Developing Moral Character

Essays on Automaticity, Agency, and Responsibility

Jeroen Rijnders

A thesis submitted for the degree of Doctor of Philosophy

Submitted June 22nd 2018 Disputation January 18th 2019

University of Oslo

Faculty of Humanities

Department of Philosophy, Classics, History of Art and Ideas

Centre for the Study of Mind in Nature

Primary Supervisor

Prof. Dr. Olav Gjelsvik

University of Oslo

Secondary Supervisor

Dr. Elinor Mason

University of Edinburgh

Table of Contents

Abstracts

Abstract (English): Imagine a person, say, an employer, who values gender equality, yet intuitively disqualifies female candidates in a job interview. Or someone else, a judge, who explicitly disapproves of racism, but nevertheless more readily perceives Black people as culpable of alleged crimes. And another, a teacher, who avows opposing classism, while tending to unconsciously evaluate the boys with a strong working-class accent as less talented.

Cases such as these are not difficult to imagine, since they frequently occur throughout all of our everyday lives. What is more difficult, however, is how to theorise about what they mean for the concept of being a 'moral agent': someone who determines their own behaviour in such a way that they can be said to have 'moral responsibility' for that behaviour. 'Automaticity', as exemplified above, has become a firmly established empirical psychological phenomenon. Most of people's behaviour is found to be driven by automatic, affective, unconscious cognitive processes - not by conscious, rational reasoning.

In this thesis, I address the debates on automaticity in both the moral psychology and the moral responsibility literature. Firstly, I critically analyse the concepts that are employed, revealing how they shape the debate as well as the data. Furthermore, I explore a novel, alternative position that relates to both debates. Starting from data on the malleability of automaticity, I argue that developing one's own 'moral character' can be conceptualised as an additional form of agency. In turn, moral responsibility for some automatic behaviour can be grounded in the opportunities one has had to engage with one's character development. I test this approach by comparing it to other theories and argue that, besides being more empirically substantiated, it performs better in evaluating a wide range of moral scenarios.

Sammendrag (Norsk): Forestill deg en arbeidsgiver som verdsetter likestilling, men som underbevisst undervurderer kvinnelige søkere ved jobbintervju. Og se for deg en dommer som eksplisitt misliker rasisme, men som likevel har lett for å oppfatte personer med mørk hud som klanderverdige når de blir anklaget for å ha begått lovbrudd. Eller tenke om en lærer som motsetter seg klassetenkning, men som ubevisst har en tendens til å anse gutter med sterk arbeiderklasseaksent som mindre begavede.

Tilfeller som dette er ikke vanskelig å forestille seg, fordi de ofte forekommer i hverdagen. Det er imidlertid vanskelig å se akkurat hva slike tilfeller

har å si for vårt begrep om å være en 'moralsk aktør'; en person som styrer sine handlinger på en måte som gjør at hen kan sies å være 'moralsk ansvarlig' for sine handlinger. 'Automatisitet' som vi ser i eksemplene over, er et empirisk veletablert psykologisk fenomen. Mesteparten av menneskelig handling har vist seg å være et drevet av av automatiske, affektive, ubevisste kognitive prosesser – ikke av bevisst, rasjonell resonnering.

I denne avhandlingen tar jeg for meg debatter om automatisitet innenfor både moralpsykologi og litteraturen om moralsk ansvar. Først gir jeg en kritisk analyse av konseptene som brukes, og viser hvordan disse former debatten så vel som dataene. Videre utforsker jeg en ny, alternativ posisjon som relaterer seg til begge debattene. Med utgangspunkt i empirisk funn angående automatisitetens formbarhet argumenterer jeg for at utviklingen av ens egen 'moralske karakter' kan forstås som en ytterligere form for aktørskap. Følgelig viser jeg at moralsk ansvar for enkelte automatiske handlinger hviler på ansvar man har for utviklingen av ens moralske karakter. Jeg tester denne tilnærmingen ved å sammenligne den med andre teorier og argumentere for at det, i tillegg til å være mer empirisk underbygget, gir bedre forklaringen av en rekke moralske scenarier.

Introduction

"Keep on learning, and soaking up game,
We gon' make mistakes, we gon' go through some thangs,
Keep on growing, keep on soaking up game,
If something ain't working, don't be afraid to change."

- Dead Prez[1]

This doctoral thesis is comprised of four essays that chiefly concern the role of the concepts of *automaticity* and *character* within morality. While the essays are written as independent works and should initially be read as such, they are unmistakably closely connected to one another, each successive one much drawing on the prior.

In this introduction, I start by shedding some light on the motivating reasons for selecting the topic, its general significance, future research possibilities, and an (immensely succinct and rough, but hopefully intelligible to non-philosophers) explanation of the thesis' central concepts. Subsequently, I survey each essay's main content while identifying some of their links and placing them in the context of the entire project. I close with an acknowledgement section.

Aristotle, moral psychology, and metaethics

In a community centre in Amsterdam, 'volunteer philosophers' (yes, there is such a thing) introduced me to Socratic dialogues, a methodology for philosophical conversations, as a method to critically explore the beliefs and values that guide one's life.[2] Moreover though, it was a gateway into their virtue ethical creed, as I was soon participating in study-groups on ancient text interpretation, reading the modern gospel of Leonard Nelson, and became a parishioner who tutored others and organised

[1] Among the many choral expositions on systemic poverty, racism, and sexism, socialist campaigning, and advocating a healthy lifestyle including sports and veganism, on this hook Stic.Man champions the importance and possibility of working on one's self-development. Especially against the background of the politically valenced message of the duo, their call for self-development relates to the central point of this thesis; that actively developing one's own moral character is crucial to moral agency and moral responsibility. Dead Prez (2012) Learning, Growing, Changing, on Information Age (CD), M1 and Stic-Man (Prod.), Krian Music Group.

[2] See *Socratisch Café Amsterdam*, founded by *Het Nieuwe Trivium's* Jos Kessels, Erik Boers, and Pieter Mostert, preserved and spread through The Netherlands by Karel van Haaften and Tom Sengers.

philosophical congregations myself. But then, when I had only just started more thoroughly studying my newfound interest in virtue ethics by enrolling in philosophy, I was shook by the words of Elizabeth Anscombe. With a sobering observation the British philosopher promptly tempered my enthusiastic dive into Aristotelian scholarship.

In her seminal essay *Modern Moral Philosophy*, in the context of boldly criticising the lack of foundation of all major moral philosophical traditions, Anscombe wrote that, before we can profitably do moral philosophy, we first need to get a better grip of the human psychology that is involved in this. (Anscombe, 1958) This requires examining concepts such as 'action', 'intention', 'wanting', 'pleasure', and 'virtue', which underlie any further thinking about morality. That message has stuck with me ever since, impelling me to put devoted Aristotelianism on hold and shift my focus to the fields of moral psychology and metaethics. As such, although not further drawing much on her work, the current thesis is in its birth very much a result of this, say (but not out loud, as she would have despised the term), 'Anscombian duty'. In addition, Aristotle relentlessly continues to inspire much of my thought, although current inquiries into this are, sadly, best described as mere dabbling.

Automaticity

Now, there are countless 'fundamental psychological concepts' that need investigation, so which to choose? The ones that are central to this thesis are *automaticity* and *moral character*. Automaticity is a psychological phenomenon that consists of fast, automatic, unconscious, and affective mental processes that can influence people's judgement-formation, decision-making, and action-guidance (hereafter jointly 'behaviour'). (Bargh & Chartrand, 1999; Evans & Stanovich, 2013; Sloman, 1996; Stanovich & West, 2000) This influence is rather profound. Over the last two decades, there has been an explosion of research on automaticity in behavioural, developmental, social, and cognitive neuropsychology. Much of this research is taken to evidence that automaticity is so ubiquitous, that most of human cognition may be automatic. As two central figures in this literature, social

psychologist Jonathan Bargh and neuropsychologist Tanya Chartrand, write, “most of a person’s everyday life is determined not by their conscious intentions and deliberate choices, but by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance.” (Bargh & Chartrand, 1999, p. 462)

Especially in relation to morality, automaticity constitutes an intriguing, fundamental, and troubling matter. Automatic states and processes, typically acquired through socialisation into one’s culture, can drive a person to exhibit morally problematic behaviours, which go against the beliefs and values one may explicitly endorse, and do so even despite efforts against it. For example, while a school teacher may cherish egalitarian values, implicit attitudes (a type of automaticity) that they may foster deep down concerning sex, race, and especially class can nevertheless cause them to evaluate a lower-class, Black, or female student’s essay more negatively, more readily notice their transgressions, and fail to perceive their ambitions. (Auwarter & Aruguete, 2008; Battle & Lewis, 2002; Dee, 2005; Downey & Pribesh, 2004; Rist, 1970) Automatic processes such as implicit biases, unconscious stereotypes, motivated cognitions, and emotions, can drive a person to block the consideration of relevant information to change one’s beliefs or behaviours, be swayed by irrelevant situational factors (e.g. that it is sunny), judge someone as less qualified to hire, more easily judge someone as responsible for a crime, more readily (mis)identify someone as holding a weapon and shooting them, and even impact the harshness of sentencing in the court of law. (Blair et al., 2004; Blair et al., 2002; Chaiken et al., 1996; Graham & Lowery, 2004; Greene et al., 2001; Moss-Racusin et al., 2012; Payne, 2005) As such, from relatively minor to more severe effects and the aggregation of all, automaticity has a crucial role in driving morally problematic behaviours, and what is more, a crucial role in reinforcing systemic inequalities in our societies.

To provide some insight into how such research is done (in the lab, besides analysis of real world data), social psychologists, for example, test people’s automatic attitudes towards some group through ‘implicit association tests’, by measuring patterns (speed and accuracy) in how one pairs photos of people (e.g. Black-White or male-female) with certain words (e.g. positive-negative or science-humanities; “helpful”, “hard-working”, “dangerous”, or “physics”, “literature”, “economics”). (Banaji & Greenwald, 2013; Greenwald & Banaji, 1995) Alternatively,

neuropsychologists use ‘functional magnetic resonance imaging’ and other techniques to observe which brain areas are active and in what way they are so when, for example, one tries to block an operant bias, or when one implicitly stereotypes compared to other implicit attitudes. (Gilbert et al., 2012; Ochsner & Gross, 2008)

So, how does moral automaticity relate to moral philosophy? Traditionally, moral psychology and action theory (sub-fields of moral philosophy) have been dominated by what we may generally call a ‘rationalist’ paradigm. Rationalism holds that ‘moral agency’, the idea that a person is the ‘agent’ of some behaviour, someone who ‘performs actions’ rather than one who ‘occurrences happen to’, is due to the person’s conscious reasoning in some way determining the behaviour. (Kohlberg, 1973; Korsgaard, 2008; Piaget, 1932; Velleman, 2000) Reasoning is emphasised because through the involvement of reasoning the person can intentionally guide the behaviour so that it is rational and moral, in accordance with one’s beliefs and values. In turn, venturing into metaethics (a further sub-field of moral philosophy), most theories of moral responsibility employ some such notion of moral agency to justify our practices of evaluating a person as ‘responsible’ and ‘blameworthy’ for some behaviour. (Levy, 2005; Smith, 2005)

As should be evident, the empirical findings that people’s moral behaviour is largely driven by automatic processes rather than by reasoning, as we saw above, has immense bearing on the concepts of moral agency and moral responsibility. This has led to a range of authors arguing for some version of the *automaticity challenge to moral agency*, which holds that; since moral behaviour is mostly driven by automatic processes rather than reasoning, and agency is marked by the latter, people typically lack moral agency. And based in that thesis, they have reinvigorated varieties of a ‘sentimentalist’ position, which stresses the role of emotions in moral behaviour, downplaying the role of reasoning as rationalists have it. (Blasi, 2009; Doris, 2002; Haidt, 2001; Haidt & Bjorklund, 2008; Nichols, 2004; Prinz, 2007) Subsequently, automaticity has also sparked a debate on responsibility, say, the *automaticity challenge to moral responsibility*. Since people may lack agency over their automatic behaviours, this may exclude a large portion of behaviour from the sphere of responsibility. (Levy, 2016; Smith, unpublished)

Moral character

Returning to the thesis at hand, the central question that runs through all the essays is how best to understand the 'nature' of automaticity; what automaticity is. For, while the philosophical debates above are informed by empirical data, that data is, in turn, informed by the philosophical concepts that we use.

One crucial such concept here, is *moral character*. While clearly greatly inspired by Aristotelian and contemporary virtue ethics, in which character is one of the central notions, the scope of this thesis, unfortunately, does not allow me to explore Aristotle or virtue ethics. I employ moral character as an umbrella term for a large family of mental phenomena. As such, on a very loose definition, a person's moral character is constituted by phenomena ranging from one's beliefs, values, affective attitudes, and behavioural dispositions, to one's skills, sensitivities, and rational capacities, among other things. With that, most automatic states and processes, such as implicit biases and unconscious stereotypes are meant to be part of one's character.

There are several further key features of character. For one, character operates automatically. 'Acting from character', is behaviour that is 'spontaneous', 'intuitive', or in other words, 'automatic'. With this, I aim to provide a general discussion of automaticity, embracing the various forms it can take. Secondly, all the factors listed above are highly intertwined with one another. For example, the implicit sexist attitudes one holds are typically tied up with beliefs one holds 'deep down' (e.g. "men lack empathy and are thus bad parents") and values one holds (e.g. "men deserve less parental rights than women"). Thirdly, one's character is an important determinant of one's moral behaviour, influencing one's judgements, decisions, and actions, and typically doing so fairly quickly, and with little effort and reasoning. Fourthly, the principal feature of character for the sake of this thesis is that a person can develop one's own character. One can change the beliefs, values, biases, and such, which one fosters deep down, through effortful self-development over longer periods of time.

This brings us to stating the main objective of this thesis project: To explore the role of an agent's capacity to develop one's own moral character in the light of moral automaticity. To be more precisely, two parts can be distinguished. Firstly, exploring whether character development can serve a role in a possible defence of

moral agency from the automaticity challenge. And secondly, exploring whether character development as a mode of agency can serve a role in grounding moral responsibility for behaviour driven by automaticity. In this exploration, I aim to acknowledge the empirical data as well as the philosophical tradition, and the various positions on each topic, in order to arrive at the type of 'foundational work' that Anscombe saw necessary for further progress in moral philosophy. Bringing these literatures together, I hope, will provide such solid ground (although doing so may, at times, require lengthy elaboration of evidence and arguments, unavoidably making the thesis fairly sizeable. I chose to utilise the medium of a thesis to bring together plentiful material, which can subsequently be shortened and divided for journal publications).

Background and future

To conclude, after having returned to the Anscombian philosophical motivation I started out with, there is an additional dual motive that is worth mentioning.

A private reason for this topic is that the developmental trajectory of character and automaticity has much affinity with people's general, personality development trajectory, which has been hard-fought for me. From an initial shoddy formation through the environment one grows up in, to later arduously developing oneself to become ever-slightly more virtuous and flourishing, I have a deep, personal connection with much of the content in this thesis. Moreover, in turn, this is a political motif for the topic. While my development and venturing into academia may have been hard-fought, it was also in many ways fortunate, and I am reminded everyday of the lack of opportunity that many others have due to similar factors, which forces solidarity.

For example, even here at the University of Oslo in Norway, which is widely considered as one of the most egalitarian societies (with one of the most extensive social support systems and highest social mobility through offering nearly free education to everyone), students from non-skilled working class backgrounds are 30-35 times less likely to obtain any degree than those from educated families, not even going into advancement to a doctoral level. (Hovdhaugen, 2013, §3.11/6.2) The

‘socioeconomic glass ceiling’ is not a mere effect of financial factors, or of (typically lacking, in the case of socioeconomic background) admission policies. In addition to economic capital, it is importantly impacted by ‘social’ and ‘cultural capital’ as well, influencing, for example, one’s relationships, mannerisms, accent, appearance, and shared references, among many other factors – factors which, it should be clear through this thesis, can strongly shape people’s behaviours and thus reinforce the increasing class-divide. (Byrne, 2015; NSF, 2015; Pain, 2014; Sutton Trust, 2013; HEA, 2013)

As such, researching the morality of automaticity allows me to address some of the psychological mechanisms and philosophical concepts that are key to these, and many other forms of systemic inequality and injustice that have played an important role throughout my life and my environment, such as classism, racism, sexism, and homophobia. As a ‘blue collar scholar’, I see it as a virtue to connect academia to intersectional solidarity.

Hopefully, in the future, I will have the opportunity to continue researching many possible further topics related to this thesis. For one, there is an important tension between ‘individual’ moral responsibility, explored here, and ‘collective’ moral responsibility. With socialisation into one’s culture being a prominent factor of character development, it would be worth addressing character development from the collective perspective, in addition to investigating the dynamic between individual and collective responsibility in such systemic, formative matters. Secondly, with most research focusing on race, sex, and sexual orientation, ‘poverty’ remains underexplored as a factor in automaticity, agency, and responsibility. The developmental approach explored here may offer a bedding for examining and incorporating this more. Thirdly, a further extension of the developmental approach may involve exploring the ways in which people can be convinced of certain morally important information, since automaticity shows this is often not a rational process. Moreover, besides convincing, it is yet another matter how people may become motivated to engage with developing their moral character. And finally, as last future exploration, I would much enjoy connecting the thoughts explored here to Aristotelian and contemporary virtue ethics, in order to investigate in what way they can inform one-another, further refining the related ideas.

The four essays

The first essay, 'A Blooming Impasse in Moral Psychology', is a critical examination of the moral psychological literature (both in philosophy and psychological sciences) on automaticity and how it leads up to the *automaticity challenge to moral agency*. The goals of this essay are quite modest; overview, structure, and analysis.

In the first and second section, I overview the discourse of, respectively, the initial automaticity research and the sentimentalists who take this as a challenge to the rationalist notion of moral agency, and the rationalist replies in order to defend moral agency. In these sections, I aim to impose more structure on the often vague and evasive debate, attempting to pin it down as a more concretely defined argument consisting of a series of premises and conclusion. I present the main empirical claims and group them as two sets of empirical premises; the *primacy of automaticity*, and the *frailty of reasoning*. Subsequently, I add a normative premise; the *deliberative standard of moral agency*. From this, I present the conclusion of *the lack of moral agency* as proposed by some. (Blasi, 2009; Doris, 2002; Haidt, 2001; Haidt & Bjorklund, 2008; Nichols, 2004; Prinz, 2007) In the second section, this argument is slightly changed, mainly based on alternative data in favour of rational reasoning, but not much, for reasons discussed in the next section. (Dreyfus & Dreyfus, 1991; Hogarth, 2001; Holroyd & Kelly, 2016; Kennett & Fine, 2009; Musschenga, 2011; Narvaez, 2011; Pizarro & Bloom, 2003; Sauer, 2012; Snow, 2006)

In §3, I aim to 'negatively' contribute to the debate by criticising it through a conceptual analysis of the philosophical notions that are employed. Through breaking down the (often only implicitly stated) parts of the normative premise that can be discerned throughout the debate, I tease out a set of 'conditions' that form the standard for moral agency. I attempt to show how these conditions do not only make up the normative premise, but also function as paradigm for the empirical research. Moreover, I argue that the agency condition, as a unified set of conditions, is so strict that it severely limits the conceptual space for agency, whereby empirical research finds little of it, and the norm cannot be met through philosophical re-interpretation of the findings. One of the main ways in which this notion of agency shapes the debate, is by construing the automaticity challenge as an 'operant' challenge, one that takes place within evaluating what cognitive processes are dominant while operant in

determining some moral behaviour. This construal conceptually places the development of cognitive processes, say, the configuration of one's automaticity, largely outside of the discussion.

