's essay on pre-reflectivity – can lapse into misrepresenting Sartre in this way. To this end, Eshelman (2010, p.39, original emphasis) claims that "for a future action to be existentially possible, I must *be able to* imagine myself as able but not necessarily having to exercise it through a future action." My main worry is that Eshleman's description of pre-reflective possibility necessitates an explicit act of self-reference by explaining agentic possibilities *via* the potential for imaginative acts (i.e. 'I must be able to imagine…'). However, as I indicated in the previous chapter, pre-reflective awareness is not self-referential, but rather it *has to be* what it is aware of. One can see this by recognising that – since the subject maintains an ongoing and unproblematic sense of being intentionally aware in a particular modality – positional consciousness must *already* be structured by a pre-reflective sense of agency. Hence, pre-reflectivity spans across intentional acts and constitutes a structural feature of consciousness as such. In this light, Sartre's description of possibilities cannot consist in imaginative acts, unless pre-reflective experience as a whole is imbued with imagination; a nuance that would require further investigation.

That being said, the claim established in this section is that first-person experience is structured by frameworks of possibilities. Properly speaking, this pre-reflective sense of agency belongs to a non-positional consciousness; since, without it, the barriers between distinct intentional modalities are blurred or rendered inoperative. As a result, the Sartrean account of agency is necessary to describe the subject's overall phenomenology. On this view, the very mode of being for unreflective consciousness is to be spontaneously and continuously projected towards its possibilities. To be *is* to commit oneself; one *exists through* one's possibilities. However, the idea that one has to be one's possibilities may still seem inaccessible, and one may even wonder what such an awareness involves; in section 7.3, I shall shed light on this complexity by describing the ways in which pre-reflective awareness is both active and world-directed. In doing so, I will position pre-reflective agency as a basic, unitary, phenomenon.

7.3. A Basic, World-Directed, Form of Agency

At this point, I will combine the conclusions of the previous two sections, to highlight their shared implications. As I recognised in section 7.1, Rawlsians may criticise my approach for its putative dependency on metaphysics. My response to this, however, showed that it is possible to tease apart a phenomenological thesis about agency from a metaphysical commitment to freedom. Building on this in section 7.2, I demonstrated that my view concerns a phenomenological claim about the overall structure of experience. Specifically, that pre-reflective consciousness involves a sense of agency; that is, a non-positional awareness of frameworks of possibilities. Drawing on Ratcliffe, I supported this claim by arguing that, without this pre-reflective sense of agency, the distinction between different intentional modalities breaks down. In doing so, I have situated Sartre's account of agency as a plausible phenomenological thesis – or, in Rawlsian terms, an appropriately general fact about persons – which can be included in the range of information that *Theory* utilises. If my criticisms in the previous chapters demonstrate the need for viable alternatives (*via-a-vis* the underpinnings of Rawls's argument from stability), what the sections above clarify is that the Sartrean approach is indeed an attractive replacement.

With that clear, I will now investigate an important aspect of Sartre's view; namely, that the frameworks of possibilities that structure one's awareness also correlate to how the world presents itself as practically organised. This point requires further development for three reasons. First, as I mentioned in section 7.2, it is difficult to conceptualise what Sartre means with his claim that we are our possibilities. Acknowledging that being one's possibilities involves actively responding to environmentally-solicited demands may resolve this obscurity. Second, in chapter 8, my reconstruction of the congruence argument establishes various links between the Rawlsian organisation of society and normative forms of self-awareness. In light of this, I want to anticipate my analysis from a straightforwardly phenomenological perspective, by emphasising the inseparability of one's possibilities from the world. Last, Sartre has been criticised for advancing an abstract form of 'absolute freedom' whereby consciousness, instead of emerging with the world, imposes itself upon it.[55] In this way, demonstrating that pre-reflectivity is a unitary, world-orientated, phenomenon, can serve to weaken this charge considerably.

[55] I shall discuss this issue again in the next section, the examples that I use there are from Grossman (1984) and Smith (1998).

With these points in mind, consider Sartre's (2018, pp.75-82; 187-188) example of spontaneously writing a sentence. As I write now, I do not need to be explicitly aware of each letter that I put down, just as I do not need to be explicitly aware of my fingers stretching along the keyboard as I type. Instead, my awareness is placed on the world as it appears '*out there'* – the sentence that I'm forming, the words that I use, the keyboard that I type on; these are all part of the same concrete relation. But what gives the world this structure if not my possibilities? I do not feel intractably compelled to write this sentence or that its precise formulation is inevitable. As my example in the previous section demonstrates, my non-thetic awareness of this sentence does not *realise* it or else – to borrow Sartre's imagery – consciousness would 'fall asleep'.[56] Instead, the sentence stands as a regulative standard for my action (Sartre, 2018, p.74). In sitting in front of my keyboard, I am confronted with the possibility of typing. Whilst I am typing, I surpass this sentence towards the meaning of the chapter as a whole, and so on. My point is that my possibilities push beyond themselves, towards a future world, and at values and ends that regulate my action; a "possibility…cannot be limited to being only a thought, as a subjective mode of consciousness", but rather, "can only be established externally" (Sartre, 2018, p.155).

Building on this, my second point is that the possibilities that I *am* are ordered within an evaluative hierarchy; there is a *form* to my projects, where values and meanings become embedded within complex, interwoven, structures (Sartre, 2018, p.607). Practically speaking, if this were not the case, then I would experience my actions as having no distinguishable pattern or direction. However, whilst it is possible for me to stop typing, this nevertheless feels to me more difficult than continuing to write this sentence. Descriptively, I exist in a world where my current actions – and the kinds of possibilities and values that they afford – not only push towards a future, but also exercise a kind of pull on me correlative to my intentions. As Sartre (2018, pp.279-280, original emphasis) puts it, these possibilities are not "disclosed as having to be actualized *by me*" – that is, through an explicit act of self-reference – but instead manifest themselves "to unreflected consciousness by a direct, personal urgency that is *lived* as such, without being either connected to a *someone*, or thematized." In this vein, I project myself towards these possibilities by being them, and in doing so, commit myself to them – phenomenologically, this commitment is disclosed *via* a lived form of exigency.

[56] The thetic/non-thetic distinction is broadly similar to the positional/non-positional distinction. As Webber (2002, p.49) notes, however, there is an operative difference: whilst positional consciousness singles out its object, thetic consciousness characterises the object singled out. One example that Sartre (2018, p.11) gives of non-thetic consciousness is the act of spontaneously counting.

From these two points, it is possible to see why pre-reflective awareness is also a non-thetic and pre-thematic consciousness of the world. First, if my possibles were realisable within a "subjective mode of consciousness", then they would not be *possibles*, and so – as in the previous section – experience as a whole would lack structure. Second, if my possibles were given as completely external to me, then they would not be *my* possibles: one would not be able to explain this sense of agency as a structural feature of consciousness, nor why I experience my possibles through a lived, immediate, form of urgency. In the *War Diaries*, Sartre puts this point using the notions of transcendence, immanence, and ends (ends, like the meanings and values that I have been discussing here, of a non-positional consciousness):

> For an end can be neither entirely transcendent with respect to the person who posits it as an end, nor entirely immanent. If transcendent, it would not be *its* possible. If immanent, it would be dreamed but not willed. The linking of the agent to the end thus presupposes a certain bond of the 'being-in-the-world' type, in other words a human existence (Sartre, 1999, p.107, original emphasis)

Focusing on the language of possibility, it seems that I can only pursue my projects within a world that demands something from me; a world that, so to speak, does not meet my expectations. In this way, I do not experience my actions as being elicited by states internal to mind, but as responding to the requirements of a resistive world (Ratcliffe, 2024, p.3; Sartre, 2018, p.279). By acting, I realise my possibilities as I go along, causing more possibilities to arise in my wake. The crucial point is that I am not making decisions within an isolated consciousness, *but actively responding to what is lacking in my environment* (Ratcliffe, 2024, p.3, emphasis added). In turn, my experience of what is lacking in the world depends on my projects, just as a blank page is empty of the words that I am yet to write; as Sartre (2018, p.285, original emphasis) puts it, "*absence* is disclosed within the world as a being to be actualized, in so far as this being is the correlative of the possible-being *that is missing from me*."

Given the tendency of scholars to criticise Sartre for advancing an abstract and unencumbered account of agency, I want to emphasise that both pre-reflective frameworks of possibilities and the world are given as a basic, spontaneous, unity. To this end, it is worth considering the drawbacks associated with the metaphor of 'superimposition' that Neuber (2021) employs in their – again, otherwise generally faithful – analysis of pre-reflectivity. As Neuber (2021, p.141, original emphasis) puts it, pre-reflectivity involves the "*superimposition* of the intentional object of that future consciousness onto the intentional object of the current consciousness", which then generates structures that are "intrinsically motivational." Whilst this reading offers a relatively close description of some of Sartre's ideas (e.g. the temporal aspects of consciousness), the issue is that it treats existential possibilities as something over and above the world, as abstractions that are imposed upon it *via* one's intentional awareness. Yet, this misrepresents the issue. The world, as a condition of its intelligibility, is *already* practically organised by a pre-reflective consciousness that is its own possibilities; whilst my possibilities are given spontaneously as structures of value, meaning, and exigency *within* the world. In this way, non-positional consciousness is not aware of its possibles *via* an act of self-reference nor by adding its possibles onto the world; instead, it only comes to know them through the world as it appears 'out there':

> Thus, in what we may call the 'world of immediacy' — given to our unreflective consciousness — we do not appear to ourselves first, in order to be thrown *subsequently* into our undertakings. Rather, our being is immediately 'in situation', which means it *arises* within our undertakings… We are therefore revealed to ourselves within a world populated by requirements, in the midst of plans that are 'in progress'… (Sartre, 2018, p.78, original emphasis).

In this way, my choices do not 'impose' a structure on the world because they arise as a *spontaneous unity with the world*. Accordingly, the world is organised according to the things that matter to us, with degrees of salience and equipmental frameworks that make one's environment intelligible (Ratcliffe, 2024, p.3). As Sartre (2018, pp.279-280) writes, "this equipmentality is not subsequent, or subordinated, to the structures that were previously pointed out [i.e. the lacks within the one's environment and my possibilities]: in one sense, it presupposes them, and in another sense, they presuppose it." Hence, the projects that I pursue pre-reflectivity, the demands that they give rise to, and the practical organisation of my environment, are all contemporaneous with one another. To speak of anteriority, interpolation, or abstraction, would be to muddy the waters significantly.

7.4. Voluntarism, Responsibility, and Fundamental Projects

So far in this chapter, I have accomplished three overarching aims. First, I have defended the relevancy of Sartre's account of agency to the Rawlsian argument from congruence by drawing a distinction between metaphysical theses and general facts about persons and the world. Second, I have provided independent support for Sartre's account of agency by arguing that it is a structural feature of consciousness, thus defending it as phenomenologically plausible. Finally, I have highlighted an important feature of Sartre's account – that one's sense of agency is world-orientated – so as to anticipate my contribution in chapter 8, clarify some of the relevant phenomenology, and dismiss a prevalent misinterpretation of the Sartrean approach. I will now consolidate two major points in this thesis. First, that Sartre's account of agency is ongoing and non-localised. The upshot of this point is that Sartre's view is a meaningful alternative to voluntarist conceptions of agency, and thus to Rawls's original approach. Second, I will clarify how pre-reflectivity invokes the norm of responsibility, both to support my arguments earlier in this thesis, and again, resolve an important point of contention within Sartrean scholarship.

In chapter 5, I demonstrated – *via* his notions of rational life-planning and self-reproach – that Rawls maintains a rationalistic and voluntaristic account of agency. I also argued that such an approach mislocates the source of our agency. In the context of this more detailed account of pre-reflectivity, I want to reaffirm and expand on this claim. Though this will serve to clarify my argument overall, it is also necessary because the charge of 'staccato voluntarism' is one that has often – though erroneously – been levelled against Sartre (Webber, 2012, p.327). For example, Grossman (1984, p.263) describes Sartre's position as "so absurd that it is hard even to formulate it" since it maintains that "man is free because he can make himself be whatever he wants to be, merely by looking at himself in a different way." Whist Smith (1998, p.30, original emphasis; 2011, p.327) claims that – on this Sartrean picture – "we have complete autonomy over our mental states…though we are not determined by the external world, we are determined… by our private, mental goings-on."

In this way, it would be a meaningful problem – both for my discussion in chapter 5 and my arguments in this chapter – should Sartre be committed to voluntarism. After all, this competing reading of Sartre, if correct, would cast doubt on my phenomenological claims thus far, and (when taken further) bring into question my bracketing of metaphysical issues in section one. Hence, considering Grossman's (1984, p.263) and Smith's (1998) misinterpretation of Sartre remains instructive, especially since the former attributes this 'absurd' view to *Being and Nothingness* because of its apparent endorsement of a localised form of voluntarism:

> Sartre's contention is…that there are no past mental states, but only present mental states which may be about other mental states which, for some reason, appear to be past mental states (Grossman, 1984, p.263).

In this reading, Sartre views agency as *individuated mental acts* wherein the present consciousness is inexplicably and unrestrictedly free. Note also that this criticism shares some resemblance with the metaphor of superimposition that I discussed above, as both arguably involve the interpolation of an undetermined consciousness onto the world. Grossman and Smith, however, take this issue further. Instead of making a descriptive error, they are advancing the stronger claim that Sartre maintains a volitional, gratuitous, notion of agency.

However, a more careful reading of *Being and Nothingness* reveals that Sartre maintains an account of agency that should not be understood through deliberate acts of will, but *via* complex, non-localised, and pre-reflective frameworks of projects. Crucially, there are patterns and (sub)structures to these projects. As I will show, one's previous commitments become ingrained in hierarchies of values, which can prove just as salient – often more so – than our capacity to freely override them. That is, for Sartre, there is a *price* to freedom. An example taken from *Being and Nothingness* supports this point directly and undermines the concerns advanced by Grossman and Smith:

> I have gone on an excursion with some friends. After several hours of walking, I am growing tired, and eventually my fatigue becomes oppressive. At first I resist and then suddenly I let myself go: I give in; I throw my bag down on the side of the road, and I drop down beside it. I will be reproached for my action, with the implication that I was free, which means not only that nothing and nobody determined my action, but also that I could have resisted my fatigue, done as my fellow-travellers did, and waited for my rest until we reached our stop. I will defend myself by saying that I was too tired. Who is right? (Sartre, 2018, pp. 594-595).

In this example, the first thing to note is that Sartre is not concerned with basic physiological reactions/states. As his brief contrast demonstrates, the hiker's fellow-travellers – who are in similar states of fatigue – are capable of continuing their expedition. Hence, I take it that Sartre (2018, p.596) is not concerned with fatigue seen exclusively as "a factual given." Instead, the hiker's fatigue must be understood as *a particular response* to this facticity. To see this, recognise that the hiker was 'fatigued' before this became thematised within their awareness, which was otherwise active and world-directed; "correlated with this non-thetic consciousness, the roads are revealed as interminable, the inclines as *more difficult*, the sun as more intense, etc." (Sartre, 2018, p.595). However, an important change comes about when the hiker attempts to retrieve and thematise their non-thetic bodily awareness through a positional act of consciousness: their fatigue is now *too much*, and their body feels too weak to continue walking. In this way, Sartre's (2018, p.596). focus is on the hiker's choice of collapsing into this state of affairs, of losing trust in their legs, i.e. of directing their consciousness towards their fatigue "in order to live it":

> I am still not *thinking* my fatigue…there comes a moment, however, when I try to consider it and to retrieve it … a reflective consciousness is directed on my fatigue in order to live it and to confer upon it a value and a practical relation to myself. It is only at this level that my fatigue can appear to me as bearable or as intolerable (Sartre, 2018, p.596, original emphasis).

Here, the crucial point is this: *for voluntarists, the explanation above would prove sufficient*. The hiker reflected on their situation and, through a reflective consciousness, chose to give in. To borrow Grossman's (1984, p.263) initial phrasing, the hiker made his fatigue too much "merely by looking at himself in a different way." Understood in this light, and on the terms initially set by the example, the fellow-travellers were right to cast aspersions. Just as the hiker elected to throw their bags away and sit down, he could have chosen to continue down the path. At least in Grossman's interpretation of Sartre, the matter is settled; the hiker's sense of agency could be functionally isolated to the actions of reflective consciousness, wherein their fatigue explicitly appears unbearable. *However*, this would only be true if the hiker's choice resisted a deeper level of analysis. Here, the implication is that, if the hiker's response can be elaborated on, then "it might be explained within the perspective of a wider choice, in which it is integrated as a secondary structure" (Sartre, 2018, p.596).

To see that the hiker's fatigue does permit of further analysis, recognise that someone similarly situated to the hiker could have acted differently on the basis of their broader projects; "thus my companion lives his fatigue within a larger project that is a trusting surrender to nature…it is only in and through this project that his fatigue can be understood and can have a meaning for him" (Sartre, 2018, p.597). In this way, the fellow-traveller experiences their fatigue as an invitation to continue because it is grounded in their trust and respect for nature. Adopting the same form of regressive analysis, it is also the case that the hiker's fatigue ought to be interrogated further by exploring it within the context of their relationship with their body. Seen in this light, the hiker's way of responding to their fatigue is embedded within a project to reappropriate their "body and…presence to the world through the other's acts of looking" (Sartre, 2018, p.599). *Mutatis mutandis*, the same applies to the relationship between seemingly isolated moments of choice and agency as a feature of non-positional consciousness; independent actions emerge in the context of broader commitments, without their necessarily being a linear or deterministic connection between them.

With this in mind, whilst voluntarists will maintain that the hiker must be reflectively or deliberatively aware of their choices in order to make them, Sartre's (2018, p.604) position is that whilst "we are…conscious of them [our projects] …this consciousness itself must be limited by the structure of consciousness in general." This entails the hiker cannot "gratuitously" – that is, by a detached act of consciousness – change their response (Sartre, 2018, p.607). Given that agency is non-localised, *the choices that one makes reverberate throughout the whole structure of one's projects*. For the hiker not to collapse into their fatigue, they would have to alter their project of bodily reappropriation; a project that, as I will discuss later on, integrates the other projects that they can pursue. Grossman and Smith are wrong to claim that the agent can choose to alter their situation *via* a gratuitous, unencumbered, volitional act. In fact, Sartre points this exact error out within the *War Diaries*, when he highlights the fact that our commitments are sustained within a non-thetic consciousness, without having to be *explicitly* or concurrently realised by each individual act:

> If I examine myself at this moment, I know and have proof that there exist in me a certain number of full, effective willings….yet these are not empty volitions, nor are they full volitional acts which once existed and are now lying dormant…does the error, perhaps, not come from the fact that the will is usually considered as an act of consciousness, brief and localized in time?...Which comes down to the same thing as saying that consciousness – usually not voluntary – may in certain conditions take on a voluntary structure (Sartre, 1999, p.36).

In this way, the practical organisation of the world with its values and demands correlate to hierarchies of values within a non-positional consciousness; crucially, this is not merely the context for agency but its very mode of persistence. If I *am* my possibilities, then I sustain them throughout my being, and not in isolation from one another, but as a totality.

7.5. A Non-Localised Sense of Agency, A Global Form of Responsibility

Having argued that Sartre advances a non-localised and world-directed account of agency, I shall now investigate his claim that this sense of agency involves a globalised form of responsibility. As Sartre (2018, p.718) puts it, "the essential consequence of our previous remarks is that man, being condemned to be free, carries the weight of the whole world on his shoulders: he is responsible for the world and for himself, as a way of being." Importantly, this position has been criticised by the likes of Warnock (1966) and Detmer (1986), who suggest that the notion of responsibility seems to be incompatible with Sartre's view as a whole. That is, to be responsible, one has to *know* or *intend* what one has done. The implication here is that we are only responsible when we have a higher-order awareness of our actions, which cannot be accounted for within a pre-reflective sense of agency. As Warnock and Detmer write:

> One of the main difficulties in understanding Sartre's theory of human action is to make out where he wishes to say that we are responsible…He

> certainly thinks…that we confer certain values upon things…But do we necessarily *know* that we are doing this? (Warnock, 1966, p.114, emphasis added).
>
> Sartre's theory of responsibility is, characteristically, extreme… his purpose is to show that when what we allow to happen is bad…not only are we responsible for this badness, but…we are *guilty* in the face of it, that we are to *blame* for it (Detmer, 1986, pp.169-170, original emphasis)

The question, then, is whether the norm of responsibility can properly belong to a non-positional consciousness. To simplify this discussion, I will answer a modified version of this problem; that is, in what way(s) does the Sartrean view of pre-reflective agency involve the norm of responsibility? Hence, I am more concerned with *how* Sartre's view holds that we are responsible globally, rather than whether or not it ought to. [57]

To grasp the relationship between total responsibility and pre-reflective agency, recognise that this sense of responsibility must be – *pace* Detmer –prior to explicit and isolated feelings of guilt and blameworthiness. To this end, note that if such feelings were required for Sartre's account, then responsibility would consist in an intentional relation to a previous action. Yet if this was the case, I would act first and *then* be intentionally aware of my action through a particular form of affective disclosure (e.g. guilt or shame) – the result being a modified version of staccato voluntarism; on which, instead of having to make decisions at each moment, I would have a sense of responsibility *after* each act. Put succinctly, this would be a deeply fractured notion of global responsibility, which would consist of contiguous moments of feeling responsible. Against this, responsibility ought to be understood within the context of one's pre-reflective sense of agency; as Sartre seems to suggest in the following:

> Moreover, this absolute responsibility is *not an acceptance*: it is merely what the consequences of our freedom logically demand (Sartre, 2018, p.718, emphasis added).
>
> I am abandoned in the world, where this does not mean that I linger, passive and forsaken, within a hostile universe, like a plank floating on water, but, on the contrary, that I find myself suddenly alone and without aid, committed within a world for which I bear the entire responsibility, *without being able – whatever I do, and not even for an instant – to separate myself from this responsibility* (Sartre, 2018, p.721, emphasis added).

[57] One should keep in mind here the connection with voluntarism. For example, if the Sartrean notion of global responsibility requires positional acts of consciousness, then this may introduce voluntarism into his theory of agency as a whole. Here, my intention is not to provide an exhaustive description of Sartrean responsibility, but rather quell the concerns of those already sympathetic to Sartre, such as Warnock and Detmer.

> I never encounter anything but myself and my projects, so that in the end my abandonment…consists merely in my being condemned to be fully responsible for myself (Sartre, 2018, p.721).

With these quotes in mind, I think there are three main ways that the Sartrean notion of responsibility ought to be cashed out. First, the experience of responsibility is a component of being *committed* to a particular project. As I demonstrated in section 7.3, our possibilities exercise a kind of pull on us; through acting, the commitments that we have already made are integrated in increasingly entrenched and valued projects. This implies that we do not experience our actions neutrally. Rather, as Sartre (2018, p.719, original emphasis) puts it, "everything that happens to me is *mine*." This is to the point that, if I attempt to distance myself from my possibilities – "to make myself passive within the world, to refuse to act on things and on others…" – then I am still undertaking an evaluative stance that still invokes the norm of responsibility; my refusal *is* my commitment (Sartre, 2018, p.721).

Second, at the bottom of this Sartrean picture, consciousness is understood to be thoroughgoingly active, thus entailing a unique form of responsibility. To this point, recognise that – though there are patterns to my projects – the values and commitments that constitute them are still only held in "suspense" rather than concretised (Sartre, 1999, p.133; 2018, p.652). There is no inertia, no certainty that my commitments will be realised, instead consciousness is always in the process of renewing, reaffirming, acquiescing, and rejecting its possibilities (Thomas, 2010, pp.161-180; Moran, 2001). To use the earlier example, the hiker could have stopped at any point, such that the possibility of reassessing their commitment remained constant. The norm of responsibility arises because consciousness is, in many ways, self-motivating. I experience my freedom as the "discontinuity" between the situation, my motivations, and the act; I *could* write this sentence, it *would be* the case that this chapter is complete, etc. (Sartre, 2018, p.76; 1999, p.134). Hence, in the 'gap' between the motive and the act, I find myself to be responsible:

> In…any act… the possibility remains of calling the act into question, in so far as my act points towards more distant and more essential aims, along with its ultimate meanings and my essential possibilities (Sartre, 2018, pp.75-76).

This leads to the final way that Sartre's account of agency invokes the norm of responsibility: I am responsible for the overall structure that my life takes, with the beliefs, dispositions, actions, and patterns of behaviour that I affirm and practice over time (Ratcliffe, 2024, p.5). To see this, consider the earlier point that one's possibilities call for further elaboration through broader and more complex projects. For example, the hiker's reaction to their fatigue is not derived from their physiological state but from their relation to their body; the significance of this sentence is not its constitutive words, but its place within the work as a whole; and so on. Yet, in tracing these projects back further and further – in line with a non-localised account of agency – one finds that the whole structure proves groundless; though they permit deeper levels of interrogation, there is no ultimate foundation to our projects. Instead, at the bottom of it all, *one simply has to choose oneself within the world*:

> In truth, the meaning of all these minor passive expectations on the part of reality, of all these banal and everyday values, derives from an initial project of myself, which is akin to my choice of myself within the world. (Sartre, 2018, p.79).

Hence, the final sense in which Sartre takes pre-reflective agency to invoke the norm of responsibility is that one is responsible for one's "fundamental project".[58] That is, a project that integrates and gives shape to all of the projects that one can pursue. One's possibilities can all be drawn back to explain one's more fundamental project, whilst one's fundamental project renders one's secondary projects intelligible. The crucial point is this: it is the structure as a whole, the unification of my projects within an organised movement, that I am responsible for. Just as my commitments become entrenched over time, I also cannot go back and change the whole thing at a whim. Instead, I have to undertake processes of interrogation and reintegration of my commitments, whilst occasionally – in extreme situations – I have to reassess the foundations of my life in a way that feels baseless and sudden. In such situations, it is the total structure that is at odds, and one only has oneself to gamble with; "I could have done otherwise, agreed: but *at what cost*?" (Sartre, 2018, p.595, original emphasis).

With this in mind, I can respond directly to Warnock's and Detmer's concerns, i.e. that Sartre utilises a notion of responsibility incompatible with his account as a whole. As my analysis has demonstrated, Sartre takes responsibility to be a feature of non-positional consciousness. Instead of requiring explicit knowledge, as Warnock and Detmer suggest, Sartrean responsibility involves three interconnected ideas about the nature of (pre-reflective) commitment. First, consciousness exists through commitment, meaning that my projects are

[58] I will discuss this notion again in the next chapter. It is worth noting, however, that Sartre (2018, p.79; 436; 602); uses the notions of "fundamental", "initial", "global", and "ultimate" project throughout *Being and Nothingness*. There are distinctions between these notions, though they are occasionally unclear or underdeveloped. Given that I have a specific use-case in mind, I shall stick with the one that conveys my point best. Here, and in chapter 8, I adopt the term "fundamental project" because I want to emphasise how this broader project serves to shape and integrate the other projects that one can pursue.

experienced *as mine*. Second, all of one's mental states are accounted for *via* the dynamics of affirmation, acquiescence, rejection, and so on. Third, all of one's projects are integrated within a fundamental project or one's choice of oneself within the world. Understood in this way, my experience is structured by a pre-reflective sense of agency in a world that demands something from me; only I am responsible for the way that I respond. In chapter 8, I will utilise the arguments made thus far to establish my positive contribution: Rawls's argument from congruence reimagined according to the Sartrean account of pre-reflectivity.

Chapter 8

Reflection, Fundamental Projects, and Practical Self-Conceptions: A Sartrean Sketch for An Alternative Congruence Argument

8.0 Introduction

As I clarified at the start of this thesis, my position when it comes to Rawls's account of stability in *Theory* is broadly reconstructive. So far, however, I have been challenging Rawls's primary response to the problem of stability: the congruence argument. When understood as the coherence between the collective view of justice and the individual pursuit of a good life, the congruence argument must adequately articulate the individual perspective; yet, as I have shown, it is precisely on this issue that *Theory* falls short. Rawls's position on rational life-planning does not capture what it means to be an agent, or a unified self, capable of formulating and pursuing a conception of the good. To summarise the critical portion of this thesis within a single claim, Rawls misrepresents the 'individual perspective' by construing it in rationalistic terms. From a Sartrean standpoint, *Theory* privileges reflective forms of awareness, without accounting for its basis within pre-reflectivity.

Whilst I have been largely critical of *Theory* so far, it is also clear that Sartrean phenomenology *is* capable of saying something fruitful about Rawls's argument from congruence. Though initially formulated in rationalistic terms, the operative themes within Rawls's argument – deliberation, agency, self-unity, and so on – are all open to further interrogation and development from a broadly phenomenological perspective. To recognise this, note that I have criticised the working parts of Rawls's argument (e.g. life-planning and objective rationality) without claiming that congruence between goodness and justice is an impossibility. My reconstructive claim is that exploring stability as a practical problem invites phenomenological insight, and it is from this point of view that Sartre can expand on the 'individual perspective' that Rawls's original argument appeals to. In other words, the congruence argument calls for a Sartrean foundation.