While few would actually defend the unified set, each condition in itself is intelligible. Nevertheless, tacit commitment to it holds the debate in a deadlock. Thus, I conclude, searching for a way to meet the automaticity challenge will initially require scrutinising the concept of moral agency.

The second essay, 'Moral Agency, Automaticity, and Character', (as well as the third), has a bolder aim as part of a doctoral thesis by, being framed as exploring a 'positive', novel account as alternative to the existing discourse. However, as noted in the essays, nothing hangs on these being independent accounts; while some might argue that they are not sufficiently distinct, and thus 'merely' expand existing accounts, my objectives are achieved if the principal points I argue for are embraced, under whatever title that may be. Their framing as distinct accounts serves to spell-out their central points as clearly as possible by considering their most strict instantiation and pushing them as far as possible.

In this second essay on moral psychology, after an introduction to automaticity (§1), I start out by proposing a *tripartite model of moral agency*, on which agency is seen as not one, monolithic phenomenon, but coming in three related, but distinct modes (§2). The first mode of exhibiting agency is *deliberative agency*, which can be thought of in traditional terms of conscious, rational deliberation as driving-force of one's moral behaviour.

The second mode, which I elaborate in §3, is *moderative agency*, which involves an agent's reasoning regulating the behavioural influence of their automaticity. For example, as one moderative strategy holds, an agent can consciously notice one's initial biased intuition and subsequently impede its further influence on one's behaviour. This mode, I argue, is what most of those who defend (a rationalist notion of) agency, in the face of the automaticity challenge, talk about (see essay 1). By appreciating their accounts as forms of moderative agency, I argue, we can most fully appreciate their work as a fruitful contribution to substantiating that mode, rather than refuting them for failing to defend agency on the whole. To this latter point, I discuss five weaknesses of moderative agency; reliance on scarce cognitive resources; difficulty in estimating correction-strength; reliance on strategies

that agents often lack environmental opportunity for; the difficulty of foreseeability that is required due to its specificity; and the perpetual need to be exhibited due to mostly aiming at mere behavioural effects.

The third mode is, elaborated in §4, is *developmental agency*, which is not concerned with determining some moral behaviour (action-focused), but rather with the development of one's own moral character (agent-focused). *Moral character* I define as an umbrella term for a family of automaticity phenomena; mental states and processes such as implicit biases, unconscious stereotypes, intuitions, attitudes, dispositions, habits, beliefs, values, even perceptive and attention patterns and capacities, among other things. As such, actively developing one's own character, one can 'reconfigure' the content of one's automaticity. For example, through reflecting on the lack of grounds of one's sexist biases, one can actually diminish the possession of such biases. And since one's character is agentively shaped, I argue, behaviour that is driven by it is also agentive in turn.

I continue to make the case for developmental agency as plausibly the most important of the three ways in which one can exhibit agency. To this end, I discuss three crucial advantages it has. Firstly, it may be more rational, tracking moral truth better, partly due to not having to take place in the 'heat of the moment' of a moral situation. Secondly, it can be exhibited more frequently, as it is not plagued by the scarcity of cognitive resources, one of the main challenges for operant agency, since one can engage in development at virtually any time. And thirdly, conceiving development as a form of agency opens up a whole range of new ways in which people can exhibit agency. Expanding on this latter point, the end of the essay reviews a collection of developmental strategies, which I categories as *cognitive* and *experiential* strategies. For example, learning about automaticity, taking tests to discovered one's own biases, and reflecting on the (lack of) grounds of beliefs one holds can influence the configuration of one's moral character. And interaction with counter-stereotypical people from stigmatised groups, especially *derivatively*, for example by reading about them, and *imaginatively*, purely fictional exercises, can change one's moral character as well.

The third essay, 'Moral Responsibility, Automaticity, and Character', addresses the automaticity challenge in relation to theories of moral responsibility. Moreover, the account explored here attempts to translate the moral psychological account explored

in essay 2 to the responsibility literature, by drawing on the empirical plausibility of automaticity, tripartite agency, and moral character. As said above, the aim of this essay is bolder as well, being framed as a novel account, although the points I argue for may be assimilated in existing accounts.

To start, I introduce the automaticity challenge to moral agency, how it relates to moral responsibility, and the two main theories of moral responsibility; volitionism and attributionism. These theories address what the conditions are in virtue of which it can be justified that an agent is responsible for some moral behaviour they exhibit. Volitionism, in short, holds that agents are responsible for some behaviour in virtue of some conscious choice or other form of 'control' being involved. (Fischer & Ravizza, 1998; Levy, 2005, 2016; Mele, 2006; Rosen, 2004; Vargas, 2013) Attributionism, in contrast, sees agents as responsible when some behaviour reflects the agent's 'deep self' in the right way. (Adams, 1985; Arpaly, 2003; Faraci & Shoemaker, 2010, 2014; Scanlon, 1998; Sher, 2009; Smith, 2005, 2008)

In §2, I start by sketching the outlines of an alternative, *developmentalist* account. First, I discuss a hypothetical moral scenario in order to introduce the main elements of a developmentalist evaluation of responsibility. For this, I introduce tripartite agency and moral character, and, concerning the latter, further expand on agency as the 'opportunity for character development', which involves both rational capacities as well as environmental circumstances.

The main body of the essay, in §3, puts the developmentalist view to the test by applying it in evaluating a range of moral scenarios, most of which are found in the literature. Through the first few scenarios, I compare the explanatory power in contrast to attributionism. Attributionism, I argue (partly drawing on established critiques by others), cannot satisfactory account for differences in agents' involvement in or opportunities for becoming who they are as a person, for example due to their sociocultural environmental background. The main reason for this is that attributionism grounds responsibility in the character one happens to possess, and telling a further story about responsibility for the acquisition or development of this would require volitionist control notions that it aims to keep out. As developmental involvement and opportunity can be important factors of responsibility, this is a deficit of attributionism, which developmentalism does not face.

Expanding the set of scenarios, I compare the developmentalist and volitionist explanatory power. 'Structuralist' type of volitionist accounts, I argue, also fail to

appreciate involvement and opportunity for developing one's moral character. By focusing on an agent's current agentive capacities, the background of such capacities is disregarded. 'Historical' volitionist accounts perform much better in this respect, by invoking a notion of 'indirect control', which pertains when an agent's failure to meet responsibility conditions can be 'traced' back to an earlier point in time where responsibility can be located. However, due to central commitments of volitionism, such indirect control is necessarily quite limited in terms of the 'foreseeability' of future situations, the 'range of factors' that may be included (development of character), and the 'behavioural and temporal demarcation' (numerous behaviours over an extended period of time). In the of these shortcomings, developmentalism performs better.

In §4, I propose a sharper definition of developmentalism, based on the features teased out in throughout the earlier discussion. To conclude, consider one last, famous, moral scenario concerning Highsmith's *The Talented Mr. Ripley* personage Herbert Greenleaf, to exhibit how developmentalism can provide rich evaluations of moral responsibility.

The fourth and last essay, 'Discrimination in the Bedroom', is an experimental implementation of the ideas explored in the prior essays to applied ethics through a detailed case study. In order to push the idea of character development in agency and responsibility farthest, the case under consideration involves an almost entirely unexplored form of automaticity, in which virtually all types of discrimination interact, making it an ideal case for a rich, intersectional analysis.

As the topic is practically unexplored, §1 is a fairly lengthy introduction. I start by describing what I take 'sexual preferences' to be (attitudes concerning potential sexual or romantic partners) and how some of these may be morally problematic; when targeting traits such as race, sexual orientation, sexual identity, physique, and class. For example, people hold preferences such as "no Asian women", "no effeminate men", or "only men over 1m80". (Coleman, 2011; Emens, 2009; Halwani, 2017; Thomas, 1999; Zheng, 2016)

In §2 I attempt to make the strongest case for what could be a classical liberal view. This position draws on a classical view of cognition, and a liberal ideal of personal freedom. Applied to sexual preferences, it argues that problematic preferences are; a non-moral matter (they affect partner choices, which is a private

domain, and do not influence anything else); moral but justified (the function of preferences is to select); and/or moral but excused (preferences are beyond agentive control). (Callander et al., 2012; Halwani, 2017; Matheson, 2012; Mills, 1994; Watts, 2012)

In §3 I start by reviewing models of cognition that could serve as basis for the classical liberal view, arguing that the 'fixed configuration' and 'modular operation' seems implausible. I then continue by considering alternative, more plausible models, which appreciate the malleability and interconnectedness of cognitive states and processes. Based on that foundation, I explore a dynamic developmental view of sexual preferences. I argue that, since sexual preferences are crucially tied up with other attitudes (e.g. racial sexual preferences with general racial biases), they influence much other behaviour, such that they cannot be treated as domain-specific and thus non-moral. Moreover, even operating within the domain of sexuality, they cause many harms that are too grave to dismiss as non-moral or justified by a function argument. And thirdly, according to empirical research, sexual preferences seem to be fluid and thus within the scope of one's developmental agency. All together, I conclude, certain sexual preferences are morally problematic and within the moral responsibility of the agent.

Acknowledgements

First and foremost I wish to especially thank my supervisor, Olav Gjelsvik at the University of Oslo and co-supervisor Elinor Mason at the University of Edinburgh. I am immensely grateful for the abundance of detailed feedback you have given me, and at least as grateful for the support and care you have shown through frequent meetings in order to guide me through the process of researching and writing a thesis.

In addition, I wish to thank many others in (and around) academia, who have contributed to my thesis through reading and commenting on drafts, discussing ideas with me, helping out administratively, and making me feel more at home at university. Thank you to Solveig Aasen, Andres Brekke Carlsson, Florent Le Chuiton, Lars Christie, Dag-Erik Eilertsen, Beate Elvebakk, Jola Feix, Hilde Finje, Eirik Finne, Mirela Fuš, Frøydis Gammelsæter, Nick Hughes, Torfinn Huvenes, Kristoffer Jakobsen, Guy Kahani, Daniel Kelly, Anne Løddesøl, Ingvild Bergom Lunde, Ainar Miyata-Sturm, Ole-Martin Moen, Grethe Netland, Nick Novelli, Jon Anstein Olsen, Kim Pedersen Phillips, Anne-Lise Bækholt Pound, Monica Roland, Eduardo Saldaña, Håvard Kroká Saunes, Theodore Scaltsas, Maria Seim, Sascha Settegast, Feroz Mehmood Shah, Maureen Sie, Øystein Sjøtveit, Angela Smith, Robert Sparrow, Caj Strandberg, Aksel Braanen Sterri, Kristoffer Sundberg, Austeja Tamaliunaite, Lina Tosterud, Leo Townsend, Franco Trivigno, Pål Ulleberg, Joost Vecht, Tillmann Vierkant, Sara Kasin Vikesdal, Sebastian Watzl, and Yelena Yermakova. This includes the University of Oslo, the University of Edinburgh, the Centre for the Study of Mind in Nature (CSMN), the Philosophy, Political Theory, Psychology, and Economics Club (PPPE), the Practical Philosophy Working Group (PPWG), the Mind and Cognition Group, and the Philosophy of Sexuality Research Group (PHoK).

A special thanks to Frederick Nathanael, who has shown incredible care and skill as a therapist throughout these years, helping me not merely to get through the process of writing a thesis, but to continue growing as a person while doing so.[3]

[3] While the majority of both students and academics suffer from mental health issues, it is still a much-stigmatised issue with, consequentially, hardly any supportive policies. (Eisenberg et al., 2009; RAND, 2017; Hysenbegasi et al., 2005; Ibrahim et al., 2013; Watts & Robertson, 2011)

Lastly, I wish to thank my family, Jaap Nijhoff, Yvette Willems, Roland Willems, Anja Rijnders, Rob van der Staay, Wim Heldens, Teun Dijkstra, Jacques Willems, and Edith Zwaaga-Willems. Finally, unquestionably categorised among my family, my greatest gratitude goes out to Jelle Bruinsma, who has been the most loving and loyal friend throughout all these years and continues to support me personally, motivate me academically, inspire me politically, and simply overall promotes ‘soaking up game’.

A Blooming Impasse in Moral Psychology

A Conceptual Analysis of the Automaticity Challenge to Moral Agency

Abstract: *Based in empirical findings, the automaticity challenge to moral agency has been advanced as a challenge to rationalist theories of moral agency. The challenge holds that, since affective, unconscious processes mostly determine people's moral behaviour, they typically lack conscious, rational control. This topic has been hotly discussed over the last years, with various authors arguing in favour or against it. Nevertheless, the debate seems to be stuck in an impasse, where two camps disagree on whether there is empirical room for a little bit more or less agency. Through conceptual analysis, I aim to show that the debate is in this impasse due to tacit philosophical commitments to a restricted concept of moral agency, which underlies and restricts the space for debate.*

Introduction

"So, I pull over to the side of the road.
I heard, 'Son, do you know why I'm stopping you for?'

'Cause I'm young and I'm Black and my hat's real low?
Do I look like a mind reader, sir? I don't know.
Am I under arrest, or should I guess some more?'

'Well, you was doing 55 in a 54.
License and registration and step out of the car.
Are you carrying a weapon on you? I know a lot of you are.'"

- Jay-Z[4]

The lyrical account above of Jay-Z's own, lived experience in the 1990s in New Jersey provides a vivid instance of an only fairly recently unearthed, yet highly prominent class of moral behaviour; behaviour driven by *automatic* cognitive processes. To illustrate, we can discern various cognitive processes that are at play, which influence the police officer's perception, judgement-formation, decision-making, and action-guidance in his targeting the maestro on the basis of generalised characteristics rather than individual traits or behaviours. It has been documented at length that such racial profiling was a common practice in New Jersey of the time. (Farmer & Zoubek, 1999) Moreover, it is increasingly clear that such processes are a common characteristic of people's moral cognition in general. Through unconscious stereotypes, implicit biases, and prejudice of Black people are violent and criminal, people more keenly discern their violations (Blair et al., 2002; Eberhardt et al., 2004), more often misidentify something they hold as a weapon (Correll et al., 2002; Payne, 2005), are more willing to shoot them (Kahn & Davies, 2011), and even pass criminal sentences more readily and severely. (Blair et al., 2004)

A large body of empirical findings on human cognition, such as the above, has been taken to show that most of people's behaviour is driven by what is called *automaticity*. Automatic cognitive processes are affective, unconscious, and effortless

[4] Jay-Z (2004) *99 Problems*, on The Black Album (CD), Rick Rubin (prod.), Roc-A- Fella Records and Def Jam Recordings, U.S.

processes that causally determine someone's judgements, decisions, and actions. This phenomenon is thought to challenge the classical rationalist notion of agency, the idea that people determine their behaviour through conscious, controlled reasoning. Especially in relation to moral behaviour, this has become a much-debated issue, known as the 'automaticity challenge to moral agency'. While the debate may appear to mainly be an empirical matter that potentially has radical philosophical implications, in this essay I take a step back to analyse the philosophical premises. More specifically, the central question here is what the philosophical conception of 'agency' is that is employed throughout the automaticity debate, and how this particular concept influences the debate.

I start with an overview of the automaticity challenge, discussing what the main claims are, and how various authors support a version of this challenge (§1). Subsequently, I discuss the main arguments that various opponents of the automaticity challenge put forth (§2). From this, I argue that the debate seems to be at an impasse between two positions on what frequency of moral reasoning can be supported with empirical data. I then turn to a conceptual analysis of the particular concept of agency that is employed throughout the debate (§3). This analysis aims to reveal the, mostly tacit, conceptual commitment to a very strict set of conditions of agency, which severely limits what type of reasoning processes can be embraced as constituting agency. I argue that it is due to these philosophical commitments that the debate is at an impasse concerning the empirical data. To conclude, I propose that any proper progress regarding the automaticity challenge will not as much hang on new data within the current conceptual framework, as much as on a revision of how agency is conceptualised.

§1: The automaticity challenge to moral agency

In this section, I will provide an overview of the literature on automaticity, discussing the background of the research tradition, the empirical findings, categorising the claims they purportedly substantiate, and spelling-out how this leads to the *automaticity challenge to moral agency*.

1A: Studying human behavioural cognition

Over the last two decades there has been an explosion of research on the automaticity of human cognition in behavioural, developmental, social, and cognitive neuropsychology. Much of this research is taken to evidence that automaticity is so ubiquitous, with the majority of people's mental processes occurring nonconsciously, that most of human cognition may be automatic. As social psychologist Jonathan Bargh and neuropsychologist Tanya Chartrand write, "most of a person's everyday life is determined not by their conscious intentions and deliberate choices, but by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance." (Bargh & Chartrand, 1999, p. 462) Now, to be clear from the start, none of such statements are meant as a categorical claim that conscious reasoning processes are entirely absent, but rather that, while various sorts of processes are probably jointly involved at any time, automatic processes have the upper hand. Moreover, while recognising that cognitive processes often do not exhibit all of the defining features in concert, such that they cannot neatly be identified by the co-occurrence of all of the dichotomous features of a certain processing system, these concepts still provide a useful paradigm for understanding moral cognition.[5] (Bargh, 1994; Schneider et al., 1984; Shiffrin, 1988)

Especially in relation to moral judgement-formation, decision-making, and action-guidance (hereafter jointly referred to as *behaviour*)[6] automaticity constitutes an intriguing, fundamental, and troubling matter. For most of the last century (if not longer), moral psychology and action theory was dominated by a rationalist paradigm. One of its most prominent advocates of rationalist moral psychology was Lawrence Kohlberg who advanced a cognitivist model based on the earlier work of Piaget, with further succession in, for example, Turiel and Rest. (Kohlberg, 1973; Nucci & Turiel, 1978; Piaget, 1932; Rest et al., 1999) Rationalist approaches emphasise the role of conscious moral reasoning as causally determining moral behaviour. At the turn of this century however, alongside the focus on automaticity in psychology in general,

[5] Critical analysis of the applicability of these concepts is part of this essay. In essay 2 of my doctoral thesis, I argue for further conceptual changes, in particular different notions of 'control'.

[6] While some may specifically discuss one specific process, I take discuss judgements, decisions, and actions as a cluster, because of their intimate connections, such that the former two processes ultimately relate to the latter, and especially the latter is most morally relevant. Moreover, as many of the authors mentioned in this essay acknowledge, as a function of their connectedness, the three processes are all susceptible to similar automaticity challenges.

moral psychology has been marked by a renewed interest in alternative theories to rationalist ones, which instead focus on emotions. This has resulted in the conception of *moral automaticity*, the idea that automatic, emotional processes drive most of people's moral behaviour, not conscious reasoning. This view is substantiated by a range of empirical findings that are connected to one another. Before going through these, consider this brief scenario as a paradigm case of moral automaticity, to make more concrete what kind of phenomenon we are discussing.

1B: A scenario of moral automaticity

> *Sara is working as a personal trainer at a gym. Since the person usually conducting interviews with potential employees is sick, Sara was asked to fill in. This came unexpected to her, as it is not part of her function. Without time to prepare, Sara receives the candidates. While evaluating each person, both during the interviews and while deciding whom to hire afterwards, she is mainly driven by the feelings they invoke in her. Bjørn has a glorious impression on Sara, as she is smitten by his stern and muscular appearance. Shanice, on the contrary, comes across as unfitting, for, despite her superior résumé and the gym's equal opportunities policy aiming at gender balance, which requires several more women trainers, her ethnic background elicits negative attitudes in Sara. To make things worse for Shanice, while it is her turn for the interview, the beaming early-spring sun that helped make Bjørn seem so gleaming is momentarily blocked from brightening the office by a passing cloud. Nevertheless, when Sara later on meets with the human resources manager and is asked about her choice, being unaware of the personal and situational factors that influenced her, she voices various constructed reasons for her preferring Bjørn, although these were in fact hardly involved in actually shaping her judgement and decision.*

1C: Modelling moral automaticity

To get an initial grasp what automatic processes are we can look at two-system theories of cognition (also known as 'dual-process theories', among many other names), which form the basis of most of the automaticity literature. (Bargh & Chartrand, 1999; Evans & Stanovich, 2013; Stanovich & West, 2000) Two-system theories divide human cognition into the automatic 'system 1' and controlled 'system 2' processes, which are defined in opposition to one another, as can be seen in table 1.