Before I begin this reconstruction, however, recognise that it will undercut Rawls's transition from *Theory* to *Political Liberalism*. To explain, Rawls (2005, p.xvi) abandons the argument from congruence in his later work, *Political Liberalism*, because he comes to think it would require a "comprehensive doctrine"; that is, a substantive moral or philosophical doctrine (rather than purely political conception), which all persons must support. Accordingly, the endorsement of a conception of justice on the grounds of a shared comprehensive doctrine, for Rawls (2005, p.xvii), conflicts with the fact of "reasonable pluralism", i.e. that there will be a diversity among person's ends within society, leading to a diversity in the ways in which they make sense of the world. As Rawls writes in the introduction to *Political Liberalism*:

> …The account of stability in part III of *Theory* is not consistent with the view as a whole [justice as fairness] … A modern democratic society is characterized not simply by a pluralism of comprehensive religious, philosophical, and moral doctrines but by a pluralism of incompatible yet reasonable comprehensive doctrines. No one of these doctrines is affirmed by citizens generally… The fact of a plurality of reasonable but incompatible comprehensive doctrines—the fact of reasonable pluralism—shows that, as used in *Theory* the idea of a well-ordered society of justice as fairness is unrealistic (Rawls, 2005, pp.xv-xvii)

More specifically, Freeman (2003, p.304) claims that Rawls's abandonment of part III of *Theory* is due to its Kantian underpinnings: the "unrealistic…[expectation] that citizens in a well-ordered society will all agree on the supreme intrinsic good of autonomy, or even the intrinsic good of justice." Construed in this light, Rawls's original argument presumes that all persons have a form of true nature – a Kantian 'noumenal self', if you like – that they will desire to realise given the constraints of deliberative rationality. As a result, Rawls responds to the practical problem of stability through a comprehensive doctrine; or at least, this is what the later Rawls says of his earlier writings.

I think that Freeman is correct in his analysis. For the most part, the Kantian background to the congruence argument caused Rawls to abandon it. To see this, note that when Rawls (2005, p.xvii) introduces *Political Liberalism*, he does so by purposefully distancing himself from Kant's work: "their [Kant's and Hume's] beliefs belong to what I refer to as comprehensive as opposed to political liberalism." Yet if this is the case, then the congruence argument can be rescued. Rawls abandons the congruence argument in his move from *Theory* to *Political Liberalism* in order to show that the motivational argument for justice as fairness can be established on grounds independent of Kant. However, in taking this approach, he fails to appreciate the possibility that congruence can be pursued without these assumptions. In this way, Rawls throws out the baby with the bathwater (and all because the bathwater was turning the baby into a Kantian, or so he thought). Here, Mendus makes an identical point, that the congruence argument does not depend on adherence to a comprehensive doctrine (e.g. Kantianism for *Theory*):

> Some commentators…[view] the Kantian interpretation and the appeal to congruence as the point at which *A Theory of Justice* goes badly wrong. They therefore urge Rawls to jettison the Kantian interpretation, abandon the aspiration to congruence, and thus render the original theory immune to allegations of comprehensiveness… [I] suggest that that is over-hasty. The sense of justice can be shown to be for the good of the agent who endorses it without degenerating into absurdity and without invoking an objectionably Kantian doctrine of the true self (Mendus, 1999, p.75).

Elaborating on this point further, Mendus claims that the failure to see alternative approaches to congruence arises out of a crucial mistake: thinking of the question of congruence as an *either/or*, i.e. that the agent will either endorse a comprehensive doctrine or not. I agree with her assessment. This ultimatum is false because *it fundamentally misconstrues the effect that justice has on a person's conception of the good*. As in chapter 4, the 'either/or' approach to congruence must assume that a person's good is somehow static and fully-formed; "it [congruence] can be attained only by postulating some objective good, independent of… the agent's actual desires" (Mendus, 1999, p.70). Yet, the agent's good is never completely determined, but is an ongoing and open phenomenon. Correspondingly, to show that justice and goodness are congruent, one must view it as a dynamic issue that concerns the agent directly, *not* attempt to prove the harmony between one worldview (e.g. a comprehensive doctrine) and an objectified version of agent's interests (e.g. objective rationality). The possibility of an alternative interpretation of congruence remains, and the problem of stability must be revisited.

As I illustrate in this chapter, the problem of congruence is *whether justice can be a part of one's life*, and whether the endorsement of justice as fairness is an appropriate response to the conditions of (the well-ordered) society. In both instances, the Sartrean description remains dynamic and precludes any Kantian underpinnings; the congruence argument is a practical response to a practically significant issue, rather than a rationalistic response to an idealised issue. I do not view the agent's interests in static terms, nor will I impose any deliberative procedure over and above the basic claims that we can make *vis-à-vis* the structures of reflective and pre-reflective consciousness. The Sartrean-Rawlsian approach permits of a phenomenological, rather moral or metaphysical, background (recall my argument in section 7.1). Consequently, I aim to support the view of congruence without comprehensiveness, to elaborate on how persons are motivated towards the maintenance and endorsement of the well-ordered society.[59]

In this final chapter, I sketch out my positive thesis: a Sartrean-Rawlsian approach to the argument from congruence. I divide the congruence argument into two parts; (i), a reflective endorsement of the Rawlsian society correlative to a particular form of self-conception, which I call '*reflection without rationalism*', and (ii), the inclusion of justice within a (form of) fundamental project, as a way of integrating and structuring one's other projects in accordance with the demands of justice. In this chapter, my primary focus is on the first of these accounts, *reflection without rationalism*. There are two reasons for this. First, a focus on reflection fits with Rawlsian orthodoxy, at least as I have presented it in this thesis. Second, the analysis required to fully support (ii) is beyond the scope of this thesis, given that one's pre-reflective experience – as Zahavi (2014, p.12) suggests – is often difficult to fully articulate. However, it should be kept in mind that both (i) and (ii) are interconnected. As I have emphasised in chapters 6 and 7, reflection is an extension of one's pre-reflective commitments – and my approach to congruence is designed with this fact in mind.

This chapter proceeds as follows. In section 8.1, I argue that reflection is valuable independent of the Rawlsian account of deliberative rationality (or, for that matter, any procedural account of rationality). I demonstrate that reflection allows the agent to respond to disruptions within the world by reorganising their commitments. Building on chapter 6, I claim that when the agent reflects, they view themselves *from a perspective analogous to their view of others*. In this way, reflection responds to disruptions within the world by establishing how they (such disruptions) relate to a particular self-conception that the agent upholds. In section 8.2, I restrict reflection to the context of justice and goodness. Here, I draw on Sartre to argue that the function of reflection is to reorientate oneself in an ostensibly political world, thereby revealing that one's pre-reflective experience is inextricably moulded by the political.

In section 8.3, I move this discussion to the Rawlsian framework and argue that justice as fairness takes for granted the fact that all persons maintain a practical self-conception, which serves to substantiate the content of reflection within part (i) of the congruence argument. From this, in section 8.4, I then argue that justice as fairness appeals to a practical self-conception that all persons are able to share. I argue that the assurance of the congruence argument is that

[59] Here, the phrase 'congruence without comprehensiveness' is taken from Mendus's (1999, p.57) claim that Rawls's argument can be pursued in "a way which ensures congruence without implying comprehensiveness."

the conditions of justice will come to mirror this practical self-conception. Thus, the subject's understanding of what it means to formulate and pursue a conception of the good (i.e. their view of themselves as a practical agent) fits with – and is expressed by – the organisation of society within justice as fairness. By recognising this feature of the Sartrean-Rawlsian approach, persons are motivated towards the upkeep of justice as fairness.

In section 8.5, I point out an important nuance, that further develops the points that I have made in this introduction. According to the arguments I present in this chapter, and have presented earlier in this thesis, *proving* the harmony between justice and goodness is not the aim of the congruence argument (that is, outside of the reflective affirmation of one's sense of justice). Nevertheless, I briefly explain why pursuing the issue of congruence remains important – and I go so far as sketching out, in broadly Sartrean terms, a possible way of describing the congruence between justice and goodness from the viewpoint of the agent. I argue that justice and goodness are congruent when one's sense of justice can be affirmed *within one's choice of oneself within the world*. Here, I draw on Sartre's idea of a fundamental project. Specifically, I argue that both justice and the good are projects regulative of other projects. To maintain a conception of the good is to be committed to a project that integrates and structures one's experience – the same, one should note, can be said for justice. Ultimately, my reconstruction of the congruence argument has two parts. First, reflecting on justice and goodness reveals that the conditions of the well-ordered society match up to one's practical self-conception. Second, having established this society, one's sense of justice will then constitute a way that one exists through the world.

8.1. Reflection, Reorientation, and Self-Awareness

In chapter 6, I argued that a person's life plan cannot be definitive of their good, because it is an extension of what they already take to be good and thus transforms it. As I emphasised there, this does not mean that reflection is not beneficial, or that reflecting on oneself and one's circumstances cannot be both normatively and epistemically useful. In light of this, I call part (i) of the Sartrean-Rawlsian congruence argument '*reflection without rationalism'*, since I will extol the benefits of reflecting on justice and goodness without guaranteeing that the outcome of this reflection will be a procedurally 'rational' one.[60] In fact, this marks the core distinction between mine and early Rawls's approach to this matter: whilst Rawls utilises rational procedures to *ensure* the agent's endorsement of justice as fairness, my concern is with the

[60] Throughout this chapter, I will refer to the entire argument in sections 8.1-8.4 as 'reflection without rationalism'.

benefit that reflection has *for the agent*. My approach maintains that reflecting on the relationship between justice and goodness within the well-ordered society will lead the agent to affirm the principles of justice in their own lives; crucially, it is in that *process* that reflection contributes to the argument from congruence, and not, as Rawls (1971, p.417) puts it, in "the outcome of… reflection".

In our day-to-day lives, we generally feel *prompted* to reflect. We reflect in order to 'step back' and reassess our commitments from a different perspective; and we do so because we have encountered some form of disruption, a problem to be solved.[61] Such disruptions occur when the patterns and expectations set by our unreflective commitments are called into question, meaning we must reflect in order to discover some 'new' or 'different' way to proceed. It is here that I locate the benefit of reflection. My claim is that reflection *allows the agent to reorientate themselves* so that their projects and commitments can be brought back into relative harmony. To convey the broad strokes of my position on reflection, consider my example below:

> Having lived with her for many years, Solomon has always considered Georgia to be a good person; however, seeing Georgia mistreat an animal that she was meant to look after, he comes to doubt this belief. The patterns of behaviour, including the sense of trust and affection for Georgia, break down. As a result, Solomon naturally wonders: is Georgia really a good person? In attempting to answer this question, Solomon deliberates over the recent situation and how this connects to what he already believed about his housemate. More likely than not, he will note that Georgia has not exhibited this kind of behaviour before, that she has generally been quite sweet to animals, and that her behaviour towards humans (for example, that time she looked after Natalie when she was sick) is often exemplary. At the same time, Solomon cannot doubt this new evidence: Georgia mistreated an innocent animal, showing a level of cruelty completely at odds with what a good person ought to do.

Whilst I will continue this example momentarily, there are already several relevant themes for describing the structure and the benefit of reflection. First, there are Solomon's pre-reflective commitments, which – as I have argued in the previous chapters – the act of reflection presupposes. Having lived with her for so long and given their closeness, it is plausible to claim

[61] In this section, I take inspiration from Larmore's (2010) description of reflection and its benefits in *The Practices of the Self*.

that Solomon's sense of a practically organised world, and the structures of possibilities that such a world offers up, depended on his relationship with Georgia.[62] Second, there is a disruption to Solomon's pre-reflective commitments that prompts an act of reflection, e.g. Georgia's treatment of animals. Solomon's relationship with Georgia – and the interpersonal, affective, and behavioural patterns associated with it – is brought into question. He is now unsure as to whether Georgia is a good person, whether he can trust her, and whether his previous understanding of her character is warranted. Third, reflection is Solomon's way of resolving this disruption. Descriptively, he reflects in order to come to grips with the newfound state of the world, in a way intended to resolve the uncomfortable realisation that Georgia may not be trusted; that she may not be a good person.

In this example so far, Solomon has yet to solve his problem. Instead, *he has made it explicit to himself*: he previously took for granted his relationship with Georgia, but now he has reason to doubt this commitment owing to her disreputable actions. Crucially, there is a fundamental tension that reflection seeks to resolve, a tension between the agent's (e.g. Solomon's) prior commitments and an actual or expected change in the *status quo* of the world (e.g. Georgia is now not to be trusted). To see that reflection involves responding to actual or expected disruptions within the world relative to our projects, consider a broad description of planning. When we plan, *we attempt to negotiate the world in advance*, and more often than not, this is because we have lost touch with the path that we were originally following. Reflection serves to rediscover this path or, if necessary, present alternatives to one's explicit awareness. Even if we have not directly encountered a disruption, we still anticipate certain problems arising within our plans and attempt to resolve any conflict that they might have with a given project. *Reflection thus aims at 'retrieving' our projects so as to make this tension explicit*, enabling the agent to deliberate over the appropriate way to proceed.

There remains a final question: how does reflection serve to resolve the tension between one's pre-reflective commitments and a disruption (or expected problem) within the world? This seems especially important given, as I mentioned in section 6.3, that Sartre (2018, p.75; 592) views reflective deliberation as 'rigged in advance'. As in that section, I part ways with Larmore's (2010, p.84; 91) interpretation of Sartre; namely, that reflection is necessarily deceptive or even inefficacious (serving only to 'distract' the agent from realising their ends). In contrast, even though I agree that reflection is rigged, I hold that *the act of reflecting also*

[62] For an account of pre-reflective agency that directly explores its interpersonal dimensions; see Ratcliffe's (2024) article 'We are our Possibilities: From Sartre to Beauvoir to Løgstrup'.

involves a form of reorientation. A person can pursue a multitude of diverse – even partially conflicting – projects at once; by reflecting, we seek to reorganise these projects (and the ends that they aim at), which can involve the affirmation of one project at the cost of another. As a result, reflection impacts the agent's actions by enabling them to realise a way to achieve, reaffirm, or relegate already established commitments. With this in mind, I will continue my example of Solomon and Georgia, in order to show *how* reflection contributes to the restructuring of the agent's projects:

> Having made explicit to himself the problem – that his previous commitments (e.g. his trust in Georgia) are in tension with the newfound state of the world (e.g. Georgia's actions) – Solomon reflects in order to reorganise his ends. Subsequently, he asks himself: 'should I continue being roommates with Georgia? Should I report her to the authorities? Do I want to be friends with someone who mistreats animals?' Although Solomon's reflection is aimed *at resolving a tension within the world* – broadly, the fact that he used to trust Georgia, who he discovered is cruel to animals – his way of dealing with this tension is to determine whether his commitments fit with *a particular view of himself*. The solution then comes to the fore: Solomon is a vegetarian, he is kind to animals and believes that they should be treated well, so he cannot continue living with Georgia.