Automatic processes are emotional processes that operate largely outside of conscious awareness and do so with little effort and high speed. Controlled processes, in contrast, are marked by deliberative reasoning, which is conscious, slow, and effortful. Now let us specify each of the main claims about automaticity as a set of theses about moral cognition, which will together lead up to the automaticity challenge.

	System 1	System 2
Dual-Process Theories:		
Sloman (1996)	Associative system	Rule-based system
Evans (1984; 1989)	Heuristic processing	Analytic processing
Evans & Over (1996)	Tacit thought processes	Explicit thought processes
Reber (1993)	Implicit cognition	Explicit learning
Levinson (1995)	Interactional intelligence	Analytic intelligence
Epstein (1994)	Experiential system	Rational system
Pollock (1991)	Quick and inflexible modules	Intellection
Hammond (1996)	Intuitive cognition	Analytical cognition
Klein (1998)	Recognition-primed decisions	Rational choice strategy
Johnson-Laird (1983)	Implicit inferences	Explicit inferences
Shiffrin & Schneider (1997)	Automatic processing	Controlled processing
Posner & Snyder (1975)	Automatic activation	Conscious processing system
Properties:	Associative	Rule-based
	Holistic	Analytic
	Automatic	Controlled
	Relatively undemanding of cognitive capacity	Demanding of cognitive capacity
	Relatively fast	Relatively slow
	Acquisition by biology, exposure, and personal experience	Acquisition by cultural and formal tuition
Task Construal:	Highly contextualized	Decontextualized
	Personalized	Depersonalized
	Conversational and socialized	Asocial
Type of intelligence:	Interactional (conversational	Analytic (psychometric IQ)

(Table 1: *The terms for the two systems used by a variety of theorists and the properties of dual-process theories of reasoning*. From: (Stanovich & West, 2000, p. 659).)

Starting with conceptualising intuitions, some crucial features are theses concerning speed, unconsciousness process, affective processing, situational triggers, and the overall constitution.

Being formed quickly, intuitive responses have primacy over slower conscious reasoning processes (the *speed thesis*). (Bargh & Chartrand, 1999; Fry & Hale, 1996; Greenwald & Banaji, 1995; Haidt, 2001) What is more, speed goes along with the processing being emotional and unconscious (the *unconscious thesis* and *affect thesis*). As Haidt writes, "moral intuition can be defined as the sudden appearance in consciousness of a moral judgement, including an affective valence (good-bad, like-

dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion (…): One sees or hears about a social event and one instantly feels approval or disapproval." (Haidt, 2001, p. 818) Neuroimiging studies find that brain regions related to emotions are more active during moral judgement-formation, while not during evaluating merely socially awkward situations (e.g. when observing unjust punishment, or unfair monetary divisions). (Berthoz et al., 2002; Heekeren et al., 2003; Phan et al., 2002; Singer et al., 2006) Moreover, automatic processes drive moral cognition to such an extent that they are both necessary and sufficient for doing so (the *constitutive thesis*). (Damasio, 1994; Greene et al., 2001; Wheatley & Haidt, 2005) As Jesse Prinz concludes from such data, "moral judgments and emotions seem to coincide in the brain (…). The natural explanation of these findings is that moral judgments are constituted by emotional responses." (Prinz, 2007, pp. 22-23)

Often, such unconscious, emotional processes come in forms such as implicit biases and unconscious stereotypes, which can run against a person's consciously, endorsed values (the *stereotype thesis* and *bias thesis*). For example, priming of racial stereotypes leads to judging Black people as more culpable for crimes, and endorsing harsher punishment. (Graham & Lowery, 2004) And science faculty (male and female) show implicit biases in evaluating job applicants as more competent and hireable when a résumé carries a male name. (Moss-Racusin et al., 2012)

Furthermore, automatic processes are triggered and shaped by often-irrelevant external factors (the *situational thesis*). For example, social psychological studies show behavioural effects by a wide range of situational manipulations such as priming (e.g. thinking about old age) and banal environmental features (e.g. finding a penny, or sunshine). (Doris, 1998)

Conceptualising moral reasoning, the automaticity literature emphasises its frailty in thesis concerning truth-tracking, resources, frequency, causality, and instead describe its post hoc and social role.

To start, moral reasoning often does not track moral truth well (the *truth-tracking thesis*). This thesis is actually a combination of many other theses. As described above, the speed of intuitive processing limits the time for reasoning, emotional processes such as implicit biases and unconscious stereotypes can deflect reasoning, and irrelevant situational factors can influence reasoning. Truth-tracking

can also be frustrated by unfair cognitive processing due to distortions, errors, and emotions (the *partisan cognition thesis*). Cognitive distortions can have an impact; relatedness motives cause people to form judgements that are in harmony with other people, rather than mainly accurate. (Chen & Chaiken, 1999) And coherence motives cause people to want to maintain important beliefs avoid cognitive dissonance, for example through exhibiting confirmation biases; uncritically accepting evidence in support of one's prior belief, while over-critically scrutinising opposing evidence. (Chaiken et al., 1996) Cognitive errors can also impact reasoning. For example, people cannot reason well about earlier judgements, people do not reliably detect correlations, are more influenced by vivid, concrete data than by pallid abstract data, engage in wishful thinking, and conscious reasoning can actually reduce the quality of judgements compared to intuitive judgements. (Bishop & Trout, 2004; Horton, 2004; Johansson et al., 2005; Wilson & Schooler, 1991) And besides cognitive distortions and errors, emotions can also make reasoning partisan (extension of the affect thesis, including unconscious stereotypes and implicit biases). For example, neurological research on reasoning by Greene et al. shows that moral reasoning in a moral situation can be thwarted by emotions that block the consideration of many relevant factors. (Greene et al., 2001) A further limitation to reasoning is that people typically lack introspective insight; they do not have accurate conscious access into what factors actually shape and trigger their behaviours (e.g. sunshine), and even fail to appreciate it when confronted with the factors (the *introspection thesis*). (Schwarz & Clore, 1983; Sie, 2009; Wegner, 2002) In conclusion, as Haidt sharply draws from such data, moral reasoning "is more like a lawyer defending a client than a judge or scientist seeking truth." (Haidt, 2001, p. 820) Ironically, however, ignoring all of the frailties above, people still typically tend to be overconfident about their reasoning capacities, which actually adds to their incapacity to receive and learn from feedback. (Bishop & Trout, 2004)

Another crucial frailty of reasoning is that it is an effortful and demanding process that relies on cognitive resources that are scarce (the *scarce resource thesis*). (Chaiken, 1987; Galinsky & Moskowitz, 2000; Monteith & Voils, 1998) And due to scarce resources, among other factors, the usage of reasoning is limited, so that people exhibit conscious reasoning only very infrequently (the *infrequency thesis*). (Haidt, 2003; Kühn et al., 2014; Perkins et al., 1991)

Due to its frailties, reasoning is causally quite ineffective, not much causally determining one's judgements, decisions, and actions (the *causality thesis*). For example, intentional effort to suppress one's implicit biases is found to 'rebound', and awareness of the lack of supporting reasons for one's initial judgement does not result in revising it, but merely in 'moral dumbfounding'. (Haidt et al., 2000; Huebner, 2009) As such, reasoning is only a modest predictor of moral behaviour, while emotion has a much stronger relation. (Damon & Colby, 1992; Hardy & Carlo, 2011)

Instead of the causal behavioural role, the automaticity literature holds that reasoning is mostly exhibited *post hoc*, after having formed a judgement or performed an action, to explain and justify one's behaviour rather than causally determining it, and thus as a confabulatory process (the *post hoc confabulation thesis*). (Haidt, 2001; Nisbett & Wilson, 1977) However, according to some this is actually the main role of reasoning, not causally determining behaviour or searching for truth, but operating in the social sphere to explain and justify oneself, defend against criticism, evaluate another's arguments and information, and influence others' behaviour (the *communication thesis*). (Musschenga, 2011; Sperber & Mercier, 2012)

The intuitive system	The reasoning system
Fast and effortless	Slow and effortful
Process is unintentional and runs automatically	Process is intentional and controllable
Process is inaccessible; only results enter awareness	Process is consciously accessible and viewable
Does not demand attentional resources	Demands attentional resources, which are limited
Parallel distributed processing	Serial processing
Common to all mammals	Unique to humans over age 2 and perhaps some language-trained apes
Context dependent	Context independent
Platform dependent (depends on the brain and body that houses it)	Platform independent (the process can be transported to any rule following organism or machine)

(Table 2: *General features of the two systems*. From (Haidt, 2001, p. 818).)

To summarise, moral behaviour is driven by automatic processes, which are emotional, unconscious, quick, and directed by irrelevant environmental factors and harmful cultural biases. Moral reasoning is deficient, as people lack introspective insight, truth-tracking is thwarted by cognitive distortions and errors, reasoning is causally ineffective, requires scarce resources, is exhibited infrequently, and mostly a *post hoc* confabulation.

1D: The automaticity challenge to moral agency

This brings us to formulating the *automaticity challenge to moral agency*. The automaticity challenge is advanced under various names (also the *frail control hypothesis*) and by various authors who all share a core commitment to the prominence of automaticity. Many of the authors have self-proclaimed that their models are built on the work of David Hume, and as such automaticity has led to the rise of several new or neo models of moral behaviour (therefore, I will generally refer to defenders of the automaticity challenge as sentimentalists, in opposition to rationalists, although not all neatly fit in these labels).

To start, social psychologist Jonathan Haidt advances the 'social intuitionist model' (SIM), as plausibly the most discussed challenge and alternative to rationalist models. (Haidt, 2001; Haidt & Bjorklund, 2008) The SIM stresses that moral behaviour is mainly driven by intuitions, and reasoning mostly plays out socially, in discussion with others. As a second example, a cognitive psychological model that has gained traction in decision theory is the theory of 'bounded rationality'. (Baron, 1993, 1995; Blasi, 2009; Cosmides & Tooby, 2004; Gigerenzer, 2008; Sunstein, 2008) Given the strenuous circumstances of decision-making (limited cognitive capacities and external and internal pressures), conscious rational decision is impossible or too costly (in time and/or cognitive resources), such that evolutionarily developed 'heuristics' are more economical; quick, unconscious response processes to minimal information. Thirdly, in what could be called the philosophy of social psychology or cognition, John Doris makes a case for situationism, a theory that especially emphasises the influence of situational factors in determining people's behaviours through triggering one intuition rather than another. (Doris, 1998, 2002) Finally, in philosophy of action, Shaun Nichols and Jesse Prinz' both champion neo-sentimentalist theories, arguing that uncontrolled, emotional states constitute moral judgements, not critical, rational reasoning, as rationalists claim. (Nichols, 2004; Prinz, 2006a, 2007) While all versions of what automaticity is and how it challenges moral agency vary somewhat, all of the theories above share a clear core, which can be formalised as follows:

The Automaticity Challenge to Moral Agency

Empirical premise i	*The primacy of automaticity*: Unconscious, emotional processes are causally determinative of most of people's moral behaviour (judgement-formation, decision-making, and action-guidance).
Empirical premise ii	T*he frailty of reasoning*: Conscious moral reasoning is often rationally deficient and not directly causally determinative of moral behaviour.
Normative premise	*The deliberative standard of moral agency*: Moral agency is marked by deliberative control, the process of conscious, rational reasoning fairly directly causally determining most of one's moral behaviour.
Conclusion	*The lack of moral agency*: People typically lack moral agency over their moral behaviour.

The lack of agency conclusion is detrimental to the human moral project, for example because a moral psychological concept of moral agency itself, in turn, is a foundational premise of (most) theories of moral responsibility. As such, were our concept of agency to turn out to be empirically unattainable, it would cripple the justificatory grounds of the practice of holding people morally responsible for their moral behaviours.[7]

Now, as any attentive reader will instantly remark, the lack of agency conclusion follows not merely from the two empirical premises discussed above, but crucially relies on a further normative premise about what moral agency is, which has thus far only been casually mentioned. It is for this exact reason that I intentionally draw out the premises so explicitly, because while both the empirical and the normative premises have crucial functions in the automaticity challenge, only the empirical ones are elaborated, and the normative one is merely mentioned, regardless of the amount of work it is doing (both in the entire challenge as normative premise, as well as in the empirical premises). In section 3 I will analyse the normative premise. First, after having set out the automaticity challenge now, we will look into the reception of the challenge.

§2: A defence of moral agency

In this section, I will discuss the replies that authors have articulated in response to the automaticity challenge, in defence of moral agency. I will present the types of

[7] In essay 3 of my doctoral thesis I explore theories of moral responsibility in the light of moral automaticity.

replies, point out trends in these, and discuss how this may meagrely ameliorate the challenge.

2A: Replies to the automaticity challenge

There is a vast range of replies to the automaticity challenge, both from the psychological sciences as from various fields of philosophy. While each author contributes in unique ways to the debate, there are two critical features that seem to trace through many of them, defending a notion of deliberative agency originating in the rationalist tradition as mentioned earlier. As such, I will discuss these authors as rationalists (although not all neatly fit the label, as with the sentimentalists who defend the automaticity challenge).

Firstly, the same psychological and philosophical concepts are employed in fairly the same sense. With that, the normative premise is undisputed but just bought into as the standing criterion for agency. Furthermore, the concepts permeate the empirical data and discussion. I will analyse these issues in section 3. The second marker of the defences is that, instead, they go at the challenge head on, disputing the empirical premises. As such, this typically involves employing the same empirical philosophy methodology of drawing on scientific research to support certain claims. Authors bring forth different data, which purportedly refutes the primacy of automaticity (*EP1*) and/or the frailty of reasoning (*EP2*), culminating in a persistence of the possibility of deliberative agency (*NP*).

Furthermore, these claims are brought to bear as substantiating imposing a normative requirement on agents to carefully endeavour to determine their behaviour through conscious reasoning; reiterating the *NP* even more forcefully.[8] To illustrate the rationalist position, the following overview of important contributions to the defence of agency is meant not as an exhaustive representation and discussion of each author's work, but to highlight the shared usage of concepts and focus of claims that is at the core of each author's view.

[8] In essay 2 of my doctoral thesis I explore an alternative response strategy, which instead draws on a different normative premise and focuses on the development of automaticity. Some of the authors discussed in the present essay seemingly belong to this class. However, as the current analysis aims to show, these authors ultimately rely on rationalist notions of deliberative agency. For that reason, in the mentioned essay, I classify these strategies as 'moderative agency', not the required 'developmental agency'.

Starting with some examples from psychology, Augusto Blasi has been a prominent critic of the automaticity challenge in general, and Haidt's SIM in particular. Blasi scrutinises the data the automaticity challenge relies on, arguing that reasoning frequency in the before-mentioned dumbfounding studies, for example, varies widely between types of moral judgements (e.g. sexual taboos are judged intuitively, while fairness less so). (Haidt & Bjorklund, 2008) Secondly, Blasi makes the positive claim that, "the evidence concerning the use of moral reasons and reasoning is far broader and stronger," which he substantiates by pointing to three other critical authors, the 'informal experience of each of us', and the 'perhaps thousands of studies' that sprung from the cognitive-developmental paradigm. (Blasi, 2009, p. 414) With this, while he acknowledges the existence of automaticity, he upholds the possibility of moral reasoning to control it, which is a crucial part in his influential model of moral identity. Albeit with a somewhat different concept of reasoning (a 'recursive, iterative', rather than a 'linear, top-down' process), Blasi clearly attacks *EP1* and *EP2*, writing that, "the real question (…) is whether moral reasons have a determining influence in the final judgment adopted by the person." (Blasi, 2009, p. 418) He answers this question positively, arguing that agents have to assume 'ownership' and control over their judgement.

> *"This involves, as a first step, creating between oneself and the judgment some space in which to operate, distancing oneself from the judgment and relation to it as an object of consideration and reflection (...) to exercise control over it (...) through a reasoned analysis of its origin, of the elements, including emotions and intuitions, by which it was constructed, and of the personal motives that might distort its meaning and corrupt its validity. The quickly formulated judgment – as also the intuitions and the emotions, or the stereotyped automatic associations that led to it – may already be there; but one has the power to go back to it, and accept it or reject it according to one's criteria of validity and truth. One could even exercise some control over one's spontaneous emotions and intuitions, perhaps not to the extent of eliminating them or preventing them from occurring, but in the sense of evaluating them, concluding that they are undesirable, and wanting not to have them." (Blasi, 2009, p. 423)*

While Blasi recognises that automatic processes occur, he stresses that, after being triggered, they can nevertheless be controlled through conscious reasoning. This

control is the distinct human psychological capacity that separates judgement from action, such that, ultimately, moral reasoning can still drive action.

A quite similar approach to automaticity is found in the extensive work of psychologist Robin Hogarth. Hogarth propagates using the principles of the 'scientific method' to make out whether some behaviour is properly substantiated. The first way to do this is assessing the validity of one's intuitions; how it originated, whether it is successful, and how important it is. Secondly, people have "to learn to apply scientific reasoning intuitively." (Hogarth, 2001, p. 226) This involves critical observation of the facts, background assumptions, and patterns, speculation about explanations, testing the intuition and exploring alternatives, and exploring generalisability of an intuition. Hogarth acknowledges that automatic processes operate quick, but argues that one can impose so-called 'circuit breakers'; markers that force one to interrupt the process from running to completion to allow for reflection. (Hogarth, 2001, p. 240)

Another noteworthy contribution is by Darcia Narvaez, who brilliantly explores various developmental factors, but nevertheless maintains that in decision-making, intuitive judgements and conscious reasoning *generally* come together, when "a person monitors and interprets many signals, such as emotional reactions, (…) current goals and preferences, mood and energy, environmental affordances, situational press, contextual cue quality, social influence, empathic response, logical coherence with self-image and with prior history." (Narvaez, 2011, p. 38) Built on Blasi's work, the developmental psychological model advanced by Atkins, Hart, and Donnely holds that, "genuinely moral action cannot derive from wholly automatic affective processes and must include to some degree reflective consideration of information and lines of action." (Atkins et al., 2004, p. 66) In the same mind, cognitive scientists Dreyfus and Dreyfus stress the need for different forms of deliberation, for example 'involved reasoning' to determine which intuition is most appropriate. (Dreyfus & Dreyfus, 1991) And similar focus can be found in the work of various other researchers in developmental, personality, and behavioural psychology. (Colby & Damon, 1992; Moshman, 2005)

In philosophy, Bert Musschenga has written a detailed overview of how positions concerning moral reasoning by various psychologists and philosophers relate to one another, in order to assess the standing of different claims and interpretations.

Acknowledging that intuitive judgements can be bad, and that moral reasoning can be flawed, Musschenga makes a case for the possibility of making reasoning better through training. A crucial insight he advocates is that beyond the impossibility of neutralising emotions, it is undesirable, as they have an important epistemic function. Rather, people have to learn "to see when emotions direct our attention to morally relevant features of a case, and when they hinder us pass a balanced judgement." (Musschenga, 2011, p. 80) To make reasoning more truth-tracking, for example, Arne Næss' 'pro aut contra' survey strategy can be utilised, whereby arguments are carefully weighed against one another, avoiding confirmation bias. (Naess, 1966) Framing effects can be avoided by awareness of one's lack of knowledge and expertise, or even minimised or eliminated by counter-framing. (Druckman, 2004) And with the knowledge of how and how much one is biased, people can appropriately alter their responses. (Wilson & Brekke, 1994) Among the main claims of Musschenga is that moral reasoning does not merely occur in the case of absent, weak, or conflicting intuitions, but also, for example, when intuitions conflict with one's moral values. Moreover, the automaticity literature can enhance reasoning, since, "if we have already so much insight into the biases and errors of human judgements, we should be able to design strategies for improving human reasoning." (Musschenga, 2008, p. 139)

Jules Holroyd and Daniel Kelly advocate three ways in which agents can have control over their behaviour. Firstly, an agent can intervene in the operant automaticity. Secondly, before being in a moral situation, one can prepare one's own cognition to trigger certain intuitive responses, or, thirdly, select an environment that will trigger certain intuitive responses. (Holroyd & Kelly, 2016) The cognitive preparation strategy is mainly built on research on 'implementation intentions', plans that one can form for oneself as to how to respond when encountering certain anticipated stimuli. (Gollwitzer, 1999) Hanno Sauer similarly invokes Gollwitzer's research, among some others, to argue for the possibility of '*ex ante* education', influencing one's intuitions before they operate. (Pizarro & Bloom, 2003) In addition, Sauer proposes '*ex post* education', "the ability to reflectively monitor one's cognitive operations and alter them according to standards of rationality or reliability deemed appropriate by the reflecting subject." (Sauer, 2012, p. 268) For example, when informed that their judgements may have been affected by the sunny weather, people can counter-balance that influence. (Wilson & Brekke, 1994) This approach draws on

the work of Jeannette Kennett and Cordelia Fine, who discuss similar concepts, 'up-front control' and 'after-the-fact correction', and hold that, a "loser examination of the interaction between automatic and controlled reflective processes in moral judgment (…) makes room for the view (…) that genuine moral judgments are those that are regulated or endorsed by reflection." (Kennett & Fine, 2009, p. 78) What is crucial for the regulation of one's automatic responses is whether a person is motivated. A final remarkable philosopher that has to be mentioned here is Nancy Snow, who refines Bill Pollard's account of habitual actions and rationality. Along with Pollard, Snow maintains that moral responsibility for automatic behaviour hangs on the ability to control it, or, in other words, for the behaviour to be part of one's agency, "something the agent does, rather than something that merely happens to him." (Pollard, 2003, p. 415) Not control in the sense of initiating the automatic behaviour, but rather to intervene. "Habitual actions are responsible in the sense that they are under the agent's intervention control. I can intervene to stop or redirect the action sequence." (Snow, 2006, p. 552) Moreover, for intervention control a "conscious effort or wilfulness is needed to interrupt or redirect" some behaviour. (Snow, 2006, p. 550)

2B: Disputing the automaticity challenge

Both the psychological and philosophical rationalist responses to the automaticity challenge express a strong faith in deliberative agency. While authors propose their own terminology, the content of their proposals are essentially similar, and often even rely on the same empirical sources. Unified, the rationalist position counters the automaticity challenge by advancing the following two alternative empirical premises.

Instead of *EP1*, rationalists acknowledge that automaticity exists, but claim that once automatic judgements are elicited, agents have the ability to, what we can call *moderate* them; inhibiting their continuation (blocking automaticity from being constitutive), reflecting on them (e.g. introspectively accessing them and evaluating their situational and biased origins, their correctness, and desirability), and controlling whether they influence behaviour.

Instead of *EP2*, rationalists acknowledge that moral reasoning can be flawed, but claim that with agents who are skilled in reasoning and motivated to behave morally are nonetheless able, through hard work, to consciously reason rationally, tracking moral truth (avoiding or countering cognitive distortions and errors), and directly causally determine their moral behaviour.

As such, an alternative conclusion is arrived at, in which the *NP* is met.

The Rationalist Defence Against the Automaticity Challenge to Moral Agency

Empirical premise i'	*The moderation of automaticity*: Automaticity exists, but conscious deliberation can moderate its operation and causal influence of behaviour.
Empirical premise ii'	T*he resilience of reasoning*: Moral reasoning can be flawed, but skilled, motivated, effortful, conscious, moral reasoning can be moral truth-tracking and directly causally determinative of moral behaviour.
Normative premise	*The deliberative standard of moral agency*: Moral agency is marked by deliberative control, the process of conscious, rational reasoning fairly directly causally determining most of one's moral behaviour.
Conclusion'	*The perseverance of moral agency*: People can exhibit moral agency over their moral behaviour.

2C: Revising the automaticity challenge

Now, there are two main reasons for which the rationalist empirical reply may not hold up as successfully as this.[9] Firstly, the empirical premises are simply unsettled. Just as rationalists argue that the automaticity evidence is not as solid as alleged, and that there is stronger evidence for moral reasoning than alleged, these claims themselves are not as empirically solid either. It is simply a tragic feature about the current state of the debate that many of the empirical claims currently lack conclusive evidence concerning their precise specification. As Blasi has pointed out (while subsequently nevertheless drawing on empirical premises favouring rationalism), the empirical premises are still unclear, because many of the details are understudied and/or too difficult to test due to their complex nature and vague description (e.g. distinguishing 'genuine reasons' from 'confabulations'). "We have no idea of how frequently intuitions are missing, or are too weak, or contradict each other; we don't know how frequently our moral judgments have real consequences of one kind or another for ourselves or others; we don't know how frequently we engage in moral

[9] I am here not as much concerned with criticising these arguments as I am with deconstructing their conceptual foundation. In §3 of essay 2 of my doctoral thesis I more elaborately scrutinise the validity of these arguments.

reasoning in the context of a conversation, or whether we have internalized the dialogical context of moral reasoning, and rely on it systematically." (Blasi, 2009, p. 416)

Secondly, while the rationalist defence aims to counter many of the theses of the automaticity challenge (e.g. constitutive, introspection, truth-tracking, and causality), the scarce resource and infrequency theses seem to persist. As Kennett and Fine explain, the reason why they consider moderative processes as 'controlled' is that these processes depend on available resources for controlled cognitive processing. (Kennett & Fine, 2009, p. 92) There is a range of research on this issue, evidencing that people's ability to control automatic processes depends on the availability of what is called 'controlled processing resources', 'working memory capacity', or 'attentional resources', for example to suppress ethnic stereotypes or inhibit primed traits. (Barrett et al., 2004; Govorun & Payne, 2006; Monteith et al., 1998; Payne, 2005; Thompson et al., 1994) Besides social psychology, also research in cognitive neuroscience shows that when someone attempts to inhibit emotional responses, there is clear activation in the dorso-medial prefrontal cortex, a brain region linked to voluntary self-control processes, particularly conscious reasoning. (Blair et al., 2001; Greenwald & Banaji, 1995; Kawakami et al., 2000; Kühn et al., 2014) Since the cognitive resources that moderative reasoning relies on are scarce, agents can only exhibit such processing infrequently (see §1). This constitutes a crucial weakness for rationalism, because deliberative agency requires a quite solid notion of control, and while the empirical premises are yet inconclusive, it does not seem to look like a very robust notion of deliberative agency will be defensible claim.

With that, we can better understand what the empirical debate is essentially about; determining how much deliberative agency people can exhibit, i.e. frequency. Admittedly, the rationalist contribution to this debate is very valuable in many respects, such as highlighting the shortcomings of empirical research, elaborating on the interaction of automaticity and reasoning, and appreciating a moderative ability of moral reasoning. But what conclusions about frequency can we draw from this? Or, more precisely, where do the automaticity challenge and an empirically plausible frequency of deliberative agency meet?

From the rationalist side, Kennett and Fine, for example, argue that it is not required that *every* moral response be the product of explicit, effortful deliberation.

(Kennett & Fine, 2009, p. 88) However, claims concerning reasoning *never* or *always* determining behaviour, and with that *entirely having* or *entirely lacking* agency, are a type of categorical claims that do not further the debate. Proponents on the automaticity side have taken responses into account, conceding that, in special circumstances, agents are able to control automaticity with rational reasoning, and that there are more such opportunities that the original automaticity challenges acknowledged. For example, Haidt and Bjorklund have admitted that people from certain highly specialised subcultures with trained skills ins 'unnatural modes of thought' (e.g. philosophers, surely!), in social interactions, or when initial intuitions are weak, conflicting, or contradictory, people can and do reflect. (Haidt & Bjorklund, 2008) As such, the automaticity challenge can be cushioned somewhat, with which we can formulate an ameliorated automaticity challenge, if you will.

Conclusion'' *The alleviated lack of moral agency*: People can occasionally exhibit moral agency over their moral behaviour.

Ameliorated automaticity holds that, rather than agents '*typically* lacking' agency, or it being *rare*, it can be exhibited *occasionally*, depending on factors such as resources and skill. This conclusion may be an alleviation of the original rarity of agency. But what do we actually conclude about agency with that? Is this a 'successful' defence of agency? Here it becomes clear why it is so important to distinguish the *EP*s and *NP* that are interwoven in the automaticity challenge.

Settling empirical questions about the frequency of certain cognitive processes only gets us so far in settling the automaticity challenge. The debate about the *EP*s may inform claims about frequency, answering *how much* deliberative agency people can and do exhibit, but this only partly answers the entire question about the status of moral agency, because it nevertheless leaves open how much is *enough*, or *what agency is*. This is what the *NP* is about. Now, the *NP* at play seems to invoke a very rigid notion of deliberative agency, involving a reasoning process that is conscious, rational, directly causal, and a high frequency, controlling a significant degree of one's behaviour. Given this NP, it is implausible that the debate about the *EP*s, even with an ameliorated automaticity challenge, will result in a final conclusion (if there is such a thing) in which moral agency is successfully defended. At best (for those who champion moral agency), the automaticity debate is at an impasse, in which authors

quarrel about the *EP*s being a tad harder of softer, all within the framework of the deliberative agency *NP* that at the same time restricts what type of phenomena are under discussion due to its rigidity, and keeps at bay settling any claims of agency due to its vagueness. To illuminate and substantiate this diagnosis, let us now analyse the *NP*.

§3: Conceptual analysis of deliberative agency

In this section I analyse the normative premise. I start by drawing out the conditions of deliberative agency that are (often implicitly) present in the automaticity debate. Subsequently I briefly touch on some background theory for this type of agency concept. Finally, I explore how the *NP* is involved in the empirical premises.

With this, I argue that the *NP* causes the automaticity debate to be in a state of impasse, and that a way out of this situation requires revisiting the *NP*, for which there are various good reasons, and which will allow for the insights by both the sentimentalist and rationalist camps to be properly appreciated.

3A: Deliberative agency as normative premise

The normative premise has the function to establish what moral agency is (a *descriptive definition*), and when some behaviour counts as being agentive and an agent as having agency (a *prescriptive standard*). The *deliberative agency* concept, present in the work of both sides discussed above, is one formulation of a normative premise, with its roots in the rationalist tradition, but in the current debate in a very strict, say, intellectualist, form.

Exemplary is Haidt envisioning the rationalist model of moral judgement to entail conscious deliberation in which "one briefly becomes a judge, weighing issues of harm, rights, justice, and fairness, before passing judgment (…). If no condemning evidence is found, no condemnation is issued." (Haidt, 2001, p. 814)

Deliberative agency draws on some idea of reasoning as the fundament of agency. For a background to such a 'reasons theory of action' we can think of Christine Korsgaard and David Velleman. (Korsgaard, 2008; Velleman, 2000) On a very rough description of their work, they distinguish 'actions', which are *done by* an agent, from other phenomena such as 'happenings', which, instead, rather *happen to* an agent. Actions are distinct because they involve the agent's capacity for reasoning in some way. Involving reasoning makes that the guiding of the behaviour can be responsive to relevant moral considerations, such that the behaviour can be ascribed to the agent as *their* behaviour. With that, action theory provides a notion of agency, which, in turn, is a necessary notion for theories of moral responsibility concerning whether a person is morally responsible *as agent* for some behaviour. While there is much interesting discussion about how to draw out the details of the theories about agency and reasoning such as those by Korsgaard and Velleman, I will not go into that further here, since their work is not directly involved in the automaticity debate, but merely serves here as an example of what a concept of agency has to provide; a concept with which we are able to ascribe some behaviour to an agent as agent.

Rather, I will analyse the concept of agency that is at play in the automaticity debate as a theory on its own, teasing out the further specifications that the deliberative agency notion adds beyond the very basic idea that an agent's reasoning is involved in the behaviour in some way. These specifics designate more precisely what type of reasoning process has to be involved, and what role the process has to play. As such, we can think of these specifics as *conditions of agency* that have to be met in order for some behaviour to count as being guided by reason, and thus as agentive.

3B: The causality condition

> *The causality condition*: For some moral behaviour to be agentive, the agent's moral reasoning has to guide the behaviour such that the reasoning concerns the relevant moral reasons (or other cognitive processes) that function as the actual causal determinants of the behaviour.

The function of the causality condition is to specify the role of reasoning, by prescribing a particular way in which reasoning has to be involved in one's behaviour. The aim of this condition is that moral reasons are not merely present as justificatory 'normative reasons', but also as causal 'explanatory reasons'. This condition can be seen as being in tension with the post hoc confabulation thesis. I take this condition to be fairly uncontroversial given an understanding of agency that involves reasoning.

3C: The control condition

> *The control condition*: For some moral behaviour to be agentive, the agent's moral reasoning has to guide the behaviour such that the agent has a sense of control over the actual causal cognitive processes.

The function of the control condition is to specify a particular type of reasoning process that is required. While many of the authors above invoke some notion of control, they propose different definitions of what control is, ranging from reflecting on and evaluating one's processing, to endorsement of active reasons, intervening on certain processes, or deliberatively selecting or choosing reasons. Nevertheless, these different definitions share the same central idea that the agent is required to be involved in the reasoning process as, say, the 'reasoner', such that the reasoning is the agent's own. With that, the control condition establishes whether some behaviour is agentive.

That control has the function of establishing agency is important to note here, because this is a different matter from behaviour being rational or, say, 'morally good' (as being responsive to the relevant moral factors, not as being a deliberative process). One does not imply the other. For example, a large part (if not the majority) of people's automatic processes may result in behaviour that is rational and morally good, but as the agent is not involved as reasoner, the behaviour is nevertheless not agentive. This distinction is important because many rationalists focus on arguing that some automatic behaviour can be rational (responsive to reasons). While this may be true, this does not meet the automaticity challenge, since the challenge is not as much a challenge to the irrationality or immorality of people's behaviour, but rather to the

typical lack of agency over the behaviour, whether the reasons the agent was involved in the reasons-responsiveness (which, surely, is more problematic with morally wrong behaviour, because then people lack the ability to control the wrong behaviour, but it is nonetheless a feature of automatic morally right behaviour as well). As such, the automaticity challenge as a challenge to agency still stands in the face of rationalist replies. For example, Snow takes up a variant of 'internalist' theories of reasons, which hold that something is a reason due to being the agent's psychological state, rather than due to its content. From there, she argues that that no consciousness of one's reasons is required for them to be rational, such that goal-dependent automatic behaviour can be rational and purposive. Nevertheless, beyond that automatic processes can be rational, to ground moral agency (in a sense that can substantiate moral responsibility), she nevertheless resorts to invoking a very direct notion of control, 'intervention control'. (Snow, 2006, p. 552)

Furthermore, another important observation is that the control condition can be seen as tying control to reasoning in contrast to understanding automatic processes as uncontrolled. This conceptual pairing is actually already present in the foundation of automaticity research and modelling. Two-system theories, which underlie most research and modelling, define automatic processes as uncontrolled and deliberative processes as controlled. This conceptual pairing dichotomises automatic and deliberative processes along with a process being controlled or not, such that a process can be *either* automatic, *or* controlled, but not both. With that, the very way in which automaticity is conceptualised already on theoretical grounds alone rules out control, and, with that, agency. In other words, even before any empirical findings on automaticity come into the discussion, agency is excluded. The result of this is that the theoretical space for agency is limited to deliberative processes only. Subsequently, this shapes the debate about the automaticity challenge, because with agency only to be looked for in deliberative processes, it is intelligible that the debate mainly concerns what the frequency of deliberative processes is. This is a way in which the deliberative agency *NP* does not only function in the evaluation of automaticity research, but already in construing the research, and with that plays a role in actually setting the debate.

3D: The consciousness condition

> *The consciousness condition*: For some moral behaviour to be agentive, the agent's moral reasoning has to guide the behaviour such that the agent is consciously aware of the actual causal cognitive processes.

The consciousness condition further specifies a particular type of reasoning process as a process that involves, beyond control, conscious awareness. Such consciousness can be thought of in different terms, such as explicit deliberation about reasons, or as introspective access into the operant reasons (and other cognitive processes). With that, the consciousness condition adds to some behaviour being the agent's behaviour, as the connection between the agent and the behaviour is marked by another feature. As with the control condition, the consciousness condition is already present as a conceptual definition on the theoretical level, not merely in the *NP* to evaluate findings on automaticity. For example, in two-systems theories consciousness is defined as a feature of deliberative processes, and thus stands in contrast to the definition of automatic processes as unconscious. Furthermore, just as the control condition does, the consciousness condition limits what type of processes can count as agentive, such that processes that operate below the level of consciousness cannot be conceptualised as constituting agency. As some rationalists have argued, conceiving of certain unconscious process as agentive would much expand the amount of agentive processes that agents exhibit, allowing recognition of a higher frequency of agency, whereby the frail agency conclusion would not be justified. (Horgan & Timmons, 2007; Sauer, 2012; Sneddon, 2007) As such, the consciousness condition not only shapes the automaticity debate, but also has to be invoked as an agency condition to substantiate the automaticity challenge to agency.

3E: The directness condition

> *The directness condition*: For some moral behaviour to be agentive, the agent's moral reasoning has to guide the behaviour such that the agent's

consciousness of and control over the actual causal cognitive processes occurs within a fairly temporally direct frame.

The directness condition is an even further specification of the type of reasoning process. Better yet, without the directness condition in addition, the causality, control, and consciousness conditions are not nearly as restrictive. Directness implies that reasoning processes that are indirect, or *too* indirect, are excluded as constituting agency.

To start with how directness restricts the control and consciousness conditions, as I pointed out earlier, many of the rationalist responses emphasise a moderative role of reasoning. What marks such moderative responses is that they all draw on strategies that function within a temporal restriction, in which the reasoning occurs either during or just prior to exhibiting the behaviour. Even those who discuss processes that avoid the consciousness condition, arguing for example for causal reasoning processes that function unconsciously, such as 'proceduralised reasons', or 'automatic goals', almost exclusively draw on such strategies that nonetheless require conscious, controlled reasoning to be exhibited within a fairly temporally direct period before the subsequent unconscious causal reasoning process (e.g. 'implementation intentions', 'ex ante education', and 'up-front control'). Through observing that only a limited type of processes and strategies is explored, we can deduct that they tacitly invoke directness as a limitation on what behaviours can be understood as agentive. What is important about this is that while such strategies may avoid the consciousness condition and its challenges in the most stringent form, they nevertheless stand in opposition to challenges such as the force of situational influences, various cognitive deficiencies, and, most importantly, the scarce resource thesis. For one, because even unconscious reasoning processes require such resources to warrant rationality, and secondly because even though expanding the time period for exhibiting controlled, conscious reasoning a bit may make available some more resources, it is nonetheless very limited. As we saw at the end of section 2, processes that rely on such resources and are susceptible to such influences are very frail. In result, abiding by the directness standard severely limits the capacity that people have to exhibit even unconscious reasoning processes. This is how directness is a crucial condition employed in the *NP*, further restricting the type of processes that can be agentive, necessary to warrant a conclusion of frail agency.