To resolve this tension, then, Solomon reflects on whether the state of the world is compatible with his particular self-conception. In terms of a structural description, it is through the act of suspending and retrieving his projects *via* reflective awareness that Solomon establishes and commits to a form of self-identification (I will say more on the form of this 'self' in a moment). As a result, this process then sets real limits on what the agent (e.g. Solomon) can or cannot do. As Sartre (2018, pp.242-243) puts it, one's reflective self-conception "may be the basis on which the for-itself determines itself to be what it has to be…this phantom world exists as the for-itself's *real situation*." Hence, by figuring out how the problem relates to a particular view of himself, Solomon reorganises his projects and sets limitations on his actions, thus restoring his sense of worldly engagement; "in so far as he decides and acts, the for-itself who exists in the voluntary mode *wills to retrieve himself*…as a spontaneous project towards this end or that." (Sartre, 2018, p.592, emphasis added)

In regard to the self-conception of reflective consciousness, as I argued in chapter 6, the self of reflection is non-identical to the subject of experience. Importantly, there is a worry here that by taking a perspective on oneself, the self that is reflected-on is both *exteriorised* and *objectified*. Indeed, as I mentioned in that critical discussion, we often attribute object-like qualities to ourselves *via* reflection. Subsequently, this thesis of non-identity between the subject of experience and the self within reflection seems to bolster Larmore's (2010, 84; 91) interpretation of Sartre: that reflection is necessarily deceptive. There is, however, a more charitable and nuanced interpretation of Sartre available, one which brings the benefits of reflection to the fore. Although fully developing this account is beyond the scope of this thesis, I maintain that one still can reflect on oneself *as a subject* – rather than "*merely* an object" – of experience (Thomas, 2010, p.167, original emphasis). To see this, consider Sartre's relatively consistent use of visual metaphor, which he uses to clarify the function of reflection:

> …That what is reflected on is profoundly altered by reflection… It knows itself to be looked at… we can find no better comparison for it than a man bent over a table, who is writing and who knows, even while he is writing, that he is being observed. What is reflected on is therefore already, in a way, conscious (of) itself as having an outside…[it] exists over there, at a distance from itself, in the consciousness reflecting on it (Sartre, 2018, p.220).

In this instance, the fact that Sartre provides this description without introducing another person within his example conveys the idea that, when reflecting, *I am like the Other to myself*. Whilst we sometimes attribute object-like qualities to the self in reflection, this does not mean that the reflected-on self is always and only an object for consciousness. Instead, the best way to understand Sartre's description of reflection is that – through the act of formulating a self-conception – *we view ourselves in a way analogous to our perspective on other people* (Thomas, 2010, pp.161-180). Hence, reflective awareness becomes an almost prototype version of our being-for-others, seen here in Sartre's distinction between a 'virtual' and 'actualised' exteriority:

> As soon as one takes up the standpoint of… reflection — the kind of reflection that seeks to determine the being that I am – an entire world appears… it is 'my shadow', what is disclosed to me when I want *to see myself…* Here we find the first draft of an 'outside'; the for-itself sees itself almost conferring an 'outside' on itself, in its own eyes: but this 'outside' is purely virtual. Later on we will see how being-for-the-Other *actualizes* the first draft of this 'outside' (Sartre, 2018, p.243, original emphasis).

When we reflect, we respond to salient features of the world by reorientating ourselves on the basis of a particular self-conception. Whilst this self-understanding is not the subject of experience, it is nevertheless directed towards oneself as *a subject*. Here, the depersonalising use of 'a' conveys the fact that the reflected-on self is analogous to the agent's relation to the

Other, and not directly identical with the subject of experience. Therefore, the agent is able to attribute properties and commitments to themselves in reflection *as if* they were viewing a subject from an exteriorised perspective. In turn, the agent can then affirm the projects that cohere with his self-conception and relegate those that do not.

In sum, Solomon's reflective view of himself is a mechanism by which he can resolve the tensions that arise in his life. After all, his day-to-day existence is marked by a multitude of commitments, possibilities, requirements, etc. Reflection then serves to resolve these difficulties, by allowing Solomon to *reorient himself within the world by reflecting on his self-conception as it would be for the Other*. As a caveat, this reflective act may not always be successful, and it ought to be understood as a process rather than a spontaneous achievement. However, it is in this tendency towards reorientation that reflection finds its benefit. Though the path may be more or less set in advance, reflection is a way of regaining one's footing. In the next section, I will deepen this understanding and argue that reflecting on the connection between justice and goodness enables the agent to make explicit the impact that justice has on their life. From a broadly Rawlsian perspective, I will then argue in sections 8.3 and 8.4 that justice as fairness is sensitive to these pre-reflective structures of meaning, entailing that the well-ordered society is uniquely situated to respect persons' claims to self-authentication.

8.2. Exploring Our Experience of the Political

In drawing out the political implications of *reflection without rationalism*, I will begin with the basic phenomenological tenet that "the world in its broadest sense is both constructed and given" (Jung, 1982, p.166). One of the main upshots of this view is that human beings find themselves 'already' – that is, unproblematically and unreflectively – situated in a complex world of meanings and values; "I, through whom meanings arrive in things, find myself committed within a world that is already meaningful and which reflects back to me meanings that I did not put there" (Sartre, 2018, p.664). As such, it is incorrect to presuppose that the 'limits' to one's sense of agency are purely material, e.g. that they consist of physical or biological constraints. Instead, my situation is also interpersonal, meaning that it involves pre-established normative structures that, as Sartre (2018, p.665) puts it, "resist and remain independent of me." For my current purposes, this 'independence' highlights the fact that one experiences the world as immediately and pre-thematically exercising certain normative demands on oneself (recall section 7.3). Rather than initially appearing as features of a reflective consciousness, then, such significations persist as 'objective' features of the world, which direct and mediate one's actions and relations. As Sartre writes:

> …directions are for the most part imperatives: 'Enter that way', 'Leave that way'; that is what the words 'Entrance' and 'Exit', painted above the doors, signify. I submit to these directions: to the coefficient of adversity in things that is engendered by me, they add a coefficient of adversity that is strictly human (Sartre, 2018, p.665).

In this quote, Sartre utilises the notion of 'coefficient of adversity' – that is, very broadly, the resistance that facticity offers against freedom – to highlight how 'human-made' structures can set real limitations on one's sense of agency. As I have suggested, and as this passage serves to exemplify, these structures appear fundamentally as part of the world – descriptively, the imperative to 'exit this way' is not only an instruction that mediates my action, but a neon sign bolted up above the door. As such, my pre-reflective sense of agency takes for granted a complex nexus of world-embedded demands and norms; "we project ourselves from the outset towards possibles, and our understanding has to be *on the basis of the world*" (Sartre, 2018, p.670, original emphasis).

These observations ought to be fairly familiar from my arguments in chapters 6 and 7. What is of greater interest here is that Sartre (2018, p.668, emphasis added) generalises this phenomenon to include the political and cultural: "I am not only thrown in the face of the brute existent; I am thrown into a world that is working-class, French, with the character of Lorraine or of the South, and which offers me its meanings *when I have done nothing to reveal them*." Just as one's immediate experience relates to objects of significance in one's environment, socially-mediated concepts and categories (say, class or nationality) are also part of the make-up of the world; they can resist or expand on one's sense of one's possibilities. Our day-to-day experience is shaped by socially mediated frameworks of meaning, along with the practices, techniques, and material structures that accompany them:

> …if I submit to this structure, I depend on it: the benefits it provides me with may dry up; it only takes some civil unrest, or a war, and suddenly the most essential products become scarce, whether I like it or not. I am dispossessed, halted in my plans, deprived of what is necessary to achieve my ends… (Sartre, 2018, pp.665-666, original emphasis).

I will now narrow down my analysis. It is clear that disruptions to political and social structures are also disruptive to the agent's mundane experience, e.g. in times of 'civil unrest' one is 'dispossessed', one's plans are 'halted', and so on. Thus, reflection is dependent on a pre-reflective sense of agency that is ultimately moulded by a political and social 'lifeworld' (to borrow Husserlian terminology). The question is this*: in the context of the congruence argument, what can reflection reveal about such a world*?

My claim is that, by reflecting on the relationship between justice and goodness, one attempts to make the ubiquity of 'the political' in one's day-to-day life explicit. [63] Importantly, this comes out in Sartre's (2018, p.134, original emphasis) brief description of class: it is only "by taking it [being working class] up within the infrastructure of my *prereflective cogito* that I endow it with its meaning and its resistance." To unpack this point, note that whenever I encounter a disruption in my life owing to my class, it is not strictly speaking possible for me to isolate this fact in reflection. Instead, I am deliberating over what Sartre might call a *totality*; I have to *be* this class, and so I will be deferred to my childhood, my education, my clothes, my use of language, and so on. Correspondingly, Sartre describes oppressed workers as unable to enact a change to their circumstances precisely because this form of reflection is rendered unavailable to them:

> The worker… does not represent his sufferings as intolerable; he puts up with them, not through resignation but because he lacks the…reflection necessary for him to conceive of a social condition in which these sufferings would not exist (Sartre, 2018, p.571)

In light of this, reflection is structured by the limits and expectations sustained by communal norms, i.e. that one's life is lived within, and in part directed by, complex social frameworks. Yet, in this same sense, reflecting on the conditions of justice is beneficial for the agent since it allows them to *reorient themselves within an ostensibly political world*. One's pre-reflective experience is saturated by the political. Insofar as reflection reorganises the world in advance, then doing so within the context of justice and goodness reveals to the agent the relative scope of the former in relation to the latter. As Rawls (1971, pp.563-564) writes, "everyone's conception of the good… [is part] of the larger comprehensive plan that regulates the community … we do not start *de novo*.". In fact, this act of reorientation seems to be directly implicated in Sartre's description here, wherein he explores the possibility of political revolution:

> It is not the harshness of a situation…which provide the grounds for conceiving of another state of things in which everyone would do better; on the contrary, it is on the day *when we become able to conceive of another state of things* that a new light is thrown on our hardships and our sufferings, and we decide that they are intolerable…(Sartre, 2018, p.571).

In a similar vein, Rawls (1971, p.7) selects "the basic structure of society" as the subject of justice "because its effects are so profound and present from the start.". Importantly, there is a

[63] I shall follow Gurley (2016, p.234) and Jean-Luc Nancy (1991, p.40) in taking the 'the political' to denote "the disposition of the community as such" or "the organizational structuring of a community as such."

deep acceptance in *Theory* that the organisation of the main political and social structures of society will reverberate throughout one's day-to-day existence. Essentially, this fact serves to implicitly motivate the main question of the congruence argument: the sense of justice is a desire that is supremely regulative, and this means that persons will have to affirm the principles throughout their lives, *how do we know that they are capable of achieving this*? In my account, reflecting on the conditions of justice will enable the agent to consider the relationship between their projects and a politically structured lifeworld. Subsequently, reflection will prove crucial insofar as one's life is *not* the subject of an objectively rational life plan and insofar as it *is* susceptible to the dynamics of change and uncertainty – for one will need to reflect in order to restore, and make sense of, one's place within the well-ordered society.

There is more that needs to be said here. All I have demonstrated so far is that, (a), acts of reflection can be beneficial for the agent in terms of bringing together their normative commitments, (b) reflection is a political act in that it tells us something about the structure of social frameworks and *vice versa*, and (c), reflecting on goodness and justice can reveal to the agent the ubiquity of 'the political' in their day-to-day, unreflective, experience. Towards the end of this discussion, I have started to suggest that justice as fairness anticipates some of these core features (in particular, the ubiquity of justice). In the next two sections, I shall substantiate the Rawlsian response to these observations. I will argue that Rawls presupposes that persons maintain a practical form of self-conception that can be articulated through a broadly Sartrean lens. The significance of this standpoint, moreover, is crucial to my reconstruction of the congruence argument: by affirming the principles of justice within one's conception of the good, *one affirms a self-conception that all persons can share in*.

8.3. The Affirmation of a Practical Self-Conception

At this stage of discussion, recall some of the main issues that Rawls's argument from congruence addresses. As I clarified in chapter 2, and as Freeman (2003, pp.286-287) points out, there are five such issues; (1), the concern that one's sense of justice is either arbitrary or the result of inoculative forces; (2), the worry that one's sense of justice involves the renunciation of personal responsibility or deference to some authoritative power; (3), the idea that one's sense of justice may be grounded in undesirable qualities, such as envy; (4), the Nietzschean argument that pursuing justice requires disavowing the higher powers and capacities of the self; and (5), the worry that one's sense of justice is a purely private disposition, which fails to account for the notions of sociality and community. Note also that these concerns are couched within the broader worry that justice is "too demanding for most people given certain tendencies of human nature" (Freeman, 2003, p.283). In this section, I will clarify how the Sartrean-Rawlsian approach to congruence answers these concerns taken together. To begin with, consider the following example from Rawls:

> …We sometimes doubt the soundness of our moral attitudes when we reflect on their psychological origins. Thinking that these sentiments have arisen in situations marked say by submission to authority, we may wonder whether they should not be rejected altogether. Since the argument for the good of justice depends upon the members of a well-ordered society having an effective desire to act justly, we must allay these uncertainties. Imagine then that someone experiences the promptings of his moral sense as inexplicable inhibitions which for the moment he is unable to justify. Why should he not regard them as simply neurotic compulsions? If it should turn out that these scruples are indeed largely shaped and accounted for by the contingencies of early childhood, perhaps by the course of our family history and class situation, and that there is nothing to add on their behalf, then there is surely no reason why they should govern our lives. But of course to someone in a well-ordered society there are many things to say. (Rawls, 1971, p.514).