Directness also restricts the causality condition. Another telling way to extract usage of the directness principle is the way that conclusions are drawn from research on intuitive judgement-formation. In dumbfounding studies, such as Haidt's, reasoning is concluded to not be causally involved in the judgement since people do not revise their judgement. However, more precisely, the conclusion is restricted to a direct role of reasoning, revising the judgement right then and there; nothing about the role of reasoning in revising judgements over a longer period of time is said with this. But apparently direct causal effect is what matters most.

Now, there is more to the directness condition than there initially seems to be, which requires further unpacking, and explains why it is actually unsurprising that directness so permeates the debate, and does this seemingly necessarily so. We can start with the observation that there is little debate on the *developmental processes* of automaticity, that is, how automatic processes and structures are formed. In two-systems theories, system 1 is conceived as being partly shaped by innate specification, provided by evolutionary factors, and partly 'acquired' through implicit learning from social and environmental experiences. (Evans & Stanovich, 2013; Stanovich & West, 2000, p. 659) Only slightly differently, Haidt champions an 'externalisation' model, proposing that various intuitions are evolutionarily built in and that social interaction with one's surrounding culture drives the emergence of certain of those innately prepared intuitions. (Haidt, 2001, pp. 826-828) While there are some important differences between acquisition and externalisation models, what is key to both is that the developmental process is a passive process, driven by evolutionary and sociocultural factors, not an agentive process in which a person is involved as reasoning agent to determine what kind of automatic processes one develops. Almost all discussion on the origin of automatic processes follows this principle, after which various theories of sociocultural learning can be added (for example, emotional training by parents, or personality forming through social embodiment. (Lapsley & Narvaez, 2004, p. 206; Prinz, 2007, pp. 268-270) As such, the development of automaticity is of little relevance to the debate concerning agency.

From this observation we can deduct that the automaticity challenge is not as much a dispute about a very broad question concerning all the processes that are involved in the entirety of cognitive processes, say, the entire chain of cognition, which ultimately culminates in behaviour. Rather, the debate merely concerns a specific part of that chain, the final part, involving only the processes that operate

while forming a judgement, making a decision, or guiding an action. In other words, the automaticity challenge is a debate about what we can call the *operational processes*. As there is no agency in development, agency is a matter of behaviour causation. Discussion on the development of automaticity is merely a premise that has a supporting role in the automaticity debate, a backstory to debate about the processes that drive the operation of behaviour, and it is here that agency has to be located. Appreciating a developmental premise, a full formulation of the automaticity challenge would read that, *once learned*, automatic processes are operationally superior to reasoning processes.

With the debate being construed as a debate about operational superiority, excluding development, the concept of agency itself is, in turn, also restricted to operational processes. As such, the construal of the debate has directness to behaviour already built into it, such that only processes that concern the final part of the entire chain of cognition, processes that are fairly directly connected to action, matter to agency. Hence, when rationalists explore reasoning processes, they are just complying well with the terms of the debate when they only explore processes within a restricted sphere of behavioural directness. And therefore, while some may speak somewhat of 'development, 'habituation', or 'automatisation', this is all necessarily within the sphere of operational processes with a quite direct behavioural effect, since that is where agency is debated, and thus that is where a role for reasoning has to be debated.

In conclusion, the construal of the automaticity challenge as an operational challenge turns on a concept of agency that implies directness. As such, we can see that directness is not only a condition in the *NP* but already a feature of the paradigm that automaticity literature is structured by. Moreover, we can see that directness is a necessary condition of the automaticity challenge, as it severely restricts the types of processes that can be explored as agentive, and thus restricts the room there is for agency such that rationalist strategies for agentive reasoning are limited to ones that have a fairly direct connection to behaviour causation.

3F: The frequency condition

> *The frequency condition*: For an agent to have moral agency overall, their moral behaviour has to be, in general, causally determined by their direct, conscious, and controlled moral reasoning.

A final condition of the deliberative agency *NP* is the frequency condition, which concerns the amount of one's behaviour that has to be agentive to speak of a person being an agent and having agency over their behaviour overall. This condition does not concern categorical standards of 'all' of one's behaviour is 'always' and 'entirely' determined by reasoning. Rather, the condition that *in general* one's behaviour is determined by reasoning, can be broken down to a degree of this all, that *most* of one's behaviour is *usually* and *largely* determined by reasoning. This condition stands in tension with the ameliorated automaticity conclusion that this is only occasionally so.

3G: Conclusion

In combination, this set of conditions makes that deliberative agency is an overly restrictive and strict standard. Only a very limited range of types of processes are eligible to count as agentive, and a high frequency of such processes is required to be exhibited for there to be agency overall. As such, deliberative agency as the *NP* contributes to the conclusion of the automaticity challenge that people typically lack agency.

What we can conclude from this is that empirical findings on people's moral cognitive processes do not straightforwardly imply the typical lack of agency due to automaticity; this conclusion only follows given the particular concept of deliberative agency that is employed, as a conscious form of reasoning that offers control over causing certain behaviour in a fairly direct manner with a high frequency. Moreover, empirical findings are likely to be in line with the automaticity proponents, since the findings are findings within that conceptual paradigm. In short, the automaticity challenge to moral agency is only a challenge to a certain concept of agency. Given

the conceptual construal of the automaticity challenge, there is very little room for agency, whereby the debate is in an impasse, quarrelling about slightly more or less room for agency within a very small conceptual margin.

It is a wholly different matter, however, whether the way that agency, and with that automaticity, are conceptualised is actually the 'right' way. There may be very different concepts of agency available. While I endorse an *NP* that involves conditions such as rationality, causality, control, consciousness, directness, and frequency, it may be possible to think of these in different terms, and as functioning on different levels, such as in the developmental process. Therefore, I expect that a successful way of dealing with the automaticity challenge will not be a head-on rebuttal of empirical premises, but rather starts from a different concept of agency.

Moral Agency, Automaticity, and Character

Exploring a Tripartite Model

Abstract: *Recent empirical research in the psychological sciences has led to the 'automaticity challenge to moral agency'. This challenge holds that, since people's moral behaviour is often driven by automatic, unconscious, affective processes, rather than conscious, rational deliberation, they typically lack 'moral agency'. I explore a 'tripartite model of moral agency', conceptualising three distinct, complementary modes of exhibiting moral agency. Central is the shift of moral reasoning from an 'operational' role in guiding behaviour to appreciating its role in the self-development of one's own 'moral character'. I argue that moral agency can be successfully defended on a tripartite model and discuss various developmental strategies.*

Introduction

"If you don't got no sauce, then ya lost.
But you can also get lost in the sauce.

You can't get born with sauce; you gotta get seasoned.
I dun 'quired the sauce."

- Gucci Mane[10]

Recent empirical research in the psychological sciences has been taken to challenge the moral psychological notion of *moral agency*, at the core of moral philosophy. Traditionally, moral agency is conceptualised as being grounded in conscious, deliberative processes, such that a person's moral behaviour is guided by one's moral reasoning. However, the data purportedly supports the view that moral behaviour is instead driven by *automaticity*; automatic, unconscious, affective processes. As such, people typically lack moral agency over their moral behaviour as reasoning agent.

In this essay, I aim to meet this challenge through exploring an understanding of agency as exhibited in three different modes rather than as one single, monolithic phenomenon. Besides *deliberative* and *moderative* agency, especially *developmental* agency is central to this model; shifting the role of moral reasoning from *operational*, concerned with guiding moral behaviour, to *developmental*, concerned with the agentive self-development of one's own moral character. As such, the driving question here is whether moral character development can be a normatively useful

[10] Gucci Mane, in a mythical dialogue, as lyrically alluded to in Translee's *Lost in The Sauce* (Translee, 2016) and CJay's *Sauce Drip* (CJay, 2016), advances three theses about personality. For one, he holds that one's personality is the main source from which further projects in life spring. Subsequently, Mane argues that one has to remain conscious and honest about the state of one's own personality (to avert getting 'lost in the sauce'), and effortfully cultivate one's own virtue through experience and study ('getting seasoned'), since this cannot be authentically obtained in another way (e.g. innately or socially). Initially, however, he maintains the inverse, that one cannot achieve anything without virtue ('lost without sauce').

Mane's view can be marvellously aligned with several key aspects of the moral character account advanced here. Since automaticity drives most of one's moral behaviour, a 'seasoned' character is a necessary requirement for agency; one is 'lost without sauce', as one cannot constantly exhibit conscious deliberation. Without such insight, one's agency is lost in automaticity; 'lost in the sauce' of, for example, overconfidence in one's rational deliberative control over operant implicit biases. Therefore, critical self-awareness of one's deep cognitive structures, and effortful self-development of one's character (in addition to environmental opportunity) are central to morality. Moreover, Mane also points at an ancient intriguing puzzle concerning the origin of agency (having to acquire sauce yourself, but being lost without already having it), which the developmental approach of this essay hopes to resolve as well, through describing the gradual coming into being of agency and self-development.

and empirically viable concept of moral agency. To answer this question, the project is both theoretical and empirical in its methodology.

Finally, a proviso concerning the essay's approach and aims. I 'positively' propose a novel account as alternative to the existing discourse. By framing it as such, I aim to explore the strongest version of the concepts that are under discussion, in order to make their central points as clear as possible. To some readers, however, some conceptual distinctions (e.g. between 'moderative agency' and 'developmental agency', or 'indirect control' and 'developmental control') may not be sufficient to understand the account explored here as an entirely distinct, alternative account. If that is the case, I bid that reader does not get hung up on the ultimate framing or identification as independent, for, even as 'mere' extension of some already-existing account, the objective of embracing the points of the conceptual distinctions would be achieved, under whatever title that may be. The main points here being that, given automaticity, the moral character development should be appreciated as an important aspect of agency, and with that the opportunities an agent has for developing one's character, in terms of one's rational capacities as well as environmental circumstances.

After a brief introduction of automaticity and the automaticity challenge to moral agency (§1), I start out with a positive account of the *tripartite model of moral agency*, giving a general description of the approach, substantiating it with an exemplary scenario, drawing the background of rational agency, and setting out the three agentive modes and concept of *moral character* (§2). Subsequently, I elaborate further on *moderative agency*, sub-categorising its instances in the literature, and critically analysing it (§3). Finally, I extensively elaborate on *developmental agency*, discussing how this is typically glossed over in the literature, why it may be considered an agentive mode, its advantages, and different strategies for exhibiting such agency including empirical substantiation of these (§4). I conclude that the tripartite model of moral agency can successfully defend a conception of moral agency that is empirically realistic through acknowledging the gravity of the automaticity literature, theoretically functional through illuminating the discourse in novel ways, and normatively useful through providing a ground for evaluative standards in moral practices.

§1: The automaticity challenge to moral agency

Here follows a quite broad albeit eminently rough introduction to *automaticity*. Over the last few decades, the automaticity of human cognition has become one of the most intensely researched phenomena at the intersection of fields such as behavioural, developmental, social, and cognitive neuropsychology. Typically, this research is based in two-system theories of cognition (also often named dual-process theories), which dichotomously defines 'system 1' processes, which are automatic, affective, and non-conscious (say, intuitions), in opposition to 'system 2' processes, which are conscious, deliberative, and controlled (say, reasoning).[11] (Bargh & Chartrand, 1999; Evans & Stanovich, 2013; Sloman, 1996; Stanovich & West, 2000) Much of this research is taken to evidence that automaticity is so ubiquitous that the majority of human cognition in judgement-formation, decision-making, and action-guidance (hereafter jointly referred to as *behaviour*)[12] is to a large extent driven by automatic processes. As social psychologist Jonathan Bargh and neuropsychologist Tanya Chartrand write, "most of a person's everyday life is determined not by their conscious intentions and deliberate choices, but by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance." (Bargh & Chartrand, 1999, p. 462) Also in philosophy, automaticity is becoming widely discussed, ranging from topics in philosophy of mind, epistemology, and to political philosophy, action theory, and theories of moral responsibility.

Especially in relation to moral behaviour, automaticity constitutes an intriguing, fundamental, and troubling matter. Traditionally, moral psychology and action theory are grounded in what we may generally call a rationalist paradigm. Rationalism holds that moral behaviour is 'agentive', properly belonging to an actor *as agent*, due to one's conscious moral reasoning being the causal determinant, such

[11] I do not straightforwardly accept the two-system theory's dichotomous definition of system 1 and 2 processes, or the initial definition of implicit biases provided here, because I believe that system 1 processes, including implicit biases, can be reasons-responsive and controlled rather than merely associative and non-propositional. I return to this later. For now, I employ the terminology as it is traditionally found in the literature, because it helps to clarify the topic and my position.

[12] While some may specifically discuss one specific process, I take discuss judgements, decisions, and actions as a cluster, because of their intimate connections, such that the former two processes ultimately relate to the latter, and especially the latter is most morally relevant. Moreover, as many of the authors mentioned in this essay acknowledge, as a function of their connectedness, the three processes are all susceptible to similar automaticity challenges.

that rationality and intentionality are involved. (Kohlberg, 1973; Korsgaard, 2008; Piaget, 1932; Velleman, 2000) In turn, also moral responsibility theories are traditionally grounded in such rationalist notions, holding someone morally responsible for some behaviour due to, for example, one's capacity to consider morally relevant factors and make conscious, deliberative choices. (Wolf, 1990) Consequentially, much of an agent's moral behaviour, the part driven by automaticity, may have to be excluded from the agent's moral responsibility, leaving large gaps in our moral practice. (Levy, 2014)

The automaticity literature refutes the rationalist paradigm. As I argued elsewhere, the empirical claims grounding moral automaticity can be seen as two sets of empirical theses concerning moral cognition, and a normative premise involving a standard of agency.

On the one hand, automatic processing, operating through situational factors triggering associated states or processes, are quicker, and its unconscious operation and emotional valence make its operation fairly robust when facing deliberative interference (the *speed*, *unconscious*, *affect*, and *situational* theses). On the other hand, moral reasoning is typically quite frail, since its operation is slow and requires scarce cognitive resources such that it can only be exhibited infrequently, and even when exhibited it often does not rationally track moral truth well due to emotional and cognitive distortions (the *speed*, *resource*, *frequency*, *affect*, *partisan cognition*, and *truth-tracking* theses). As such, reasoning is often not causally determinative of one's moral behaviour, but rather follows afterwards in order to rationalise one's behaviour (the *causality*, *post hoc*, and *confabulation* theses).

This account of moral cognition, then, is evaluated by a normative standard of moral agency in the mind of the rationalist paradigm mentioned above, although sterner, encapsulated in the following set of agency conditions. For an agent to have moral agency, most of one's moral behaviour has to be under deliberative control, meaning that it is causally determined by conscious, rational deliberation in a fairly direct manner (the *frequency*, *control*, *causality*, *consciousness*, and *directness* conditions).

Thus, we have arrived at the *automaticity challenge to moral agency*, holding that; since automatic processes rather than reasoning processes often drive one's moral behaviour, and since reasoning grounds moral agency, people typically lack

moral agency. While various authors in psychology and philosophy, under various names for the concepts, and with some content variance, have advanced the automaticity challenge, they all share most of these core commitments outlined above, and instead endorse alternative moral psychological models much in the sentimentalist tradition, emphasising that automatic emotional processes are superior and sufficient for driving moral behaviour. (Baron, 1995; Doris, 2002; Haidt, 2001; Haidt & Bjorklund, 2008; Nichols, 2004; Prinz, 2007)

The Automaticity Challenge to Moral Agency

Empirical premise i	*The primacy of automaticity*: Unconscious, emotional processes are causally determinative of most of people's moral behaviour (judgement-formation, decision-making, and action-guidance).
Empirical premise ii	T*he frailty of reasoning*: Conscious moral reasoning is often rationally deficient and not directly causally determinative of moral behaviour.
Normative premise	*The deliberative standard of moral agency*: Moral agency is marked by deliberative control, the process of conscious, rational reasoning fairly directly causally determining most of one's moral behaviour.
Conclusion	*The lack of moral agency*: People typically lack moral agency over their moral behaviour.

To somewhat further elucidate moral automaticity, let us look more specifically at one form, implicit biases, which form the focus of this essay due to being the most discussed automatic process in the literature.

Implicit biases are automatic attitudes that typically operate without conscious awareness of the agent and are difficult to control even when one is aware of them. Hence, biases influence someone's behaviour (even including perception, evaluation, and emotional responses) so that this often differs from how it would have been, were it determined instead by the agent's consciously endorsed, explicit attitudes. Implicit biases can concern social groups based on class, race, gender, sexual orientation, mental illness, physical ability, religious identity, aesthetic appearance, but also concern many other features, and negatively influence interactions with individuals who belong to these groups due to connecting them to negative properties or stereotypic traits. Studies show that biases can affect an agent's behaviour in relatively minor ways, like blinking more and choosing another seat, or in more significant ways, like job applications and criminal-sentencing decisions. (Blair et al., 2004; Moss-Racusin et al., 2012) Importantly, implicit biases are no rarity but omnipresent; virtually everyone harbours and expresses certain implicit biases, including those who belong to a particular social group themselves and those who

explicitly and sincerely avow egalitarian values. (Jost et al., 2009) Ultimately, diverse types of effects, by various agents, all factor in together in systematically continuing and reinforcing patterns of discrimination, marginalisation, and oppression.

In response to the automaticity challenge, a range of authors have criticised the paradigm and defended rational moral agency, often in the rationalist mind. The main dispute concerns the empirical claims, arguing instead that moral reasoning can be truth-tracking and causally determinative much more often, especially through what we can call *moderating* automatic processes. For example, an agent can become aware of automatic processes operating through consciously reflecting and hence rationally regulate its running further to influence one's moral behaviour or not. (Dreyfus & Dreyfus, 1991; Hogarth, 2001; Holroyd & Kelly, 2016; Kennett & Fine, 2009; Musschenga, 2011; Narvaez, 2011; Pizarro & Bloom, 2003; Sauer, 2012; Snow, 2006) These replies have contributed importantly to further advancing our understanding of moral cognition, and somewhat strengthened the case for moral reasoning, hence ameliorating the automaticity challenge.

However, it does not at all seem that agency is successfully defended with that. For one, these accounts face many critiques themselves. Most crucially, truth-tracking due to emotional and cognitive distortions, and limited frequency due to scarce cognitive resources, seem to be persisting issues (to be discussed in §3). Moreover, as I argue elsewhere (my doctoral thesis essay 1) the defences buy into the framing of the automaticity as an 'operant challenge'. Agency is conceptualised as an 'in-action' phenomenon, related to operationally determining some behaviour in a moral situation. In turn, the psychological question is which cognitive process is dominant during those operations. This conceptual framing informs the collection of empirical data as well as the subsequent philosophical interpretation of it, severely limiting the conceptual space for agency. As such, the debate is, at best, at an impasse.

In conclusion, while it is yet an open question exactly how much, how, when, and where people's moral life is automatic, there is now a large body of research and growing consensus that automaticity plays a very significant role in moral behaviour. As such, the automaticity challenge is still very much a heated and pressing issue, which targets shortcomings at the core of moral philosophy at large. This occasion

lends itself for a re-evaluation of the fundamental concepts involved, moral agency being the concept under discussion in this essay.

§2: Tripartite agency and moral character

In this section I present the *tripartite model of moral agency*. I start by introducing the core concepts of the model while disclosing some of the principal background commitments. Subsequently, I present and discuss an exemplary hypothetical scenario in order to gain an initial grasp of the model, before further elaborating on and substantiating the key concepts and claims.

One of the main claims that motivate the tripartite model is that moral agency cannot be conceptualised as one sufficiently uniform phenomenon, under one all-encompassing notion. Instead, three distinct *modus operandi* of moral agency can be conceptually distinguished, three *modes* of cognitive processing that warrant agency in one's moral behaviour; operant deliberative processing (*deliberative agency*), a combination of operant deliberative and operant automatic processing (*moderative agency*), and operant automatic processing with developmental deliberative processing (*developmental agency*). Capturing these three modes in unison calls for a tripartite model of moral agency (or 'agencies', if you will).

As such, by distinguishing three modes of agency, the tripartite model breaks with the paradigm that tacitly governs the automaticity literature, which is built on an operant notion of agency. One of the main purposes of this is to be able to appreciate all the ways in which an agent can exhibit agency, even though some of its instances may be too dissimilar from one-another to be captured under the same notion of agency. With that, the tripartite model aims to offer a way in which findings on automaticity as well as arguments in defence of rational agency can be appreciated within the same picture.

Besides the three agentive modes, *moral character* is another key notion. Moral character is invoked especially in relation to automatic behavioural processing, which may, albeit being automatic, nonetheless be agentive, due to the prior agentive development of one's automatic processes, or, character. One of the aims of appealing

to a notion of character is to bring together various distinct phenomena that are discussed within the automaticity literature at-large, such as habits, implicit biases, and unconscious stereotypes.

Besides the automaticity literature, a further background of the tripartite model that cannot be omitted is that it is unmistakably inspired by Aristotelian virtue ethical moral psychology. However, while drawing from this rich tradition, the account here is not presented as a neo-Aristotelian or virtue ethical account.

One reason for this is that such a relationship requires much more elaborate specification and contextualisation than the invoking of several seemingly shared basic concepts here allows for. Moreover, the tripartite model is presented, for now, as a neutral model of moral psychology, which may tie to various theories in metaethics and normative moral philosophy, not necessarily exclusively reserved to virtue ethics.

Nevertheless, it is worthwhile to point out the existence of association for three reasons; firstly, simply to pay homage when it is due, since Aristotelian literature has much inspired my thinking here. Secondly, virtue ethics provides to most readers a readily available framework to form an initial understanding of the tripartite model, by calling to mind a picture of character, character development, and acting from character. Thirdly, due to focusing on similar concepts, the tripartite model may face many of the same challenges that virtue ethical theories face, such that exploring relations in the future may be very beneficial in order to develop the model further.

2A: Amir

To start, consider the following hypothetical scenario about Amir as an exemplar of the three modes of agency that an agent may exhibit.

> *Amir grows up in a relatively segregated neighbourhood of a large western European city, and hence his social interactions at home, at his local school, and elsewhere, are for a large part limited to people with a similar immigration, lower socioeconomic class, and/or lower educational background. Initially (*A*), Amir picks up strongly sexist*

*implicit attitudes from his environment, and throughout adolescence he continues to possess these and typically act on them. During early adulthood (*B*), however, Amir becomes increasingly more aware of his attitudes and actions in the light of morality. He begins to actively relate to his character and behaviour more critically, evaluating and reflecting on it, discussing it with others, reading about related matters, practising employing other cognitions and exhibiting other actions for example while around friends, and attempting to moderate his moral behaviours at the meetings. Over time (*C*), Amir's implicit attitudes concerning gender change significantly, becoming increasingly egalitarian, eliciting egalitarian behaviour.*

All throughout this time, Amir is a member of a local community organisation and regularly attends meetings where affairs concerning the neighbourhood and plans for projects are discussed. At the meetings, the other members consist of both men and women, who all have opportunity to speak on the issues at hand.

During time period A*, Amir intuitively disqualifies the contributions voiced by women as less valuable and less important, he automatically listens with only little attention, has an attitude keener on picking out points of criticism, is disposed to wave away worthwhile input, and without a thought interrupts women very easily.*

During time period B*, Amir is aware of his character and effortfully works to re-evaluate, alter, and regulate it, for example through learning about admirable women entrepreneurs, explicitly telling himself to treat women fairly right before the meetings, by holding back on his initial responses during the meetings, and by focusing on the valuable parts of women's contributions, among other strategies, causing him to exhibit automatic sexist behaviour to a lower degree, but nevertheless Amir still has much of the same inclinations and attitudes as during time* A.

During time period C*, Amir engages in the meetings almost completely spontaneously, treating the contributions by men and women equally on their merit, and doing so positively effortlessly, simply acting, so to say, 'from character'.*[13]

Now, what can we say about Amir's moral agency when we analyse this scenario on a tripartite model? Starting at time period *A*, Amir's behaviour is largely driven by automatic processes. Since these automatic processes are passively acquired through socialisation, they are themselves non-agentive. As Amir is still young, he has only partly obtained a sense of rational capacity such that his deliberation is direct causal or can moderate automatic processes. Additionally, his deliberative and moderative

[13] I do not wish to suggest that sexist attitudes in some way belong to (Middle Eastern) immigrant populations, or poorer people for that matter, as sexism is prevalent among native White and wealthier people just as well. Rather, the scenario aims to stress the importance of intersectionality in the tripartite model, through showcasing that opportunity for moral development can be limited by various axes at the same time, including class, race, ethnicity, sex, and housing and schooling segregation, among many other factors.

agency are further diminished by the limited environmental opportunity of his so-called 'low opportunity environment'. As for developmental agency, this too is diminished by his capacity and environment, and his age simply temporally restricts his opportunity to have engaged in his character development. As such, we can conclude that overall at this stage of his life Amir typically lacks every type of agency over his moral behaviour.

During time *B*, as an adult, Amir has full rational capacity, albeit possibly somewhat diminished by his environment. Amir exhibits deliberative agency through, for example, consciously choosing certain moral behaviours. And Amir exhibits direct and indirect moderative agency through, for example, cognitive intervention and cognitive preparation, respectively. Amir is also actively engaged with his character development, for which he now has the capacity and time, although still being limited by his environment. However, Amir still possesses the biases he did at time *A*, such that he does not yet exhibit developmental agency.

During time *C*, when Amir exhibits behaviour that is driven by automatic processes, he exhibits developmental agency, since he was agentively involved in developing these processes over-time.

2B: Agency and rationality

One key notion that the analysis on a tripartite model above employs is *rational capacity*. While adequately going into this notion is beyond the scope of this essay, it may be helpful to mention several issues concerning the way in which agency and capacity are employed here.

Firstly, it may be noted that virtually any model of moral agency and action draws, at its foundation, on some notion of rational capacity and activity, as a way of connecting an agent to some behaviour *as agent*. This rational connection between the agent and the behaviour is not merely important in order to understand the behaviour as the agent's behaviour, but also, for example, to evaluate the agent as morally responsible for the behaviour as agent. The connection can be conceptualised in many different ways, leading to different claims regarding what rationality gives us, such as

that it makes the agent 'causal' to the behaviour, or makes the behaviour 'intentional', or makes it 'belong' to the agent.

Two very influential accounts on this issue are the so-called 'reasons theories of action' by Christine Korsgaard and David Velleman. (Korsgaard, 2008; Velleman, 2000) On such theories, 'actions' can be distinguished from other phenomena such as 'happenings', because the latter merely *happen to* an agent, while actions are *done by* an agent since one's capacity for reasoning being is in some way involved. On their accounts, as for many others, rational capacity is involved in order to ensure that moral behaviour can be guided in a way that is responsive to the relevant moral considerations. Regarding the question of what rational capacity *does*, the way in which the notion is employed throughout this paper may be thought of along the lines of accounts such as those by Korsgaard and Velleman.

Regarding the question of what rational capacity *is*, the way that the notion is used here is much in the same mind of what I take to be a widely accepted notion of 'normative competence', especially as conceived by Susan Wolf. (Wolf, 1990, pp. 121-124) On a loose formulation, rational capacity is an agent's ability to be reasons-responsive, to be aware of oneself, of one's behaviours, and of one's environment, to be able to reflect upon these all, to be receptive to moral knowledge, moral facts, and moral values, to be able to acquire and possess knowledge, among many other abilities, and to be able to bring all these traits to bear on one's moral behaviours.

Following David Brink and Dana Nelkin, I also take an agent's *environmental circumstances* to be partly constitutive of one's agency in addition to, and in interaction with, one's rational capacity. (Brink & Nelkin, 2013) As Brink and Nelkin argue in the context of moral responsibility, not merely an agent's internal capacities determine the opportunities one has, but also one's situational circumstances, such that making a multi-faceted judgement of one's responsibility for some behaviour requires information about both.

For example, jumping on a grenade may well be within a soldier's capacity, but as it is a situation that requires extreme sacrifice, this influences the agent's responsibility for doing so or not. Similarly, drawing on the case of Amir, as an adult he may possess the rational capacities to moderate his automatic behaviour and develop his character, but due to his environment this is a very effortful enterprise, such that his agency in these respects is diminished. Thus, environmental

circumstances can make some behaviour more difficult, either by requiring sacrifice or effort, and this influences the agency that an agent has over the behaviour, together with one's rational capacity.

As such, agency consists of rational capacity and environmental circumstances, which interact on each other, and jointly determine one's opportunity for agency.

A further peculiarity about rational capacity is that it can be seen as a condition for agency while an agent does not initially have agency over having capacity in the first place. Rather, capacity is, at least at first, largely a matter of biological and social luck. An agent's capacity can gradually come into being, but it can also be said to gradually increase, partly so due to the agency of an agent.

Again, drawing on Amir's scenario, he does not merely gain increased capacity say 'naturally', through biological and social cognitive maturation, but he actively works on increasing his own capacity. For example, through intentionally expanding his knowledge on sexism, Amir makes himself more receptive to sexist stimuli and better able to respond to such stimuli. With this conception of how an agent can be involved in developing their own capacity, we have already started to address character development. On the broad notion of character that is employed in this paper, an agent's rational capacity is part of one's moral character, together with a wide range of other mental phenomena. However, before elaborating on character, let us first draw out the tripartite model more.

2C: Three modes of agency

As the brief analysis above aims to show, we can distinguish three different modes in which Amir exhibits agency over his moral behaviour. I will now describe what the three modes are, how they are exhibited, and why they are distinct modes of agency. The main features of the three modes are comprised in table 1. After this, the rest of the paper will concern further explaining and substantiating these modes and their related concepts.

Agency mode	Objective	Cognitive behavioural process
Deliberative agency (*prohairesis*)	Action (direct)	Conscious deliberative processing directly causally determining some moral behaviour
Moderative agency I (*enkrateia*)	Action (direct)	Cognitive intervention on active automatic processes during the processing of some moral behaviour
Moderative agency II (*enkrateia*)	Action (indirect)	Cognitive preparation or environmental regulation just prior to the processing of some moral behavioural to prompt activation of certain automatic processes
Developmental agency (*aretê êthikê)*	Agent (development)	Moral character development over-time and resulting developed automatic processes subsequently determining some moral behaviour

(Table 1: *Tripartite model of moral agency*)

The first agentive mode is *deliberative agency*. This may be the traditionally most appreciated concept of agency, at least within a dominant rationalist paradigm. As the, say, most 'pure' form of agency, deliberative agency involves some sense of conscious deliberation that is directly causal and the dominant. There are copious different versions of this mode available in the literature, ranging from 'merely' requiring introspective access into the operant reasons of one's behaviour, to explicit deliberation and control through choice (see thesis essay 1).

While I am open to different ways of spelling out this mode, the core principle is that one's deliberative cognitive processing mainly drives one's moral behaviour. Moreover, the relation to one's behaviour is fairly direct. In Aristotelian terms we could think of this mode as *prohairesis* (conscious choice) resulting from *boulesis* (rational reflection) as virtually the sole significant factor determining one's behaviour.[14] Such agency is especially useful in novel, complex, and weighty decisions where one can calmly contemplate and inform oneself. To illustrate this mode of agency, one can think of, for example, a teacher who grades students' papers by carefully assessing the relevant factors, or a judge in the court of law who ponders over the weight of various moral and legal considerations in order to form a ruling, or parents who are deliberating whether to send their child to the higher-appraised White homogenous, or the lower-appraised more ethnically-mixed school.

The second mode of agency is *moderative agency*. This mode is marked by an operant interaction of deliberative and automatic processing. This means as much as that deliberative processing in some way works on or with some active automatic processes, in order to regulate the latter's influence on some behaviour. In

[14] This is only one interpretation of Aristotle's concepts, which is merely meant for the purpose of clarifying the agentive modes, without entering into discussion within Aristotelian scholarship.

Aristotelian terms, one may think here of behaviour that is determined through *enkrateia* (mastery) over one's occurring *pathos* (emotions, including, for our discussion, automatic processes such as implicit biases) through one's deliberation. Note that, with enkratic behaviour, the agent still possesses the automaticity, but it just does not determine one's behaviour. While moderative agency is defined by interaction of deliberative and automatic processes, there are various ways in which this interaction can take place, such that we can distinguish different sub-modes. The first distinction we can make is between directly or indirectly moderating the influence of automatic processing.

Direct moderative agency is exhibited through *cognitive intervention*, an agent's deliberative processing intervening on certain automatic processes that are activated, in order to regulate whether or how they further influence one's behaviour. This mode of agency shares its operational directness with deliberative agency, but is distinct due to involving interplay between deliberative and automatic processing rather than mainly 'pure' deliberative processing.[15] As an illustration of this, one can think of a teacher who pays attention to operant biases and adjust the evaluation of papers in accord, or Amir who can notice and block interruptive inclinations driven by sexism.

Indirect moderative agency is also exhibited through deliberatively regulating the behavioural impact of automatic processes, but doing so just prior to the latter being activated and operant. As such, this sub-mode is still action-focused, but relates to it somewhat less directly. Indirect moderative agency can itself be distinguished in two broad categories. *Cognitive preparation* involves prepping one's own cognition such that certain automatic processes will be likely to be activated. For example, a teacher can, before starting the task of marking papers, formulate the goal to oneself to do so in an egalitarian manner, and Amir formed the intention to ensure paying attention to valuable content in women's discussion contributions. The other form of indirect moderative agency, *environmental regulation*, involves engineering the situational context in which one's moral behaviour will take place in such a way that when one later processes some moral behaviour, the situation is likely to elicit certain desired automatic processes (or not elicit unwanted ones). For example, a teacher can

[15] To conceptually distinguish three modes of agency in terms of the mechanisms that are involved, I employ the term 'deliberative agency' (and behaviour), rather than 'direct agency' (and behaviour). With this, I hope avoid confusion with concepts such as 'direct control', which typically means 'direct operational impact on behaviour', because such operational directness does not exclusively apply to deliberative agency, but also to 'direct moderative agency'.

anonymise students' essays so that there is no personal information that may trigger implicit biases that may subsequently influence the marking, and Amir can push for an equal gender-representation in meetings causing women to feel more empowered to speak and men more compelled to acknowledge each individual woman's contribution.

The third mode of agency is *developmental agency*. In this mode, an agent consciously engages with the configuration of one's automaticity, or rather, the development of one's own moral character. In turn, one's moral behaviour is mostly driven by one's moral character. As such, the operational phenomenology of the determination of behaviour is automatic, but the agency is located in the developmental stage earlier on where reasoning is imbued in the configuration of the automaticity. In Aristotelian terms, again, one may think of behaviour from *aretê êthikê* (virtuous character), or with *phronesis* (practical wisdom), without operational prohairesis, boulesis, or enkrateia, but instead against a history of conscious character development. In order to further define developmental agency, it may help doing so by focusing on several differences with other modes of agency (especially indirect moderative agency).

1. Other agentive modes are action-focused (either direct or indirect), having the determination of some moral behaviour as its objective. Developmental agency, instead, is agent-focused, concerned with the formation of the agent's constitution (developmental).

2. Other agentive modes are enkratic, meaning that while one's automaticity does not operationally determine one's behaviour, the configuration of one's automaticity is typically retained. Developmental agency, in contrast, reconfigures one's automaticity (this point is elaborated later).[16]

3. Other agentive modes, even indirect moderative agency, are exhibited 'in-action', either within or just-prior to some moral situation (3a, *temporal*

[16] Admittedly, indirect moderative agency *can* have an impact on the configuration of one's automaticity, but all modes of agency can. Rather, what matters to distinguish them is what the main aim of each mode is (and, in addition, the effectiveness of achieving that – as the developmental side-effects of deliberative and moderative agency is not very effective, I argue in doctoral thesis essay 2).

demarcation), through one or several clear behavioural instances (3b, *behavioural demarcation*), that are performed within a limited time-frame of each other and of the ultimate moral behaviour (3c, *temporal demarcation*), in order to ultimately determine one or one or several specific moral behaviours (3d, *foreseeability*). Developmental agency, on the other hand, can be exhibited 'offstage', not within, but rather 'detached' from any moral situation, at virtually any time and place, say, during one's free time, home-alone (3a), and is done through numerous behaviours (3b), which are performed over an extensive period of time (3c), which, in turn, has an effect on a wide range of a person's ultimate moral behaviours that are driven by one's character (3d). (See my doctoral thesis essay 3)

4. Other agentive modes employ 'operant reasoning', reasoning in-action in a moral situation, which is thus vulnerable to all the automaticity challenges. Developmental agency can employ 'detached reasoning', in less hostile environments, framed to be less emotionally heated, with less time-pressure, and less scarcity of cognitive resources (this point is elaborated later).

5. Other agentive modes have a limited 'factor range', only relating to factors that can be clearly specified and comprehensibly determined, such as some specific knowledge, or a certain state-of-mind. Developmental agency targets one's moral character, which comprises a wide variety of factors and their interrelations with one another, such as emotional dispositions, attitudes, values, and skills (this point is elaborated later).

Such are the basic features of developmental agency, which will be elaborated throughout this essay.[17] Illustrating this point, one can think of a violinist, who

[17] Some readers may, despite this all, nevertheless find the conceptual distinction between developmental agency and indirect moderative agency insufficient, rather perceiving the former as sub-species of the latter. While there is undeniably some likeness, I do think the differences mentioned here sufficiently distinguish the modes. Nevertheless, as this point is so central to the essay, I will say a little more.

Especially coming from the moral responsibility literature, one may see both as forms of 'indirect control'. However, concepts from one literature often do not neatly map onto another literature. The notion of 'control', for example, is defined in countless different ways. The three agentive modes are initially explored from within the moral psychological literature. As argued elsewhere (my doctoral thesis essay 1), the automaticity challenge to moral agency is framed as an 'operational challenge', in which agency is conceptualised as in-action phenomenon, and the question is what cognitive process is dominant during the operation of cognitive processes while determining some moral behaviour. Moderative agency fits within that paradigm, as seen by the many authors who champion forms of it. Developmental agency falls outside of such a paradigm, as it does not employ

develops their skill not as much on-stage, while giving a performance, but rather off-stage, through spending large amounts of time practising, at home, in other training settings, playing, listening, discussing, and even reading about music. Translating this to morality, a teacher can read about educational performance of students from various class backgrounds, attend workshops on the functioning of biases concerning ethnic groups, and decorate one's wall with images of intellectually and culturally admirable working-class people. Amir can reflect on and evaluate his attitudes, watch a documentary on women entrepreneurs and inventors, and practice other thought-patterns and reactions. All of this can be done at any time and is done principally to develop one's moral character (not in order to determine some specific moral behaviour in a specific moral situation).

2D: Moral character

Besides the three modes of agency, the notion of *moral character* is very central to the tripartite model, such that more clarification of what this exactly entails is required.[18] I will here describe what the nature of character is, how character functions in (behavioural) processing, and what the development of character involves. Having clarified character and the other main concepts of the tripartite model, the rest of the paper will be concerned with substantiating this view.