Importantly, this description accords with my analysis so far. Rawls anticipates that citizens within society are liable to experience disruptions to their projects, causing them to reconsider the value and the origin of their sense of justice. As I indicated towards the end of the previous section, Rawls (1971, p.573) accepts that committing to justice is an often very demanding task; "it is true enough that for the sake of justice a man may lose his life where another would live to a later day." Yet, the description in the indented quote also agrees with my more substantive point so far: what comes to focus when the agent reflects on the conditions of justice are the manifold social structures that serve to carve out their place within the world; 'class', 'the contingencies of childhood', 'family history' and so on. The question then arises as to whether the agent can affirm the role of these structures in their own lives, or whether they constitute structures that are, in a broad sense, both alien and alienating.

I think there is something deeply Sartrean about Rawls's main sentiment in the indented quote. At first glance, one would be forgiven for thinking that – as Rawls's hypothetical individual laments – there is no reason why the organisation of social structures governs so much of our lives*, they just do*; they are the same structures that we take for granted in our daily experience. However, Rawls's argument is that we can come to affirm these structures – and the principles that organise them – *as part of our own projects* (for more on this, see section 8.5). Accordingly, the agent is not a passive recipient of an already politically-structured lifeworld, but instead has *to exist through this world* and so in an important sense, *sustains it.*

On the Sartrean-Rawlsian view, one's projects are not passively formed within an assumed situation, but instead actively give that situation its meaning and structure.[64] As Sartre writes:

> …The for-itself cannot be a person – which is to say to choose the ends which it is – without being a man, the member of a national community, a class, a family, etc. But these are abstract structures that are maintained and surpassed through its project (Sartre 2018, p.680).

Whilst I will return to this point in section 8.4, I contend that the Sartrean underpinnings to *Theory* arise because of Rawls's attempt to appeal to *the foundation* of one's contingent preferences and inclinations. As Freeman (2003, p.294, emphasis added) puts it, "if the sense of justice can be shown to belong somehow to our nature, then Rawls can contend that, by affirming it, *we exercise a capacity that is fundamental to our being*." Indeed, this explains and adds substance to Rawls's (1971, p.574) overarching motivation for acting in accordance with the principles of justice; "in order to realize our nature we have no alternative but to… preserve our sense of justice as governing our other aims." Hence, *Theory*'s approach to the problem of congruence is, in a sense, to directly call on the powers of the self, i.e. "our nature as a free and equal rational being[s]" (Rawls, 1971, p.256; 519; 572; 574). Consequently, the worry that one's sense of justice is arbitrary, that it is dependent on an authoritative power, or that it leads to the diminishment of the self, is bypassed. Instead, Rawls (1971, p.515) appeals to a capacity for agency, which he takes to maintain a certain priority over the "natural contingencies and accidental social circumstances" of life.[65]

Yet these claims must be contextualised within the current discussion on reflection, given my argument that talk about 'true nature', 'noumenal selves', or similar Kantian-inspired claims, are unnecessary to the congruence argument overall. Crucially, such talk invokes 'comprehensive doctrines' in the guise of claims about persons and is excluded in Rawls's move to *Political Liberalism*. In juxtaposition to these doctrines, *reflection without rationalism* concerns the agent's practical response to the problems concerning their sense of justice within the well-ordered society (for example, Rawls's worry that one's sense of justice will be seen as a 'neurotic compulsion'). In such instances, the agent will be reflecting on a relevant self-conception to examine whether justice coheres with their normative commitments. My claim is that the agent reflecting on the conditions of justice within the well-ordered society will see that it matches up with their practical self-conception and so will want to affirm their sense of

[64] The discussions in sections 7.3 and 7.4 of the previous chapter serve to bolster this claim.

[65] Here, keep in mind my broad account of stability in chapter 1. Much of this thesis has served to exemplify the importance of agency to the Rawlsian picture.

justice as part of their good. Crucially, from the Sartrean-Rawlsian perspective, there is no 'true self' that leads the agent to affirm their sense of justice, no metaphysical background upon which congruence is established. Rather, there is only a self-conception that the process of reflection takes for granted. More substantially, *reflection without rationalism* appeals to the agent's *practical understanding of themselves* in order to demonstrate that justice as fairness suits that particular form of self-conception – without the need to attribute to persons a 'noumenal self' or an 'objective' version of their good. Here, Freeman's discussion helps to substantiate this claim from a Rawlsian standpoint:

> Rawls's idea is that, from a practical point of view… we normally see ourselves and each other not just in terms of our particular identities, ends, and commitments; more fundamentally we conceive of ourselves and others as free moral agents capable of determining our actions, adjusting our wants, and shaping our ends – all according to the requirements of practical principles… (Freeman, 2003, p.295) [66]

With this in mind, the notion of a practical self-conception is best understood as *the view of oneself as being capable of formulating and authenticating one's idea of the good*. To see this, recognise that the worries that I have described above *via* Rawls and Freeman primarily concern whether one's *particular* identity, or one's current commitment to justice, is the result of indoctrination/other means of self-abnegation. Put in Rawlsian terms, the worry is that the agent is unable to self-authenticate their sense of justice, which – until now – has had a major role in shaping their projects and commitments. On my account, the agent responds to these worries by reflecting on whether the prevailing social institutions support their capacity to fashion their own ends, in order to then discern how such institutions relate to their projects. Should the agent discover that the conditions of justice fit with their practical self-conception, then they have reason to incorporate the Rawlsian sense of justice within their actual good, as defined by their contingent desires and interests.

Moreover, in undertaking this process of reflection, persons within the well-ordered society will be able to affirm justice when viewing themselves as *active participants* in the political schema. At least in broad terms, the problem that motivates the congruence argument

[66] It is worth reemphasising Rawls's mistake here, in line with my analysis in this section and in the introduction to this chapter. In particular, if Freeman's description is right, then Rawls moves from the idea that we all have a *view of ourselves* as 'free moral agents' towards the claim that "persons are, *by their nature*, free, equal and rational" (Freeman, 2003, p.295, emphasis added). Yet these two claims are not equivalent. In light of this, my Sartrean-Rawlsian approach is not making a claim about the nature of persons *per se*; to borrow Sartre's (2018, p.683) phrase, "this does not belong to freedom's nature, because here there is no nature." Instead, *reflection without rationalism* upholds the more parsimonious view that persons maintain a conception of themselves as practical agents that we can explore from a phenomenological standpoint.

is the comprehensive impact of justice. As I have described it, this includes the influence that the political has on one's pre-reflective experience. I have also maintained throughout this thesis that reflection is not a neutral form of representation; rather, it is itself a form of commitment, a process that makes one's projects explicit in order to determine the appropriate way to proceed. Understood in this context, then, to reflectively affirm the conditions of justice *is a way of committing to justice within one's own life*. On *reflection without rationalism*, to say that goodness and justice are congruent is to say that the conditions of society can be affirmed by the agent when they consider how it relates to their practical self-conception. The agent then resolves any expected conflict that justice has with their projects and is thus reorientated within a world that they recognise is saturated by the political, but that nevertheless appeals to their capacity for practical agency. To summarise my argument:

1. Reflection is beneficial for the agent since it allows them to respond to problems within the world and reorganise their projects *via* a relevant self-conception.
2. Reflection is inextricably connected to the political, both in the sense that one reflects *on* a political lifeworld and reflects *within* such a world.
3. Given the scope of justice, persons are liable to consider whether their particular identity is the result of indoctrination or other means of self-abnegation.
4. Reflection can reorientate the agent by establishing whether the organisation of society is congruent with their practical self-conception.
5. Within justice as fairness, the conditions of justice are congruent with the agent's practical self-conception, meaning that the agent will reaffirm their sense of justice within the well-ordered society.

There are two main tasks that remain. The first is to build on this discussion here and apply this account of *reflection without rationalism* directly to justice as fairness, in order to demonstrate *why* persons should affirm their sense of justice as part of their conception of the good (or rather, why congruence between the conditions of justice and one's practical self-conception is appropriate motivation to affirm justice within one's life). [67] The second is to rethink the congruence argument by exploring how justice and goodness can be congruent *in concrete terms*, rather than in the agent's self-conception. In taking these steps, I will then argue that whilst my account cannot discursively 'prove' the congruence between justice and

[67] Here, one should keep in mind that – *as per* chapter 2 – the justificatory stage of justice is settled prior to this discussion (which would, to note, rule out unjust conceptions from appealing to my approach). The issue here is whether the conception of justice is stable, and specifically, whether Rawls's conception can promote the two generative properties of justice as fairness.

goodness – which is distinct from Rawls's original argument, given that he imposes an account of procedural rationality to achieve his aim – it nevertheless retains some major benefits. In particular, *reflection without rationalism* means attending directly to the perspective of the agent and recognising that justice, at the very bottom, is a way that we collectively make sense of the world; a way that we give meaning to our own lives.

8.4. Practical Self-Conceptions and the Assurance of Congruence

In my reconstruction of the congruence argument, there are two interconnected claims that motivate persons to incorporate the Rawlsian sense of justice within their conception of the good. Namely, by focusing on a practical self-conception, justice as fairness (i) accounts for the pre-reflective structures of meaning that shape one's experience of the political and (ii) maintains a view of the self that, in principle, all persons can share in. The upshot of these points is that justice as fairness is uniquely situated to recognise persons' claims to self-authentication. Importantly, this is because the Sartrean-Rawlsian approach does not only formally recognise one's practical self-conception, but accounts for it within the very fabric of society. By affirming the principles of justice, the agent is assured that their understanding of themselves as a practical agent *is confirmed by the organisation of the social world*. Although one should keep in mind the adaptions that I make to his view, Rawls puts it this way:

> …We can say that by acting from these principles persons are acting autonomously… under conditions that best express their nature as free and equal rational beings. To be sure, *these conditions also reflect the situation of individuals in the world and their being subject to the circumstances of justice* (Rawls, 1971, p.515, emphasis added).

Here, I shall start with point (ii), as it naturally leads to both (i) and to my broader argument. Importantly, the fact that justice as fairness maintains a view of the self that all persons can share comes out through Rawls's (1971, p.528) treatment of what he calls "dominant-end" views of justice. As Rawls (1971, p.561) writes, "the distinctive feature of a dominant-end conception is how it supposes the self's unity is achieved." Although Rawls (1971, p.561) does not offer a general rule of thumb for this achievement, he does consider a form of hedonic utilitarianism, wherein "the self becomes one by trying to maximize the sum of pleasurable experiences within its psychic boundaries." Rawls then criticises this account because of its failure to appreciate the separateness of persons, because of its attempt to subordinate rather than regulate persons' aims, and because of teleology's tendency to bring disunity to the self (similarly to the tensions highlighted in sections 4.3 and 4.4).

That being said, the interesting point here is not Rawls's (1971, p.558) general critique of utilitarianism – for, it might be noted, he takes its drawbacks to be fairly apparent – but his description of why dominant-end approaches to justice are appealing, why such approaches fail to realise this appeal, and how justice as fairness is distinct from such approaches. In regard to these points, Rawls writes:

> It would appear… that the turn inwards to the standard of agreeable feeling is an attempt to find a common denominator among the plurality of persons, an interpersonal currency as it were, by means of which the social ordering can be specified. And this suggestion is all the more compelling if it is already maintained that this standard is the aim of each person to the extent that he is rational (Rawls, 1971, pp.558-559).

Here, Rawls recognises a plurality of persons that naturally involves a plurality of aims; "human good is heterogeneous because the aims of the self are heterogeneous" (Rawls, 1971, p. 554). The appeal of hedonic utilitarianism is that it stands as a simplifying procedure by which anyone can order their aims and interests (and ultimately, competing aims/interests across society). Subsequently, what was originally a complex multitude of purposes can be brought down to one organisational principle: promote social arrangements that maximise pleasure. Hence, it is the primary aim of dominant-end theories to establish a currency *whereby all persons can come to understand themselves and the world on the same terms*.

The shortcoming of this approach, however, is that persons are expected to join together over what Rawls (1971, p.566) calls "a more or less homogenous quality or attribute of experience", e.g. pleasure for hedonic utilitarianism. Despite this appealing aim, Rawls's (1971, p.557) argument is that the organisational principle of maximising pleasure *is not generalisable in this way* and, therefore, cannot be used to simplify social arrangements. Persons will still have to *decide* which course of action maximises pleasure, "taking into account the full range of…[their] inclinations and desires", even though pleasure is meant to be a dominant-end that is intelligible from any perspective. As Rawls writes:

> Although to subordinate all our aims to one end does not strictly speaking violate the principles of rational choice (not the counting principles anyway), it still strikes us as irrational, or more likely as mad. The self is disfigured and put in the service of one of its ends for the sake of system (Rawls, 1971, p.566).

The Rawlsian worry regarding dominant-end conceptions of justice is that they *posit an account of the self that is not appliable, or even informative, to all persons*. At worst, this self-conception is 'disfiguring', in that it demands subservience to the collective maximisation of ends; at best, it is insufficient, as one must still organise one's own ends through a process that is not exhausted by the principle of maximising pleasure. In either case, dominant-end approaches fail to specify a notion of the self that all persons can share, by first defining 'what is good' irrespective of 'what is right' (Rawls, 1971, p.560). However, when it comes to the actual organisation of the social life, both of these notions prove to be intwined; *our interests and ends are both formed within and conditioned by society*, and so restructuring its main institutions means appealing to *the capacity* to formulate such ends:

> It is not our aims that primarily reveal our nature but rather the principles that we would acknowledge to govern the background conditions under which these aims are to be formed and the manner in which they are to be pursued (Rawls, 1971, p.560).

Herein lies my point. From the Sartrean-Rawlsian perspective, *our sense of ourselves as practical agents cannot be separated from our sense of the world more broadly*. Indeed, this substantiates the distinction between justice as fairness and hedonic utilitarianism; the latter isolates part of the good and then organises social life accordingly, the former recognises that the pursuit of one's good and one's experience of a shared world are simultaneous and inseparable. Hence, the condition for formulating a conception of good – i.e. purposive agency – must be accounted for within the main structures of society, and *vice versa*. In my reconstruction, Rawls (1971, p.499) affirms this idea by appealing to persons' view of themselves as practical agents, which serves as assurance that the organisation of society "will be rooted not in abnegation but in affirmation of the self." This brings me to point, (i): by serving as an assurance in relation to one's practical self-conception and the conditions of justice, this reflective act serves to anticipate one's pre-reflective experience within the well-ordered society. In the Rawlsian terms, this can be put as follows:

> In due course everyone will know why he would adopt the principles of justice… eventually …we do not look at the social order from our situation but take up a point of view that everyone can adopt on an equal footing (Rawls, 1971, p.516).