---

agency as merely an operant concept, or automaticity as a challenge that only (or even mainly) relates to which cognitive process is operationally dominant. Translating this to the moral responsibility literature, one could think that indirect control already goes beyond the borders of the moral psychological operant concept of agency. As I argue elsewhere (my doctoral thesis essay 3), I do not think it does, because it would overstretch the notion within the paradigm it functions (volitionist theory) in and require giving up some central principles of that paradigm.

All that said, even when one may not be convinced by those arguments, I maintain that making this conceptual distinction is justified due to its usefulness. Making conceptual distinctions, which is a large part of philosophy in general, is a project of untangling the fundamentally entangled and 'untanglable' reality, with the aim to understand phenomena better, allowing for more nuanced analysis, and in turn design better strategies to cope with them. Here, conceptually emphasising the importance of character development is the goal. And this, above all else, I aim to achieve through focusing on the differences of the concepts by exploring the strictest version of each. As said, if, in the end, others embrace character development within moderative agency or indirect control, then this can perfectly be considered as a success.

[18] Situationist theories have contested the empirical plausibility of the existence and significance of philosophical conceptions of character as explanatory of moral behaviour, arguing that such accounts are systematically mistaken in their understanding of character traits as people's individual, general, and robust dispositions, while in fact people are swayed by situational factors much more. However, the situationist challenge has been amply argued against for misrepresenting the concept of character, being empirically and theoretically incoherent, and building on shaky (interpretations of) empirical data, in addition to which evidence supportive of character has been presented. The scope of this essay does not allow me to address this matter further, but I take it not to constitute a serious challenge, as at worst the challenge and defence of character is yet unresolved, and at best the support of character is looking bright. For situationist critiques, see: (Doris & Stich, 2005; Doris, 2002; Harman, 1999; Ross & Nisbett, 1991; Sabini et al., 2001). For responses, see: (Brownstein, 2016; Krueger & Funder, 2004; Merritt, 2000; Miller, 2013, p. ch1; Nelkin, 2005; Sreenivasan, 2002; West, 2017).

The notion of moral character employed here is very broad one, and, for now, quite general. I employ moral character as an umbrella term for a large family of mental phenomena. As such, on a very loose definition, a person's moral character is the entire set, constituted by wide variety of phenomena, ranging from an agent's beliefs, desires, values, affective attitudes, behavioural dispositions, emotional reactions, spontaneous responses, intuitions, and habits, to what grabs one's attention, perceptions, sensitivities, skills, and one's rational capacities, among other things. And thus, most automatic states and processes, such as implicit biases, unconscious stereotypes, and prejudices, are part of one's character too.

It may well turn out that not all these phenomena, or not all instances of these phenomena, belong to one's character, but I will leave such more precise demarcation for a later discussion. For now, the aim is to unify a wide range of different sorts of mental phenomena in order to create a model that can account for all of these together in our theorising about moral agency. I will now put forth four reasons for this unified conception of moral character: automaticity, integration, function, and development.

The first essential and shared quality of characterological phenomena is that they all admit of a high degree of automatic processing. As such, the aim of jointly modelling these phenomena is to conceptualise moral automaticity in its entirety, involving its various forms and products. Now, to be clear, many of these phenomena can also function in non-automatic modes, such that they are not necessarily always automatic. For example, Amir's belief that e.g. "women are less intelligent" can function automatically through his implicit biases, but also in his conscious explicit deliberation. Much of the discussion in this paper will focus on implicit biases, as this is the most discussed automatic phenomenon in the literature. Nonetheless, aiming for a general model of automaticity, the discussion of implicit biases is meant to apply to other automatic phenomena just as well (although some adjustments may be required for specific phenomena).

A second feature of characterological phenomena is that they are highly integrated with each other, such that they cannot properly be conceptualised on their own. Both in the acquisition, possession, and expression of automatic mechanisms there is a high degree of connectedness. For example, Amir's implicit sexist bias involves various

beliefs (e.g. “women are less intelligent” and “men are better leaders”, values (e.g. “it is acceptable to interrupt inferior speakers” and “men should be in charge”), and attitudes (e.g. interruption propensity and critical scrutiny towards women’s contributions), which also function in connection to other phenomena. As beliefs, desires, and behavioural attitudes are typically seen as part of one’s moral character, automatic phenomena such as implicit biases, in virtue of consisting of these and other highly interconnected phenomena belong to one’s moral character too. Comparable conceptions can be found in literature on moral responsibility. For example, Angela Smith considers implicit attitudes as being embedded in rich inferential relations with the states of an agent that jointly make up one’s practical identity, tied up together through an on-going (albeit unconscious) process of identifying, evaluating, and accepting putative reasons in favour of the attitudes, which involves rational activity, such that the entirety can be attributed to the agent as being expressive of who one is. (Smith, unpublished, pp. 21-22) Similarly, Elinor Mason argues that implicit biases are tied up with one’s ‘deep motivations’, such as contempt, disgust, or a disposition such as ‘openness to accept hierarchies that favour us’. (Mason, forthcoming, p. 5) And Jules Holroyd and Daniel Kelly too set out an account of the interaction of implicit bias with an agent’s values such that the whole is reflective of the agent and can be understood as one’s moral character. (Holroyd & Kelly, 2016, p. 3)

Moreover, what is important about the interconnectedness of characterological phenomena is that an agent’s self-development, as will be discussed later, can also not be conceptualised as targeting one isolated factor, but rather has to concern a broader development such that we can speak of *character development*. For example, for Amir to develop his automatic sexist behaviour, he does not merely challenge one isolated belief or even one bias concerning e.g. women’s intelligence. Rather, Amir has to engage in a wide variety of developmental efforts concerning an equally wide variety of characterological states, e.g. reflecting on what his beliefs and values are regarding men and women in different respects such as intelligence and leadership, scrutinising these in the face of new evidence including seeking out information about sociocultural influences on the differential development of traits in men and women, and practising with new associative structures between situations and possible reactions.

Thirdly, another important feature of the automatic phenomena under discussion here is that they share a common functional role. As Lorraine Besser-Jones argues, because implicit biases produce emotional and behavioural dispositions, which influence an agent's behaviour, they belong to the agent's moral character. (Besser-Jones, 2008) One way of defining what kinds of phenomena are part of one's moral character is by behavioural impact, and in particular the way in which behavioural impact is achieved. For example, on this ground, beliefs can be taken to belong to one's moral character due to their role in shaping dispositional behavioural effects, influencing judgements, decisions, and actions in moral matters. Similarly, then, implicit biases belong to one's moral character since they too do so. Again, clarifying the point through Amir's case, Amir's implicit sexist bias functions in disposing him to certain behavioural tendencies, and it is for this reason that we can understand his bias as being part of what his moral character is.

The fourth and principal feature of automatic phenomena that supports understanding these as characterological phenomena is their developmental nature. There is an extensive body of research evidencing the malleability of automatic states and processes, ranging from implicit attitudes to habits. The rest of this paper will mainly be concerned with substantiating the idea of developing automaticity and elaborating on various ways in which this can take place. As may be clear, this approach draws on a specific underlying understanding of cognition that is not in terms of strict Fodorian modularity, but rather, as also Keith Stanovich and Richard West, one of the first to index automaticity models, recognise; "system 1 processes result from more than just innate specification. In our work on cognitive models of reading acquisition we were among the first to stress the importance of the concept of acquired modularity (…) Specifically, System 2 can strategically arrange practice so that habits of decontextualization become automatic." (Stanovich & West, 2000, p. 709)

The main point of this, for the tripartite model, is that, given that automatic phenomena can be developed, this opens up for the possibility of an agent exhibiting agency over one's own automaticity, through agentively developing one's automaticity. The possibility of agency over one's automaticity provides a very strong fundament to conceptualise one's automatic phenomena as being part of one's moral character, since it makes it possible that one's automatic states and process are due to one's agentive efforts, and as such one's automaticity can reflect one's agency.

§3: Moderative agency

The first auxiliary mode of agency, beyond traditional deliberative agency, is moderative agency. A range of authors has already advanced proposals for an extension of deliberative agency. While they typically do not frame their efforts as introducing a different agentive mode, taking them all together reveals that this is in fact what they are doing. Therefore, I will now discuss their work as a way of substantiating this distinct agentive mode. The ground for grouping these approaches together as accounts of moderative agency is that, while different authors propose different strategies, they all share the essential quality of determining some moral behaviour through a fairly direct operant impact of deliberative processing on automatic processing.

After setting out some of the main instances (and, importantly, categorising these), I will critically discuss this overall approach, arguing that, while moderative agency provides an important addition to deliberative agency, it is insufficient as an answer to the automaticity challenge. However, these approaches do point us in the right direction for a third agentive mode, the addition of which *does* suffice as defence of moral agency, as I will argue in §4.

Moderative agency focuses on moderating the influence of some operant automatic processes in determining some moral behaviour, through involvement of deliberative processing. As said earlier, this can be done in two ways, either directly or indirectly, but the core of both strategies is moderative action-focused. Various authors have advanced forms of moderative agency, which I will now briefly mention, categorising them as direct strategies and different types of indirect strategies.[19]

3A: Cognitive intervention

To start with direct moderative agency, we there are copious proposals that we can categorise as *cognitive intervention* strategies. The essence of this strategy is that an

[19] The focus here is on the critical discussion of this entire mode of agency. For that reason, the proposals by various authors are only presented in a highly summarised form. For a more elaborate explanation of them, see §2 of essay 1 of my doctoral thesis.

agent makes use of conscious deliberative effort to moderate the behavioural impact one's activated automaticity while within a moral situation.

In psychology, Augusto Blasi has argued that people can control their automatic processes through moral reasoning. This involves analysing and evaluating an automatically formed judgement or spontaneous emotion once it occurs, from which an agent can "accept it or reject it according to one's criteria of validity and truth." (Blasi, 2009, p. 423) Similarly, Robin Hogarth, champions the idea that agents can interrupt operant intuitive processes and engage in what he calls 'intuition reflection, assessing the validity of one's intuitions through 'scientific thinking'. (Hogarth, 2001, p. 219) Darcia Narvaez, in turn, argues for the possibility of agents monitoring and interpreting automatic processes. (Narvaez, 2011, p. 38) And many more such approaches can be found throughout the psychological literature, such as 'reflective consideration' of automatic affective processes (Atkins et al., 2004, p. 66), deliberatively determining the appropriateness of intuitions through 'involved reasoning' (Dreyfus & Dreyfus, 1991), or reflective 'perspective taking' (Pizarro & Bloom, 2003). Lastly, one can think of the widely popular conscious effort to slow down one's thinking as to think more critically, as promoted by Kahneman. (Kahneman, 2011)

In philosophy, Bert Musschenga explores different ways in which moral reasoning quality can be improved; partly in order regulate the influence of emotions in the case that they lead to flawed intuitive judgements. (Musschenga, 2011, p. 80) Nancy Snow proposes the possibility of 'intervention control' and 'inhibition control', the capacity of agent to redirect or entirely stop some habitual action through a conscious effort. (Snow, 2006, p. 550) Jules Holroyd and Daniel Kelly, and Hanno Sauer support and further Snow's proposal, which Sauer then names '*ex post* education', the reflective monitoring and evaluating of one's cognitions. (Holroyd & Kelly, 2016, p. 6; Sauer, 2012, p. 269) And, lastly, Jeannette Kennett and Cordelia Fine defend what they call 'after-the-fact correction', regulating or endorsing some already active automatic process. (Kennett & Fine, 2009, p. 78)

For example, through conscious, effortful deliberation of another person's position, empathic emotions can be intentionally aroused in order to oppose automatically triggered intuitions. (Batson, 1998)

3B: Cognitive preparation and environmental regulation

Concerning indirect moderative agency we can distinguish two types of strategies that have been proposed by various authors, cognitive preparation and environmental regulation. Similar to direct moderative agency, the focus is on operant automatic processes in determining moral behaviour. Also, just like direct moderative agency, the agent makes use of conscious deliberative cognitive effort to moderate one's automaticity. The difference is that indirect moderative agency does this not *in* the moral situation, while the automatic processes are already activated, as direct moderative agency, but rather just prior to a moral situation.

Cognitive preparation strategies involve preparing one's cognition such that certain desired automatic processes are more readily activated when encountering anticipated stimuli in an anticipated moral situation.

Holroyd and Kelly expand Andy Clark's notion of 'ecological control' as a strategy of influencing one's moral behaviour without in-the-moment deliberative processing, but instead deliberative manipulation of one's mental states "as to shape their cognitive processes, thus enabling the exercise of ecological control in the future." (Holroyd & Kelly, 2016, p. 8) Similarly, Kennett and Fine discuss what they call 'up-front control', Hanno Sauer argues for '*ex ante* education', and also in psychology David Pizarro and Paul Bloom advance prior 'cognitive appraisal shift'. (Kennett & Fine, 2009, p. 78; Pizarro & Bloom, 2003, p. 194; Sauer, 2012, p. 267)

All of these proposals mainly draw on a psychological phenomenon known as 'implementation intentions', mostly based on Peter Gollwitzer's research, in which when agents form explicit 'if-than plans' for oneself, before a situation, to respond in a certain manner to specific stimuli. (Gollwitzer, 1999) For example, participants in a study formed intentions to remain calm while facing fear-inducing images, which reduced their subsequent automatic disgust and fear responses. (Gollwitzer et al., 2009) Or, participants in a study on the 'shooter bias', concerning the bias of Black people as more dangerous and subsequently falsely identifying them as carrying a weapon, can adopt the plan, "if I see a Black person, I will think 'safety', causing less misidentification. (Mendoza et al., 2010)

Environmental regulation strategies involve manipulating or selecting the environment one is or will be in, such that it will be more plausible for certain environmental triggers to be there or not, which, in turn, will trigger certain desired or not trigger certain undesired automatic responses.

Holroyd and Kelly discuss this strategy as another form of ecological control, Sauer as another form of *ex ante* education, Kennett and Fine as another form of up-front control, Pizarro and Bloom argue for the same while calling it 'input control', and Hogarth champions 'environment selection and/or creation'.

As a very clear example of environment selection, one can think of a dieter who chooses not to walk through the grocery store section with ice cream as to avoid temptation being automatically triggered. (Schelling, 1984) Alternatively one can, for example, remove gender and racial information from student essays or job applications. Environment engineering can be done by, for example, putting up counter-stereotypical images of admired Black people on one's office wall, which counter the operation of racial biases. (Dasgupta, 2013; Dasgupta & Greenwald, 2001)

3C: Critical analysis of moderative agency

There are several problems with moderative agency as an answer to the automaticity challenge, which relate to cognitive resources, magnitude, opportunity, foreseeability, and perpetuity.

The first issue, which concerns all three forms of moderative agency, is that these strategies are very demanding of cognitive resources for deliberative processing, which are scarce, due to which these strategies can only be exhibited infrequently. Firstly, I acknowledge the success of the authors above in opposing the automaticity challenge's initial very limited conceptual space for deliberative and moderative agency and substantiating various ways in which agents do have realistic ability for these modes. However, while more than the initial automaticity challenges may have recognised, people's capacity for these modes will nevertheless be quite limited.

As Kennett and Fine acknowledge, moderative processes, like deliberative ones, depend on available 'controlled cognitive processing' resources. (Kennett & Fine, 2009, p. 92) Research on such resources, including 'working memory capacity', and 'attentional resources', among other resources involved in effortful reasoning processes, has shown these to be very scarce, such that they are easily depleted and take time to regenerate. (Chaiken, 1987; Galinsky & Moskowitz, 2000; Monteith & Voils, 1998) For example, social psychological research on suppressing activated ethnic stereotypes or inhibiting primed traits shows this. (Barrett et al., 2004; Govorun & Payne, 2006; Monteith et al., 1998; Payne, 2005; Thompson et al., 1994) And research in cognitive neuroscience on inhibition of emotional responses have similar findings, for example showing dorso-medial prefrontal cortex activation, a brain region linked to voluntary self-control processes, especially conscious reasoning. (Blair et al., 2001; Greenwald & Banaji, 1995; Kawakami et al., 2000; Kühn et al., 2014)

As such, since these agentive modes rely on resources that are scarce, the usage is limited, which makes it that agents can only infrequently avail themselves of these strategies. (Haidt, 2003; Kühn et al., 2014; Perkins et al., 1991) In other words, reliance on scarce cognitive resources causes that agents cannot exhibit deliberative and moderative agentive modes very often, such that it is unlikely to warrant agency over the majority of one's moral behaviour (given the automaticity's claim that the majority of people's behaviour is driven by automatic processes).

Another issue, which most clearly affects cognitive intervention, but other moderative strategies too, is that even when agents are aware of operant biases, they typically do not know magnitude to adjust their behaviour by. Awareness of, for example, being primed by an irrelevant factor, in addition to knowing the direction of the prime effect is insufficient for appropriate adjustment of one's behaviour. For example, awareness that a 'kindness' prime disposes one to evaluate others as more kind than one otherwise would, causes the agent to counteract the effect. But rather than annulling it, people typically overdo it, as they are unaware *how much* they are biased and thus how much to adjust their response, resulting in what is known as the contrast effect. (Lombardi et al., 1987; Martin, 1986; Martin et al., 1990)

Moreover, even when people are, in turn, aware of this difficulty, they still face great difficulty attempting to assess the respective influences. For example, even

when forewarned (a cognitive preparation strategy) about the 'halo effect' (positive evaluation of someone due to some specific trait), people still failed to correct their evaluation of a professor, unable to distinguish the influence of their liking of the professor, despite participants' self-indication that they were only minimally affected by the likability. (Wetzel et al., 1981)

A further issue, which affects mainly environmental regulation strategies for indirect moderative agency, is that the opportunity to either select or engineer one's environment is very limited. While an agent can choose to avoid the ice cream aisle, there will be numerous stimuli for unhealthy food such as advertisements throughout one's day that one cannot avoid. Moreover, most of one's environments will probably be saturated with that, which is how one acquired the issue with unhealthy food in the first place.

The same goes for, for example, racial biases. While one may sometimes influence one's behaviour through introducing counter-stereotypical stimuli into their environment, most of one's behaviour throughout the day will be exhibited while there are stereotype-triggering stimuli, since our society is filled with negative portrayals of people of colour, which is how one acquires said bias or stereotype in the first place. If we were able to engineer environments that will mostly trigger egalitarian automatic responses, automaticity would hardly be as problematic, as people would not have acquired undesirable automatic attitudes and would not have undesirable automatic processes triggered. As Godsil argues, racial anxiety is typically developed due to a homogenous environment, in virtue of which (especially positive) intergroup interaction to challenge such attitudes is unlikely. (Godsil et al., 2014, p. 50; Plant & Devine, 2003). In the light of how common racial segregation is, this is a severe problem for environmental regulation. (Powell, 2012) As a personal example, while I can, do, and should[20] counter classist, racial, and sexist automaticity that influences my moral behaviour at university, my university environment is predominantly male, especially White, and even more upper-class[21], such that the photos of admirable counter-stereotypical intellectuals on my wall will provide automatic behavioural stimuli that are most likely dwarfed by the typical university

[20] Regarding whether one 'morally should' counter biases, see essay 3 of my doctoral thesis on moral responsibility for automaticity and character.

[21] While 23% of the Oslo population is categorised as immigrant, only 14% of the students at the University of Oslo, where also students from non-skilled working class background are 30-35 times less likely to obtain a degree than those from educated families. (Hovdhaugen, 2013, §3.11/6.2)

stimuli, since most of the moral situations I encounter are not in my 'safe space', 'positive prime' office. As such, this will often not be successful as moderative strategy (although, as I will discuss later, it may be more effective as developmental strategy).

As such, environmental regulation may be occasionally useful, but hardly sufficient to warrant agency in the face of automaticity. Moreover, automaticity poses us with a problem of moral agency *within* a morally difficult situation (the difficulty constituted by automaticity and its stimuli), which cannot always be circumvented by avoiding or changing that situation, but rather requires strategies to actually deal with it.