On my account, the endorsement of a shared self-conception reflects the assumption that one will experience 'the political' as objective structures within the world, as well as the guarantee that such structures will not be arbitrarily discriminatory. To demonstrate this in more detail, I will draw on Lindahl's (2020, p.256) "phenomenological reading of legal ordering" to spell out how a shared self-conception can serve to anticipate the experience of social norms and structures within the well-ordered society.

As Lindahl (2020, p.257) puts it, the ordering of social and legal norms is a 'second-order' phenomenon because it involves a form of "practical engagement" that *enables* other "social forms of engagement with the world." Here, one can imagine a parliamentary edict that prohibits, say, walking the street with a blood alcohol concentration of .08%. In this instance, not only is a legal norm established, but also a "*pragmatic order*…which configure[s] certain kinds of relations between different places, subjects, times, and act-contents" (Lindahl, 2020, p.258, emphasis added). By combining these relations, the parliamentary edict then turns into a genuine restriction on possible courses of action that I can take (i.e. not drinking five pints and then going for a jog outside). That is, part of the function of legal ordering as a second-order phenomenon is that it encompasses these relations "as the dimensions of a *single* order of behaviour…by pointing beyond themselves to a pragmatic order as a meaningful nexus of legal relations" (Lindahl, 2020, p.258, original emphasis).

To expand on this discussion, recall my claim from section 8.2 that our experience of 'the political' was often unthematized and immediate, and that our awareness of social orders was typically characterised by our unreflective use of techniques, behaviours, etc. Below, Lindahl's analysis of legal ordering supports my observations:

> Under normal conditions, legal norms and the corresponding pragmatic order remain more or less unthematized and unobtrusive at the background of each event in which something appears as something to someone (Lindahl, 2020, p.258).

Importantly, Lindahl's (2020, p.258) analysis concurs with the view that one's experience of the world typically involves a background acceptance of social norms, or perhaps more precisely, "spatial, temporal, subjective, and material boundaries" that correspond to a pragmatic order. What interests me now is what happens when these boundaries and the accompanying order *are* thematised; when one's experience of the social order prompts an act of reflection. Here, my contention is that a shared self-conception within the well-ordered society lends itself to making our reaction to the thematization of social order *intelligible*. Moreover, a condition of this intelligibility is that the accepted social order is not based on arbitrary distinctions (i.e. distinctions that single-out members of communities because of their particular identity).

To appreciate this point, note that *since all are included within this account of practical identity* (i.e. shared self-conception), and since the conditions of justice match this self-conception, then the political structure of the world (in the well-ordered society) addresses in the first instance that self-conception. Another way to put it is that the ‘subject’ of the pragmatic order, to borrow from Lindahl, is in its essence a generalised self; that “instructions, labels, orders, prohibitions and name-plates are addressed to me in so far as I am just anybody” (Sartre, 2018, p.665). To see this, recall that reflection enacts a change in the normative stance of the agent by allowing them to view themselves in a manner analogous to their view of others. The upshot of this, as Larmore (2010, p.93, original emphasis) puts it, is that “by adopting a universal point of view…we can discern the reasons that *one* would have had to behave as we did.” It is here that I locate the possibility for all persons to maintain a practical self-conception since, in reflection, one *is* able to render one’s commitments intelligible *to anybody*; I am not concerned with my particular identity as such, but *the conditions for someone similarly situated to fashion for themselves their own lives*. In turn, and as I argued in the previous section, agents are then encouraged to affirm the political organisation of the world through their own projects – and indeed have reason to, given its congruence with their practical self-conception.

With this in mind, contrast this Sartrean-Rawlsian account with cases of arbitrary discrimination, wherein this sense of ‘anybody’ is instead made relative to the agent’s particular identity prior to their sense of self-authorship. In such instances, communities (and members thereof) find themselves to be discriminated against on the basis of an identity that is “*given*”. Descriptively, they experience a way of “being” than “is *undergone*…without *being existed*” (Sartre, 2018, p.681, original emphasis). Importantly, the difference is that whilst the objective structures of the well-ordered society cohere with the agent’s practical self-conception, and so can be reflectively affirmed as regulative of one’s projects; in cases of arbitrary discrimination, these structures correlate directly with the particulars of a person’s identity, which are ‘given’ rather than affirmed, and which are experienced as incommensurable with one’s projects. As a result, the shared quality of social structures – in the case of arbitrary discrimination – then becomes alienating, resistive, and objectifying; “I encounter this being at the origin of a thousand prohibitions and a thousand resistances that I come up against at every moment... in certain societies, I will be deprived of certain possibilities, etc.” (Sartre, 2018, p.681).

Therefore, an important advantage of grounding the stability of justice as fairness within a shared, practical, self-conception, is that it emphasises the need for self-realisation, without discriminating against persons on the basis of their particular identity. Accordingly, my reconstruction of the congruence argument takes steps towards fulfilling Bedorf and Herrmann’s description of the role of social orders in *Political Phenomenology*:

> The political must be shaped in such a way that different perspectives can come together… the role of the political is not simply to ensure a smooth coexistence of individuals but rather to create those conditions within which individuals can realize themselves. This is only possible if individuals can experience themselves in political action as identical to as well as different from others. Thus, the experiences of identity and difference are conditions that the political must provide if it is to enable the self-realization of individual (Bedorf and Herrmann, 2020, p.5).

In this case, *reflection without rationalism* accounts for the shared identity between persons by appealing to the mutual endorsement of a practical self-conception, which represents each person's capacity to formulate and pursue a good life. Whilst conversely, the differences between persons are accounted for by the fact that each person's life will be their own, and that all are understood as active participants within the well-ordered society. Thus, persons will want to affirm the Rawlsian sense of justice as regulative of their aims since "this sentiment *reveals what the person is*, and to compromise it is not to achieve for the self free reign *but to give way to the contingencies and accidents of the world*" (Rawls, 1971, p.575, emphasis added). Affirming a shared self-conception through one's sense of justice upholds this self-conception ('reveals what the person is'), which is prior to one's contingencies, whilst *the same can be said about the conditions of justice*. To *give way* to the contingencies of the world is to maintain a social order that discriminates on arbitrary grounds, the Sartrean-Rawlsian commitment –which the account of *reflection without rationalism* demonstrates – is that this will not be the case.

8.5. Justice as a Fundamental Project

At this point in the discussion, I am able to fully summarise *reflection without rationalism* as follows:

> Reflecting on the connection between justice (the arrangement of society) and goodness (one's conception of good) is an act of reorientating oneself within an ostensibly political world *via* a particular view of oneself. In regard to the well-ordered society, this reflective act will reveal that the conditions of the political world match up to the conception of oneself as a practical agent. As a result, by upholding one's sense of justice, one endorses a self-conception that all persons can share in; since all persons are able to view themselves as practical agents. Moreover, since this practical self-conception represents one's ability to formulate, authenticate, and pursue one's idea of the good, then one can affirm one's sense of justice within one's life, without this leading to self-abnegation.

In putting forward this Sartrean-Rawlsian approach, I have provided an account of congruence that treats the act of reflection as beneficial *for the agent*. This is a unique advantage of my approach. Instead of claiming that persons have a true nature, or imposing an objectively rational procedure over their choices, I have accepted that agents will – even within the well-ordered society, where the account of justice is already established and justified – experience disruptions which cause them to question their prevailing social institutions. I have built my approach around the natural idea that, during periods of disruption and disharmony, agents will come to reflect on their circumstances and attempt to come to some self-understanding. Moreover, I have argued that the stability of the well-ordered society will be ensured because persons will be able to reflectively endorse their sense of justice, leading them to see their institutions as independently valuable and motivating them to do their part in maintaining the organisation of society.

Nevertheless, there is an important issue that still needs to be addressed. To see this, note that I have only demonstrated that justice and goodness are congruent for the agent on reflection. However, this form of reflective endorsement – as my analysis in chapter 6 demonstrates – is distinct from the incorporation of justice within one's life more broadly. How, then, do we know that justice and goodness are congruent? Or even more fundamentally, how do we define a person's conception of the good? In a sense, this is a Sisyphean task. A rational plan of life cannot resolve this issue; nor, for that matter, do I think that this calls for an empirical or metaphysical solution. For example, the assumption that 'the good' is a property of the world implies that our conceptions of justice should be structured according to that metaphysical feature (to this, recall section 7.1). Similarly, assuming that goodness is an empirical question would either leave us with little to say about its importance (e.g. think an almost Humean view, wherein congruence 'just happens to be the case') or will succumb to the same shortcomings as the life-planning approach (e.g. persons being unable to translate their experience of the good to empirical data). Ultimately, though I do not have space to dismiss all possible interpretations, my position is this: *whether something is a part of one's conception of the good is a phenomenological question*. Something is only ever good-for-someone, such that the good can only be understood within the complex dynamics of lived experience.

If one agrees with my position, then the original argument from congruence seems almost certainly doomed to failure. To claim that justice *is* congruent with goodness is to generalise a qualitative claim that is only intelligible from a first-person perspective across a plurality of viewpoints. Nevertheless, there is something to be said about the Rawlsian argument from congruence. Whilst one cannot *prove* that justice and goodness are congruent, that does not mean that one should not *aim* at congruence. As my account of *reflection without rationalism* demonstrates, the question of justice and goodness provides a meaningful context for exploring issues of political philosophy.[68] Understood in this light, to say that one can affirm the principles of justice within one's conception of the good *just is* to make a claim about one's experience of the political. As a result, there is immense benefit in attempting to describe the congruence between justice and goodness; for one is trying to bring to attention the way that one's life is shaped by the shared constraints of the community.

Here, I will make a tentative suggestion as to how we should understand the issue of congruence between justice and goodness in broadly phenomenological terms – although I believe it is the only one implied by my arguments so far. In doing so, I will draw on the Sartrean notion of a "fundamental project". My claim is this: *one's choice of the good ought to be considered as identical with one's choice of oneself within the world*. As a result, justice is congruent with goodness *once it is included as part of that choice*. To understand this, recall from chapters 5 and 7 that one's projects refer to other projects, which are themselves dependent on an ongoing pre-reflective hierarchy of values. For Sartre, and similarly for the view that I am putting forward here, the only way to explain the totality of these projects is by recognising that they are grounded in the fundamental project that the agent *has to be*:

> In this moment while I am writing; I am not the mere perceptual consciousness of my hand tracing signs on the paper, for I am far ahead of this hand, reaching right out to the book's completion and the meaning of this book – and of philosophical activity in general – in my life; and it is within the framework of this project, *which is to say the framework of what I am*, that specific projects towards more limited possibilities are inserted, such as expounding this idea in such and such a way… (Sartre, 2018, pp.605-606, emphasis added).

In this description, a complex hierarchy of projects and values is integrated into each of our lives as a way of making sense of the world. The culmination of this process is given the form of a fundamental project, whereby deference to another justificatory project no longer remains intelligible (Sartre, 2018, p.606; Webber, 2009, p.55). To provide a roughly analogous description, upon being wronged, one might ask the perpetrator: 'why did you do such and

[68] It is interesting to note how my approach relates to contemporary Rawlsian scholarship. Weithman (2024, p.14; 18), for example, seems to accept that there is no guarantee for congruence. However, he defends the aim of congruence because it reveals the value of justice. I believe my account accords well with that description.

such?' throughout progressively broader levels of explanation. Once they have gotten through the more immediate justifications 'I needed the money...I wasn't thinking straight…etc.', one will come to an explanation that captures these values and patterns of action in a comprehensive fashion. In terms of a fundamental project, one can trace back the hierarchies of projects that I have pursued *to an encompassing project that only I can be responsible for*; "we make it exist through our very commitment and therefore we can grasp it only by living it" (Sartre, 2018, p.606; see also, section 7.4).

An additional point to note regarding one's fundamental project is that, as Sartre (2018, p.606) puts it, it is "not first conceived, and then actualized" – instead, it is *lived through*. As per my arguments in chapters 5 and 6, this means that one's fundamental project is not the result of a reflective will, nor can its totality be made explicit through reflection. Instead, my fundamental choice of myself is also my choice of the world (Sartre, 2018, p.606). The overarching structure of my projects also serves to structure my experience of my environment, such that the objects in my surroundings – with the meanings and values that are given with them – come to reflect my choices back to me:

> The value of things, their instrumental role… do nothing other than sketch out my image, which is to say my choice. My clothing…untidy or neat, sophisticated or vulgar, my furniture, my street …everything teaches me, myself, about my choice, i.e. about my being. But the structure of positional consciousness is such that I cannot reduce this knowledge to a subjective apprehension of myself… (Sartre, 2018, p.606-607).

Undoubtedly, there is more that can be said about Sartre's concept of a fundamental project, but my intentions here are specific. My argument is that one's conception of the good *consists in the kind of project within which other projects are integrated*. Although it will organise the world as my choices are reflected back to me, my conception of the good is a choice that only I am responsible for, that will encompass the other projects that I pursue, and that is foundational to both my experience and my deliberations. To hark back to Taylor's description in section 6.5, my good constitutes my 'palette of motivations', being both formed as I navigate the world and taken for granted by my current projects. Though brief, I will now apply this description to the context of social justice.

As I have mentioned throughout this thesis, both Royce (1908, p.167) and Rawls (1971, p.408) endorse the following claim: "I cannot answer the question, 'Who am I?' except in terms of some sort of statement of the plans and purposes of my life." Given my analysis in this thesis so far, the mistake that Royce and Rawls make is now clear. Royce motivates this issue with a question *that demands an explicit answer*: who am I? The error is that both Royce and Rawls take one's answer to this question as constitutive of personhood and as definitive of one's conception of the good (i.e. the systemisation of one's purposes in a life plan). Yet if we remove this issue from its place within a dialogue, we see that there is a response that cannot be posited as an answer to Royce's initial question: it is not 'who I am and what I value' because who I am *is* what I value and *vice versa*. As Sartre (2018, p.79) puts it, my fundamental project "gives rise to the existence of values, appeals, expectations and a world in general… I have no recourse, and I cannot have recourse, to any value to set against the fact that it is I who maintain values in being." It is here, as a project regulative of my other projects, that my conception of the good can be found. My good is not a *statement* of my purposes, but their integration across my life and lived experience.

In summary, my claim is that congruence between goodness and justice obtains insofar as it possible to include one's sense of justice within one's choice of oneself in the world. In this way, the fact that justice as fairness reorganises the basic structure of society and is compatible with my account of *reflection without rationalism*, is of great benefit. Justice is both demanding and transformative. As I have emphasised from the start of this thesis, to change society's basic structure is to change the people within it and *vice versa*. Though only a sketch, I hope that the account detailed here gives expression to Mendus's (1999, p.57-75) description of the connection between love and justice in *Theory*:

> I do not mean to imply that it is a mere personal sentiment or preference, much less that it is capricious or unreliable. On the contrary, drawing out the analogy between love and justice illuminates the sense in which both are (or may be) guiding principles in the lives of individuals, principles which inform their actions and, as Rawls himself puts it, 'reveal what the person is' (Mendus, 1999, p.74)

Whilst I have accepted that we cannot prove congruence between justice and goodness, I have nevertheless attempted – throughout this chapter – to offer the best approximation. Moreover, I have demonstrated that reflecting on the conditions of justice is itself beneficial. When we reflect on the relationship between goodness and justice, we reorient ourselves within a political world. The advantage of justice as fairness is that it offers the assurance that our practical self-conception – that is, a self-conception that we can all, in principle, uphold – is confirmed by the organisation of social order. Consequently, persons within the well-ordered society come together in their identity, whilst acknowledging their differences, recognising that each has the power to fashion their own lives and self-authenticate their major aims and values.

Should this reflective vision come to fruition, it is because all persons have come to affirm – as a project regulative of other projects – the shared sentiment of justice.

Conclusion

Who We Are and What We Value

9. Concluding Remarks and Summary of Argument

In this thesis, I adopted a Sartrean perspective to criticise and reconstruct Rawls's main response to the problem of stability: the congruence argument. Here, I will begin with a general summary of this thesis before considering the implications of my argument. In chapter 1, I identified the primary mechanisms that Rawls uses to establish the stability of his conception of justice, i.e. a system of enforcement and the two generative properties of justice as fairness (i.e. that citizens will view just institutions as independently valuable and transform their private ends). Moreover, I supported this interpretation of Rawls by contrasting his position with a broadly Hobbesian approach, wherein I explored the arrangements necessary for the application of the liberty principle. At the same time, I demonstrated – using McClennen's 'liminal citizens' as a thought experiment – that justice as fairness generates stability-conducive attitudes by appealing to the status of persons as purposive beings. In taking these steps, I established the background requirements for a stable conception of justice, as well as indicated – in the broadest terms – the ways in which Rawls establishes the stability of justice as fairness.

In chapter 2, I brought this discussion on stability to the specific context of Rawls's argument from congruence, i.e. the claim that a person's conception of the good can include, or be harmonious with, their sense of justice in the well-ordered society. In this chapter, I defined the congruence argument as a form of stress-test: once a conception of justice is justified (in the Rawlsian case, *via* the original position), the question arises as to whether it is compatible with certain tendencies in human nature. In framing Rawls's argument in this way, I was able to respond to Barry's counter-claim that the disposition 'to do what is right because it is right' ought to be sufficient motivation for citizens to act justly. Against this reading, I highlighted the fact that persons will have different values, desires, and purposes in the well-ordered society – including the disposition to 'do what is right for its own sake' – such that justice must be supremely regulative in light of this evaluative, epistemic, and normative diversity. Understood in this way, Rawls aims to generate justice-supportive attitudes, verify the scope of justice over our lives, and demonstrate the feasibility of his conception, by having justice be a part of citizens' respective conceptions of the good.

In chapter 3, I focused on Rawls's claim that a person's good is given by their rational plan of life. Specifically, I investigated the technical background to *Theory*'s account of goodness, using it to expand on the temporal and hierarchical nature of rational life-planning. Employing this technical analysis in chapter 4, I defended Rawls's account of the good from one of the main criticisms in the literature: the temporality critique. As I explained there, the temporality critique takes Rawls's account of life-planning to presuppose an abstract and atemporal perspective, thereby missing out on time-sensitive goods and virtues. Responding to this critique, I demonstrated that rational life-planning is a deliberative process whereby, over time and permitting for localised periods of non-planning, one reflects on and organises one's major purposes and desires. Understood in this way, there is no need to assume any form of temporal abstraction in *Theory*'s account, such that one of the main criticisms against Rawlsian life-planning fails to make meaningful headway.

Having responded to the most prevalent criticisms of Rawls's approach to stability, I then begin my own critical treatment of the congruence argument. In the latter parts of chapter 4, I argued that *Theory*'s account of deliberative rationality brings disunity to the self by maintaining that a person's good is settled outside of their capacity for self-formulation. Whilst life plans are not temporal abstractions, they are – when understood in light of Rawlsian objective rationality – an abstracted, idealised, version of a person's good. I clarified that this problem is symptomatic of a broader complication within Rawls's congruence argument, i.e. the question of whether life plans are formulated with complete information and true beliefs, or not. Whilst there are relevant reasons one may wish to suppose the former condition, e.g. it avoids agent fallibility, I demonstrated that this approach is incompatible with Rawls's conception of justice as a whole. Instead, the Rawlsian response to stability involves generating appropriate attitudes, recognising the unity that a person can form out of their lives, and appealing to their capacity for agency.

In chapter 5, I argued that *Theory* assumes a voluntaristic and rationalistic account of agency that is ultimately untenable. To substantiate this reading of Rawls, I analysed *Theory*'s notion of self-reproach and showed that the self is unified when one successfully treats oneself as a singular coherent agent, where the condition for success is actively planning for one's life. In light of this analysis, I attributed to Rawls a form of rationalistic voluntarism that reduces agency to localised, deliberative acts. Subsequently, I adopted a Sartrean standpoint to criticise Rawls's account of agency, arguing that self-reproach – even as *Theory* utilises this notion – involves a range of normative phenomena that resist a voluntaristic explanation.

In chapter 6, I argued that the good life is not given by a rational plan. Instead, our values and commitments unfold before us dynamically, without requiring the intervention of a deliberative or contemplative will. Whilst planning can provide some order to this process, by providing a perspective on pre-existing projects, several epistemic and normative errors arise when a planning approach is used to characterise a person's conception of the good. For one, we fail to account for the sense in which plans themselves are commitments, and thus extend and adapt the very thing that they are meant to contain. Connectedly, taking up a perspective on one's life means distancing oneself from it; relating to oneself as the object, rather than the subject, of experience. Hence, the ascriptions that one makes when drawing up one's life plan are not only inexhaustive, but liable to misrepresent one's main values, interests, dispositions, and so on. In the context of the congruence argument, this proves to be a particularly severe mistake. By failing to adequately characterise the good, Rawls also fails to show that it is compatible with the pursuit of justice.

In chapter 7, I advanced a phenomenological reading of Sartre's account of pre-reflective agency by drawing on the work of Ratcliffe. In doing so, I demonstrated that pre-reflectivity involves a basic, pre-intentional, and world-directed sense of possibilities. I cleared up several relevant misinterpretations of Sartre's position – perhaps most significantly, that he likewise assumes a voluntaristic account of agency – and situated this phenomenological thesis as analogous to the technical notions already underpinning Rawls's account of agency and goodness. In taking this step, I tied up some loose ends regarding the critical portion of this work, whilst anticipating my positive contribution by providing independent argumentation for the phenomenological account of pre-reflectivity. In broader terms, I established what it means to be an agent in the Sartrean sense, thus situating his account as directly relevant to the Rawlsian approach to stability.

In chapter 8, I offered a revised version of congruence that is embedded in a phenomenological account of pre-reflective agency (rather than, as per Rawls in *Political Liberalism*, a comprehensive doctrine). In doing so, I maintained the spirit of the congruence argument by investigating how justice becomes a part of one's life. On the Sartrean-Rawlsian approach, justice must be affirmed as a project that gives shape and directs – without requiring the intervention of a volitional will – the other commitments that one can pursue. Though extensive, the requirements of congruence do not stop there. In times of disruption and upheaval, reflecting on the conditions of justice is a natural and often necessary response. In such cases, one has to assess whether one's current commitments – in this case, the commitments to justice and the values representative of one's good – are compatible with a particular view of oneself. In regard to the well-ordered society, such reflective acts serve to affirm one's view of oneself as a practical agent; a self-conception that all persons can in principle share. In this way, neither justice nor goodness assumes an entirely fixed perspective; rather, they both serve to transform and sustain the values that one upholds and are open to reassessment through reflection.

Throughout this thesis, I have provided a framework through which issues of stability, and specifically congruence, can be further explored. The question now arises as to how Rawlsians should respond to my contribution. On the face of it, there are two main responses for Rawlsians. First, they could stick to *Theory*'s account of stability and expand on it in light of my contribution. As I will explain in a moment, this is the option that I favour. Whilst I have reconstructed Rawls's congruence argument, its connection to the other part of Rawls's framework will need to be analysed in response. For example, Tomlin's (2008, pp.101-116) investigation into Rawls's account of envy maintains that there is a putative connection between *Theory*'s account of stability and the justificatory stage of the original position. If this connection is warranted, Rawlsians will need to be clear about how my Sartrean-Rawlsian approach to stability impacts the broader project of justice as fairness. Crucially, this is an advantage, and not an omission, of my account overall. Even in terms of my critique of Rawls, I have shown that notions in social justice might be interrogated in light of Sartre's phenomenology. That this process may be continued further only shows that a more complete interchange between the two is both possible and worth pursuing.

The second option, relevant to late Rawlsians, is to reconsider Rawls's move to *Political Liberalism* or adapt its account of stability in order to encompass the arguments advanced here. To see why this is necessary, note that my research (i.e. chapters 7 and 8) has shown that Rawls did not need to abandon *Theory*'s congruence argument. To recall, Rawls abandoned the congruence argument because it proved incompatible with the fact of reasonable pluralism. However, having separated the Sartrean-Rawlsian approach to congruence from a commitment to any comprehensive doctrine, and having situated pre-reflective agency as a plausible phenomenological thesis, I have avoided Rawls's original concerns. Congruence is not about acceding to any particular doctrine but accounting for the scope of justice. The Sartrean-Rawlsian approach achieves this goal by attending to the ways in which reorganising society transforms our basic sense of ourselves and the world. Hence, I favour returning to *Theory* because – as this thesis has exemplified– the issue of congruence can be crucial to our theorising about justice on a practical level. Late Rawlsians must provide additional reasons for why the congruence argument should still be abandoned, or else look to incorporate it within *Political Liberalism*'s account of stability – for example, *via* Rawls's (2005, pp.133-173) notion of overlapping consensus.

That being said, it would be a mistake to think that Rawlsians need only (re)consider the commitments internal to justice as fairness as a response to this thesis. On the contrary, my account of congruence also makes clear the need for a debate regarding the normative consequences of stability. As I demonstrated in chapter 2, claims about stability constitute the grounds for what morality (i.e. justice) can conceivably *require*. Theoretically, this means assessing whether justice is compatible with various justified stipulations regarding persons and the world; more practically, it means exploring whether the rules, institutions, and institutional practices of Rawls's approach are conducive to the two generative properties of justice as fairness. In other words, once it is determined that justice is compatible with the conditions of life as we know it, there is an additional level of analysis as to what institutional structures best embody the requirements of justice *in light of those conditions*. To provide a concrete example, Hussain (2012, pp.180-200) argues that justice as fairness is not neutral when it comes to different schemes for private property-owning economies. Instead, using the Rawlsian account of stability, he advances a form of democratic corporatism, which employs the notion of political participation to generate attitudes supportive of justice (more efficiently, that is, than alternative institutional arrangements).

My point is that the Sartrean-Rawlsian approach to congruence should also be employed when investigating the normative implications of stability. That is, the notions of pre-reflective agency, practical self-conceptions, and the dynamic nature of the good, will be useful tools when considering concrete social and economic structures of the just society. For example, applying this approach to *Theory*'s argument from moral development may serve to consolidate the importance of reflecting on justice within the educative process (see, chapter 2). As Rawls (1971, p.473) accepts, those actively involved within the political structure will "often have to take up the point of view of others" in order to determine their own values and interests, as well as balance the putatively competing interests of other citizens. In a similar vein to Hussain, my approach to congruence demonstrates the need to expand this process of reflection outside of those limited few directly engaged with politics. Instead, looking at systems that adopt a more holistic approach to political participation – e.g. that seek to harmonise seemingly non-political activities within a move towards social justice – may be preferred given the account of congruence that I have advanced. After all, adopting justice as a fundamental project means fashioning one's other projects accordingly, even those that are not explicitly politically-orientated; from a Rawlsian perspective, this means giving special place to institutions and institutional practices conducive to this ongoing process.

In this way, the research undertaken in this thesis provides the groundwork for exploring the moral requirements of stability, whilst moving towards a more general reassessment of Rawlsian liberalism, involving both its practical and theoretical commitments. At the same time, this thesis has deepened the relationship between political phenomenological and theories of social justice. To this latter point, Bedorf and Herrmann (2020, p.2) claim that there is a "crisis" in that contemporary political philosophy, since it maintains a "near exclusive focus on normative principles that obscure the first person's perspective." As a result of this narrowing of perspective, they add that "the gap between what *is* and what *ought* to be has become increasingly unbridgeable and the ability to relate to life-world problems of social actors is gradually vanishing" (Bedorf and Hermann, 2020, p.2) This thesis has taken substantial steps in remedying such concerns. As I have maintained throughout, even though Rawls maintains a constructivist approach to the principles of justice, resolving the problem of stability means that these principles must be verified by the conditions of the world. That phenomenology can serve to elucidate those conditions, by directly attending to the first-person viewpoint, has been sufficiently demonstrated by my account of congruence.

In conclusion, one of the main mechanisms that Rawls uses for securing the stability of the well-ordered society is the congruence argument: the idea that the individual pursuit of a good life and the shared affirmation of justice are harmonious with one another. In his original work, Rawls failed to carry this argument through because of three main issues; (i), how he formulates his idea of deliberative rationality, (ii), where he locates the source of agency, (iii), and how he characterises the good life. I have attempted to rescue the congruence argument, which rightly serves to highlight the encompassing influence that justice has on our lives. To achieve this reconstruction, I employed Sartre's phenomenology of agency and demonstrated that a life is not a rationalised list of purposes and desires. In many ways, the good life is something that slips through one's fingers; remodelled and transformed as soon as it is grasped directly. Recognising this point does not mean abandoning the idea that society should be organised according to principles of justice. It is quite the contrary: one’s pre-reflective experience is embedded within a shared world of values and demands, permeated by the normative standard of justice. Ultimately, to strive for congruence is to aim precisely for this world, to show the compatibility of justice, agency, and the good.

Bibliography

Alford, F.C. (1991). *The self in social theory*. New Haven and London: Yale University Press.

Bagnoli, C. (2000). Value in the guise of regret. *Philosophical explorations*. 3(02), pp.169-187.

Banerjee, K. and Burcuson, J. (2015). Rawls on the embedded self: liberalism as an affective regime. *European Journal of Political Theory*.14(02), pp.209-228.

Barry, B. (1995). John Rawls and the search for stability. *Review of A theory of justice; Political liberalism*. Ethics. 105(04), pp. 874-915.

Bedorf, T. and Herrmann, S. (2020). Three types of political phenomenology. In. Bedorf, T. and Herrmann, S. (Eds). *Political phenomenology: experience, ontology, episteme*. New York: Routledge, pp.1-15.

Benson, J. B. (1997). The development of planning: It's about time. In S. L. Friedman and E. K. Scholnick (Eds.), *The developmental psychology of planning: Why, how, and when do we plan?* pp. 43–75.

Bercuson, J. (2014). *John Rawls and the history of political thought: the Rousseauvian and Hegelian heritage of justice as fairness*. New York: Routledge.

Care, N. (1996). *Living with one's past: personal fates and moral pains*. Boston: Rowman & Littlefield.

Cross, M. J. R. (2017). *Communities of individuals: liberalism, communitarianism, and Sartre's anarchism*. New York: Routledge.

Dalton, A.N., and Spiller, S. (2012). Too much of a good thing: the benefits of implementation intentions depend on the number of goals. *Journal of Consumer Research*. 39(03), pp.1-16.

Darnell, T. and Rohatyn, D. (1992). Sartre's debt to Rousseau: freedom, faith, and fulfilment. *Bulletin de la Société Américaine de Philosophie de Langue Française*. A Special Issue on the work of Jean-Paul Sartre edited by Caws, P, pp.244-263.

De Beauvoir, S. (1965). Simone de Beauvoir: an interview. Interviewed by Gobeil, M. *Paris Review*. 35, pp.23-40.

Delaney, P. F., and Ericsson, K.A. (2004). Immediate and sustained effects of planning in a problem-solving task. *Journal of Experimental Psychology*. 30(06), pp.1219-1234).

Detmer, D. (1986). *Freedom as a value: a critique of the ethical theory of Jean-Paul Sartre*. La Salle: Open Court.

Dewey, J. (1916). Voluntarism in Roycean philosophy. *The Philosophical Review.* 25(03), pp.245-254.

Dillon, R.S. (2001). Self-forgiveness and self-respect. *Ethics*. 112(01), pp.53-83.

Eshelman, M. C. (2010). What is it like to be free? In. Webber, J. (Eds). *Reading Sartre*. New York & London: Routledge.

Eshelman, M. C. (2016). A sketch of Sartre's error theory of introspection. In. Miguens, S. Preyer, G. and Morando, C. B. (Eds). *Pre-reflective consciousness: Sartre and contemporary philosophy of mind*. New York & London: Routledge.

Freeman, S. (2007). *Rawls*. London: Routledge.

Freeman. S. (2003). Congruence and the good of justice. In. Freeman, S. (Eds). *The Cambridge Companion to Rawls*. Cambridge: Cambridge University Press. pp. 277-315.

Fried, C. (1970). *An anatomy of values*. Cambridge: Harvard University Press.

Friedman, S. L. and Scholnick, K. E. (1998). *The developmental psychology of planning: why, how, and when do we plan?* New York: Psychology Press.

Galanter, E. (1966). *Textbook of elementary psychology*. San Francisco: Holden-Day.

Galanter, E., Miller, G.A. and Pribham, K.H. (1960*). Plans and the structure of behavior*. New York: Holt, Rinehart, and Winston, Inc.

Gerassi, J. (2009). *Talking with Sartre: Conversations and Debates*. Interview with Sartre, J-P. Translated from the French by Gerassi, J. London: Yale University Press.

Gewirth, A. (1978). *Reason and morality*. Chicago: University of Chicago Press.

Gewirth, A. (1982). *Human rights: essays on justification and applications*. Chicago: University of Chicago Press.

Gewirth, A. (1996). *The community of rights*. Chicago: University of Chicago Press.

Goldman, A. (1970). *A theory of action*. Englewood Cliffs: N.J., Prentice-Hall.

Grossman, R. (1984). *Phenomenology and Existentialism*. London: Routledge and Keagan Paul.

Gurley, S. W. (2016). Attention is political: how phenomenology gives access to the inconspicuously political act of attending. In. Gurley, S. W. and Pfeifer, G. *Phenomenology and the Political*. London: Rowman & Littlefield, pp.233-249.

Hermann, D.H. (1974). The fallacy of legal procedure as predominant over substantive justice: a critique of "the rule of law" in John Rawls' a theory of justice. *DePaul Law Review*. 23(04), pp.1408-1436.

Hershfield, H. E. (2011). Future self-continuity: how conceptions of the future self transform intertemporal choice. *Annals of the New York Academy of Sciences: Decision Making Over the Life Span*, pp.30-43

Heyd, D. and Miller, F. (2010) Life plans: Do they give meaning to our lives? *The Monist*. 93(1), pp. 17–37

Hill, T. E. Jr. (1999). Autonomy and agency. *William & Mary Law Review*.40(03), pp.847-856.

Hill, T. E. Jr. (2014) Stability, a sense of justice, and self-respect. In. Mandle, J. and Reidy, D. A. (Eds). *A Companion to Rawls*. West Sussex: Wiley Blackwell, pp.200-216.

Hobbes, T. (1996). *Leviathan*. Gaskin, J. C. A. (Eds). Oxford:Oxford University Press.

Hohfeld, W, N. (1917). Fundamental legal conceptions as applied in judicial reasoning. *The Yale Law Journal*. 26(08), pp. 710-770.

Hussain, W. (2012). *Nurturing the sense of justice*. In. O'Neill, M. and Williamson, T. (Eds). Property-Owning Democracy. Oxford: Wiley-Blackwell, pp.180-200.

Jung, H. J. (1979). *The crisis of political understanding: a phenomenological perspective in the conduct of political inquiry*. Pittsburgh: Duquesne University Press.

Jung, H. Y. (1982). Phenomenology as a critique of politics. *Human Studies*. 5(03), pp. 161-181.

Kojève, A. (1969). *Introduction to the reading of Hegel*. Assembled by Queneau, R. Edited by Bloom, A. Translated by Nichols, J. H. Jr. Ithaca and London: Cornell University Press.

Kruks, S. (1990). *Situation and human existence: freedom, subjectivity, and society*. Boston: Unwin Hyman.

Laborde, C. (2002). The reception of John Rawls in Europe. *European Journal of Political Theory*.1(02), pp.133-146.

Ladenson, R. F. (1975). Rawls' principle of equal liberty. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 28(01), pp. 49-54.

Larmore, C. (1999). The idea of a life plan. *Social Philosophy and Policy*, 16(01), p.96-112

Larmore, C. (2010). *The practices of the self*. London: The University of Chicago Press.

Legrand, D. (2007). Pre-reflective self-as-subject from experiential and empirical perspectives. *Conscious Cogn*. 16(03), pp.583-599.

Lindahl, H. (2020). Intentionality, representation, and recognition: phenomenology and the politics of a-legality. In. Bedorf, T. and Herrmann, S. (Eds). *Political phenomenology: experience, ontology, episteme*. New York: Routledge, pp.256-277.

Lindsay, J. (1918). Rationalism and voluntarism. *The Monist*. 28(03), pp.433-455.

Mabbott, J.D. (1953). Reason and desire. *Philosophy*. 28(105), pp.113 – 123.

MacIntyre, A. (1984). *After virtue*. Notre-Dame: University of Notre Dame Press.

MacIntyre, A. (1985). Micheal Slote, Goods and Virtues. *Faith and philosophy: journal of the society of Christian philosophers*. 2(02), pp. 204-207.

Mackie, J.L. (1984). Can There Be a Right-Based Moral Theory? in J. Waldron (Eds). *Theories of Rights*. Oxford: Oxford University Press.

McBride, W. L. (1980). *Social theory at a crossroads*. Pittsburgh: Duquesne University Press.

McBride, W. L. (1991*). Sartre's political theory*. Bloomington and Indianapolis: Indiana University Press.

McClennen, E. F. (1989). Justice and the problem of stability. *Philosophy & public affairs*. 18(01), pp.3-30.

Mendus, S. (1999). The importance of love in Rawls's theory of justice. *British Journal of Political Science*. 29(01), pp. 57-75.

Merleau-Ponty, M. (1969). *Humanism and terror: an essay on the communist problem*. Translated by O'Neill, J. Boston: Beacon Press.

Mill, J. S. (1987). *Utilitarianism*. Buffalo: Prometheus Books.

Moran, Richard. (2001). *Authority and Estrangement: An Essay on Self-knowledge*. Princeton: Princeton University Press.

Nagel, T. (1986). *The view from nowhere*. New York: Oxford University Press.

Nancy, J-L. (1991). *The Inoperative Community*. Minneapolis: University of Minnesota Press.

Neuber, S. (2021). Determined to act: On the structural place of acting in Sartre's ontology of subjectivity. In. Erhard, C. and Keiling, T. *The Routledge Handbook of Phenomenology of Agency*. London: Routledge, pp.134-147.

Ratcliffe, M, (2024). [Forthcoming]. We are our possibilities: from Sartre to Beauvoir to Løgstrup. Aho, K. Altman, M. Pedersen, H. *The Routledge Handbook of Contemporary Existentialism*. London: Routledge. https://www.academia.edu/87614000/

Ratcliffe, M. (2017). *Real hallucinations: psychiatric illness intentionality, and the interpersonal world*. Cambridge: MIT Press

Rawls, J. (1971). *A theory of justice*. New York: Harvard University Press.

Rawls, J. (2000). In. Herman, B (Eds). *Lectures on the history of moral philosophy*. Harvard: Harvard University Press.

Rawls, J. (2005). *Political Liberalism: Expanded Edition*. 2nd ed. New York: Columbia University Press.

Rawls, J. (2007). *Lectures on the history of political philosophy*. Freeman, S. (Eds). Cambridge & London: The Belknap press of Harvard university press.

Rawls, J. (2007). *Lectures on the history of political philosophy*. Freeman, S. (Eds). Cambridge & London: The Belknap press of Harvard university press.

Rowlands, M. (2016). Sartre on pre-reflective consciousness: the adverbial interpretation. In. Miguens, S. Preyer, G. and Morando, C. B. (Eds). *Pre-reflective consciousness: Sartre and contemporary philosophy of mind*. New York & London: Routledge.

Royce, J. (1908) *The philosophy of loyalty*. New York: Macmillan.

Sandel, M.J. (1982). *Liberalism and the limits of justice*. New York: Cambridge University Press.

Sartre, J-P. (1948). *Anti-Semite and Jew*. Preface by Walzer, M. Translated by Becker, G. J. New York: Schocken Books.

Sartre, J-P. (1984). *War diaries: notebooks from a phoney war 1939-40*. Translated by Hoare, Q. London: Verso.

Sartre, J-P. (2004). *The Transcendence of the Ego: A Sketch for a Phenomenological Description*. Translated by Andrew Brown. London: Routledge.

Sartre, J-P. (2018). *Being and Nothingness: An Essay in Phenomenological Ontology*. Translated by. Richmond, S. London: Routledge.

Setiya, K. (2017). *Midlife: a philosophical guide*. Oxford: Princeton University Press.

Slote, M. (1983). *Goods and Virtues*. Oxford: Clarendon Press.

Smith, C. (1998). Sartre and Merleau-Ponty: The Case for a Modified Essentialism. In Stewart, J. (Eds), *The Debate Between Sartre and Merleau-Ponty*. Illinois: Northwestern University Press, pp.25-36.

Steinfath, H. (2023). Plans, open future and the prospects for a good life. *Ethical theory and moral practice*, pp.1-14.

Stendhal. (1982). *Souvenirs d'égotisme*. Paris: Folio.

Street, S. (2005). A Darwinian dilemma for realist theories of value. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 127(01), pp. 109-166.

Taylor, C. (1985). *Philosophy and the human sciences: philosophical papers* 2. Cambridge: Cambridge University Press.

Taylor, C. (1989). *Sources of the self: the making of the modern identity*. Cambridge: Cambridge University Press.

Taylor, C. (2015). *Origins of the self and the secular age*. John Hope Franklin Center at Duke University. Available at: Origins of the Self and the Secular Age - YouTube.

Thomas, A. (2010). Alienation, objectification, and the primacy of virtue. In. Webber, J. (Eds). *Reading Sartre*. New York & London: Routledge.

Tolman, E.C. (1932). *Purposive behavior in animals and men*. Berkeley: University of California Press.

Tomlin, P. (2008). Envy, facts and justice: a critique of the treatment of envy in justice as fairness. *Res Publica*. 14(02), pp.101-116.

Waldon, S.M., Patrick, J. and Duggan, G.B. (2011). The influence of goal-state access cost on planning during problem solving. *The quarterly journal of experimental psychology*. 64(03), pp.485-503.

Warnock, M. (1966). *The philosophy of Jean-Paul Sartre*. Paton, H.J. (Eds). London: Hutchinson.

Webber, J. (2002). Motivated-aversion: non-thetic awareness in bad faith. *Sartre Studies International*. 8(01), pp.45-57.

Webber, J. (2009). *The existentialism of Jean-Paul Sartre*. New York & London: Routledge.

Webber, J. (2012). Freedom. In. Luft, S. and Oevrgaard, S. (Eds). *The Routledge companion to phenomenology*. New York and London: Routledge.

Weithman, P. (2024). Rescuing justice and stability. *Philosophy & Social Criticism*, 0(0), pp.1-22.

White, J. (2021), The take-up of Rawls's theory of the good in philosophy of education. *Theory and Research in Education*. 19(03), pp.301-307.

White, J. (2021). The take-up of Rawls' theory of the good in philosophy of education. *Theory and research in education*. 19(03), pp.301–307.

Whiteside, K. H. (1988). *Merleau-Ponty and the foundation of existential politics*. New Jersey: Princeton University Press.

Williams B. (1981). *Moral Luck*. Cambridge: Cambridge University Press.

Williams, B. (2009). Life as narrative. *European journal of philosophy*. 17, pp. 305-14.

Zahavi, D. (2014). *Self and other: exploring subjectivity, empathy, and shame*. Oxford: Oxford University Press.

Zahavi, D. (2015). Phenomenology of reflection. In Staiti, A. (Eds). *Commentary on Husserl's Ideas I*. Berlin: De Gruyter.