An issue that plagues both forms of moderative agency, cognitive preparation as well as environmental regulation, is that the specificity of these strategies target particular situations, while morally challenging situations are oftentimes not foreseeable. In some cases, an agent may well be aware of some future morally problematic situations, for example in the case of scheduled job interviews or student essays that will have to be marked at the end of the semester. In such cases, the agent can reasonable expect that specific automatic processes will operate, and based on that knowledge employ strategies to moderate their automaticity beforehand.

However, all too often, a morally challenging situation cannot be foreseen. For example, one could be asked to fill in for a colleague who was supposed to conduct the job interviews but suddenly fell ill. It seems plausible even that a large amount, if not most, of one's moral situations are not foreseeable in such a way that an agent has the possibility for prior moderation. As such, while one may prepare oneself to challenge their racial biases during a certain situation, they will still operate and influence one's behaviour outside of that. What is more, while one may foresee and prepare oneself for a certain specific automaticity challenge, say racial biases, there will often be many other operant processes, such as class biases, someone's physical beauty, the agent being hungry, the weather being sunny or not, among others.

As with the resource and opportunity issues, the specificity of moderating a particular bias in a particular situation makes that one can only occasionally foresee and exhibit such agency. While on such occasions it may be a very useful agentive strategy, it will hardly be sufficient given the ubiquity of automaticity.

A final issue concerning all three moderative strategies is that they are mainly operant, action-focused modes of agency that will need to be perpetually exhibited. Cognitive intervention, cognitive preparation, and environmental regulation are aimed at influencing some particular moral behaviour through moderating the influence of automatic processes. In other words, such modes of agency target an instance of the expression of automaticity, but not the acquisition and possession of a particular configuration of one's automatic cognitive structures. While this may sometimes be successful, such that the agent exhibits agency over their behaviour, the automatic processes that are moderated on that occasion nevertheless remain in existence as a part of the agent's cognition, whereby the agent will have to moderate them again and again in subsequent moral situations.

Admittedly, moderative (and deliberative) agency may, besides the operational effect, also impact the configuration of one's automaticity, as most of the authors discussed above are keen to mention. However, such 'mention' merely highlights a beneficial side effect, but not much more than that, because the modes are primarily operational, aimed at behavioural effects, not developmental, aimed at achieving a reconfiguration of one's automaticity. With the nature of these modes being behavioural, the developmental impact that may come about as side effect will probably not be sufficient to actually change one's automaticity. For one, there will be an abundance of stimuli in one's environment that reinforce one's current automatic configuration, as argued above. And secondly, unless the majority, or at least a significantly large part, of one's moral behaviour is under deliberative or moderative agency, one's behaviours will express one's current automaticity and thus mostly reinforce rather than challenge one's automaticity (given the various challenges for deliberative and moderative agency above, it does not seem plausible that agents are able to exhibit agency over their moral behaviour often enough so that a sufficient amount of their behaviour is in line with reflectively endorsed reasons rather than automaticity).

As such, one will have to perpetually moderate one's automatic processing. As with the platitude of feeding the poor rather than fighting systemic poverty, the root cause of the problem of automaticity is not addressed through operant, behavioural modes of agency that mainly address the effects of automaticity. With this, moderative agency will not be sufficient in actually successfully defending moral agency in the light of the automaticity challenge.

In conclusion, while many of the empirical premises in the entire discussion on automaticity are yet inconclusive, it does not seem that moderative agency in addition to deliberative agency provides a sufficiently robust notion of moral agency that can live up to the automaticity challenge. One positive take-away is that the authors above *have* strengthened operant moral agency compared to its original conception in the automaticity challenge, which has somewhat alleviated the issue, carving out more room for operant agency. But even more importantly, the discussion above, albeit not fully capitalising on it, reveals some indications for what could be a further expansion of moral agency, focused on the development of automaticity, to which we will turn next. With that, I am highly sympathetic to the authors discussed here, and rather than disputing their accounts, I think that all of them can be subsumed *within* a tripartite model. As such, these contributions can be fully appreciated as accounts of moderative agency, such that they do not fail to achieve some warrant for moral agency in its entirety by itself as a monolithic notion of agency, but instead they can be seen to successfully achieve one distinct mode of agency that has its specific strengths and role among three modes.

§4: Developmental agency

In this final section, the notion of developmental agency is further elaborated. After a brief discussion of how character development is largely overlooked in the agency literature, I will argue why development is a mode of agency in its own right, why it should even be centralised due to the advantages that are particular to it, followed by exploring a range of possible strategies in the light of empirical research in order to substantiate the notion.

4A: Standard account of character development

Character development is a topic that has received relatively little attention in the debate on moral agency in the light of automaticity. One of the two ways in which it

is addressed, is, as discussed above (§3), as a side effect of deliberative and moderative modes. However, conceived as a side effect, development is only sparsely explored in itself, as authors like Musschenga, Kennett and Fine, Sauer, and even Holroyd and Kelly are more concerned with the operant, action-focused approach to automaticity. Moreover, as also argued above, the developmental side effects are unlikely to warrant moral agency.

The other way in which development is discussed in the literature is in relation to biological and social formative sources. As elaborated elsewhere (my doctoral thesis, essay 1), this is the main form of acquisition of automaticity that is elaborated. For example, two-systems theories, which underlie most automaticity modelling, understands 'system 1' (the automatic processing system) as acquired through (mostly implicit) social learning in addition to being shaped through innate specification. (Evans & Stanovich, 2013; Stanovich & West, 2000, p. 659) The social intuitionist model similarly understands intuitions as evolutionarily prepared states, certain of which emerge through cultural formation. (Haidt, 2001, pp. 826-828) Others follow this core of biological and social factors, only differing concerning whether automaticity is quite unrestrictedly acquired or rather emerges within a pre-set range of options, and what sociocultural learning process is emphasised. For example, Lapsley and Narvaez stress personality formation through social embodiment. (Lapsley & Narvaez, 2004, p. 206) And Prinz highlights the role of emotional training by parents. (Prinz, 2007, pp. 268-270)

Moreover, biosocial development even functions as a premise of the automaticity challenge, such that the challenge is construed as an operational challenge, taking for granted that agency is an operational matter concerned with determining some behaviour, not concerned with character development, since the latter is merely a biosocial matter. This construal is unsurprising, however, since standard conceptions of automaticity understand the concept in dichotomy to control, such that a cognitive process is either automatic, or controlled. In turn, this conception relies on an understanding of 'control' as an operant, conscious deliberative process that fairly directly relates to determining some behaviour. Without going further into this, it should be clear that character development has, due to various conceptual causes, gotten little attention beyond biological and social factors.

The key problem of biosocial development is that it is not an agentive process, which especially matters when it causes morally problematic outcomes. In their

development, agents are passive *as agent*; they are not actively involved through one's reasoning to determine how one's automaticity is configured, or what kind of characterological states one develops. In the best case, one may be fortunate enough to pick up a flawless virtuous character from which one automatically exhibits moral saintly behaviours. Nevertheless, even then, this still does not resolve the frailty of agency that the automaticity challenge gets at, since that person was not involved in the development as agent. Surely, were people mostly raised to acquire virtuous characters, such lack of agency would not concern us as much. However, the automaticity challenge is more significant in the case of automaticity driving morally problematic behaviours. And with evolutionary and cultural factors being the main formative factors, it is unsurprising that in reality people passively acquire a sorely flawed automaticity, which subsequently all too often drives morally problematic behaviours. As such, that the automaticity challenge *is*, in fact, very troublesome, since agency becomes especially consequential when its frailty leads to morally problematic outcomes. And while it may be conceptually intelligible that development was excluded from discussion on agency, there are no conceptual grounds that this is inevitably so. Moreover, since operant agency modes seem to fail at fully warranting agency over automaticity, the development of automaticity may be worthwhile to include more into the discussion.

4B: Development as agency

A tripartite model understands character development as a mode of agency in its own right, not merely a side effect of deliberative and moderative agency modes, nor merely as a biosocial premise. Better yet, developmental agency may actually be emphasised as the *main* mode of agency; meaning it is the most important method of securing agentive behaviour. I will now argue what makes one's character development agentive, and why subsequent behaviour driven by agentively developed character is, in turn, also agentive.

Agency, as discussed earlier, is understood in terms of an agent's conscious, rational, deliberative reasoning, since through this capacity one can endorse certain beliefs and values for the right reasons, which makes them intelligible, and which

enables the agent to account for them. In the case of deliberative and moderative agency, reasoning determines some behaviour, and through that causal connection the behaviour is deemed agentive, as it comes forth from one's reasoning. The same can be said, I argue, for characterological behaviour, except that it is a two-step process. Firstly, an agent's reasoning determines some developmental behaviour, whereby that reasoning is not merely present in those behaviours, but also in the character that is developed due to those behaviours. Secondly, an agent's character subsequently drives some moral behaviour, in which the reasoning is present due to its presence in one's character.

Developmental agency, as such, is a form of *agency by proxy*; meaning that one's agentive reasoning is present (causally active) even while it is so in an in-direct manner ('diachronically', if you will). To some, this may sound unproblematic, since some moral behaviour being agentive due to its causal origin seems similar in both a direct or developmental route. One can think here of Clark's parity principle, which holds that a process' functional characteristics is what determine its nature; whether a process takes place 'in the head' or, for example, in a computer, if the function is the same, both are 'mental'. (Clark, 2010) Just so, the functional characteristics of reasoning determining behaviour directly or by proxy are the same.

For those who are more critical on this point it may be worth somewhat further elaborating the type of connection that agentive character development and character-driven behaviour have. As Ellen Fridland points out in her meticulous discussion, a mere causal relation between an agent's automatic cognitive processes and their personal-level, conscious, deliberative, intentional states (an agent's goals and beliefs and such) is insufficient. Rather, a certain type of cognitive penetrability is required, "a particular kind of connection: a connection where there is a meaningful or semantic interaction between content and processing." (Fridland, 2017, p. 11) Building on Pylshyn's 'semantic coherence criterion' and Macpherson's 'intelligibility criterion' as adequate definitions of cognitive penetrability, Fridland argues that the processing of a system (in our case moral automaticity or moral character) has to operate in such a way that its outputs are impacted by the agent's intentional states in a semantically coherent way. (Macpherson, 2012; Pylyshyn, 1999, p. 343) The semantic coherence criterion ensures that the automatic system's operation is sensitive to an agent's intentional states in virtue of the latter's actual semantic content, such that there is a logical relation between them. As example,

Fridland draws a scenario of a person's belief that the person ringing one's doorbell will be a cookie-selling girl scout, such that the content of that belief state influences the visual system's processing of phenomenal properties that are presented to it; for example by being more likely to produce green qualia over other colours (based on the typical green uniforms). Translating Fridland's focus on visual perception to moral automaticity, one can think of Amir's the content of implicit bias that women are less intelligent, with the subsequent expectation that some woman's contribution will be flawed, influences his automatic processing of what she says, for example rendering him more likely to pick up faults, pay less attention in general, or notices room for interruption more quickly. Such attentional, interpretative, and behavioural automatic responses do not merely have just any causal relation to one's character, but there is semantically coherent cognitive penetrability.

Now, the development of automatic processes involves a similar connection, since this is not a mere result of brute repetition, but rather it relates in a meaningful way to the content of the agent's reasoning that determines developmental experiences. On this point, Fridland argues that such diachronic connection is similar to the above synchronic one, as behaviours or states that are automatised bear a semantic relation to the content of an agent's intentional states. As such, automatic processes and their development can be, as she calls it, 'in the space of reasons. (Fridland, 2017, p. 15)

How exactly automaticity is within the space of reasons takes us to a further discussion on propositionality. Fridland takes a middle ground here, understanding them as 'proto-intelligent' rather than fully intelligent, because it is not obvious to her that automatic processes bear propositional thought. However, beyond her account of cognitive penetrability and development, there are many other, further reasons that support understanding automatic states as propositional rather than merely associative or some intermediate. For example, while imperfect, implicit states often exhibit evidence-sensitivity (Brownstein & Madva, 2012a; Mandelbaum, 2013, p. 206), while 'typical' propositional states (i.e. beliefs) often lack doing so (Mandelbaum, 2013, p. 209; 2014, p. 67). As such, evidence-sensitivity is at best a difference of degree. Moreover, development can thus also run accordingly, with at best a degree difference in the response-time, as implicit states often shape over an extended period of time rather than update directly. (Brownstein & Madva, 2012b) Going further into

the metaphysics of implicit states is outside the scope of this essay.[22] Nevertheless, regardless whether ultimately classified as intermediate or fully propositional states, Fridland's account offers an elaborate argument for understanding automaticity as agentive when deliberatively developed; through an agentive developmental process, the intentional or semantic content of an agent's conscious deliberation is ultimately causal to the characterological behaviour.[23]

4C: Advantages of developmental agency

Developmental agency is not an operant mode, focused on determining some particular behaviour within a particular moral situation, but rather concerned with the reconfiguring character of the agent. As such, developmental agency need not be exhibited within or just prior to a moral situation, but can instead be exhibited at any time. This opens up for a vast space of exhibiting moral agency, virtually including every moment of one's entire life. Recalling the violinist comparison, one spends most of one's time and effort off stage, training and otherwise learning. Before discussing possible strategies for doing this, I will now argue for three important advantages that developmental agency has in comparison to deliberative and moderative modes; multifariousness, truth-tracking, and cognitive efficiency.

The first advantage of developmental agency is that it opens up a vast array of new ways of exhibiting agency. Since developmental agency need not (but can) be exhibited within moral situations, the available agentive strategies are not limited to deliberative action, cognitive intervention, cognitive preparation, and environmental regulation. Instead, developmental agency can be exhibited offstage, and thus agentive strategies can be expanded far beyond, as there are many different ways of

[22] Also see (Huebner, 2009). For another intermediate account, see (Railton, 2014) on a representationalist account of 'learned information structures' that are 'information-value-sensitive', and also (Schwitzgebel, 2010, 2013). For opposing associative accounts, see (Gendler, 2008) and (Levy, 2014).

[23] Also interesting to note here, is the presence of related thought already in Aristotle's work, who struggled to balance behaviour from character while including some notion of prohairesis (conscious choice) with bouleusis (deliberation), such that the behaviour is 'free' and 'controlled'. (Aristotle NE 111b4-1113a14, 1139a22-b6) Cooper proposes that Aristotle's characterological behaviour should be understood as including 'hypothetical deliberation'; the behaviour is rational, as if it is deliberative, just not explicitly so. (Cooper & Cooper, 1975) Similarly, Sherman argues that prohairesis need not be conceived as behaviour resulting from explicit prior deliberation, but merely rationally and in some way controlled. (Sherman, 1989, p. 82) Lastly, following Aristotle, McDowell argues that 'conceptual capacities' become part of one's cognitive and behavioural habits, such that these capacities become 'second nature'. (McDowell, 1996, p. 84)

developing one's character. For one, this empowers every moral agent, as the action space one has becomes larger, such that one can more easily find ways to take agency over one's moral life. Secondly, agentive strategies can be fashioned that are specifically beneficial to some agents, depending on their agentive opportunities (both rational capacities and environmental conditions), which can empower disadvantaged people. In the next and final part of this section I will explore a range of possible developmental strategies.

A second advantage of developmental agency, is that moral reasoning can be more truth-tracking when exhibited offstage, outside of moral situations. When one engages in deliberation in what we can call *detached moral reasoning*, this is not as likely to be misguided due to the vulnerabilities that the automaticity literature points out concerning operational reasoning in the 'heat of the moment'.

For one, versus the speed thesis, there is no similar competition between cognitive processes to reach a conclusion as quick as possible, as there is no time pressure. Rather, there is ample time to reflect upon one's beliefs and values and such. This allows agents to more carefully and systematically engage in deliberation, which may avert cognitive distortions and errors as described in the partisan cognition thesis. Moreover, proposals for distinct reasoning strategies by Musschenga, Blasi, and Hogarth to make reasoning more accurate and rational can be fully appreciated in this setting. (Blasi, 2009; Hogarth, 2001; Musschenga, 2011) Furthermore, one can reason engage in developmental agency at any time, such that it need not be in a situation that may elicit irrelevant factors that influence one's reasoning (as goes the situational thesis).

Besides time, another possible beneficial feature for truth-tracking is that, detached from a specific moral situation, the agent may be less committed to a specific outcome, such that distorting emotions and other partisan cognitions may not be as present, allowing for a more truthful evaluation of the matter. One indication for this may be that people's reasoning about impersonal or non-moral issues is found to be less emotionally engaging, which in turn allows people to better appreciate relevant factors. (Berthoz et al., 2002; Heekeren et al., 2003; Phan et al., 2002; Singer et al., 2006) Another indication may cognitive neuroscientific studies on moral judgement-formation regarding moral dilemmas such as trolley problems, which found different processing of 'footbridge cases' (pushing someone in-front of a

trolley) compared to 'switch cases' (pulling a lever). As Joshua Greene and colleagues draw from their findings, the more up-close and directly involved someone is, fast emotional processes dominate one's behaviour more. (Greene et al., 2001) And thirdly, another indication for the superiority of detached reasoning could be taken from Hogarth's studies on people's decision-making quality in everyday lives. As Hogarth emphasises, the informational poverty concerning the appropriateness and effectiveness of one's decisions is the greatest barriers to acquiring valid intuitions, as people typically get no or little feedback for improvement (although this does not bar their confidence). (Hogarth, 2003, 2006)

While it is mere conjecture at this point, and more specific research is required, there seems to be evidence supporting the idea that detached moral reasoning is more truth-tracking than operant moral reasoning. Especially since, as I will soon discuss, developmental agency can be exhibited through, for example, hypothetical situations, in which the stakes are not as high for an agent as the outcome will not affect anyone's life, it seems likely that one will be better able to truthfully reflect on oneself. As such, I would follow Brosnan in taking deliberative reflection as a process does not merely work *with* one's moral beliefs, but rather *on* them, restoring the epistemic credentials that are otherwise undermined (due to social learning, in addition to Brosnan's focus on evolutionary origins). (Brosnan, 2011) As such, developmental agency is, for a large part, the reconfiguring of one's flawed or inconsistent beliefs and values. This requires deliberative reasoning processes that are truth-tracking, not like lawyers committed to some outcome, uncommitted judges, which is more likely to be the case with detached reasoning than with operational reasoning. As Haidt and Bjorklund themselves remark, "lawyers can be very reasonable when they are off duty, and human minds can be too." (Haidt & Bjorklund, 2008, p. 191) Developmental agency allows for such reasonableness, as reasoning does not as much compete with automatic processes to guide behaviour, but rather works on one's the reconfiguration of them.

A third, and major advantage for developmental agency is that it is not nearly as plagued by the resource and frequency theses. One crucial weakness of operant reasoning is that it is dependent on cognitive resource that are scarce, such that one can only infrequently exhibit such in-action deliberation as it depletes one's capacity. While detached reasoning is dependent on and thus depletes the same resources, it is

not bound by a specific temporal confinement. For one, there is no time pressure to engage in deliberation, since the aim is not some moral behaviour in a moral situation that demands engagement now, but rather the development of one's character. Secondly, agents thereby have the time to regenerate their cognitive resources, and choose moments when they have ample resources available to engage in developmental reasoning. As such, agents can exhibit reasoning for developmental purposes whenever they have the resources available.

The inverse of this claim is that moral behaviour is driven by automatic processes, which take less cognitive resources. If one has a developed moral character, such that one can rely on one's automatic processing, this frees up their cognitive resources for usage elsewhere, Various others have pointed out the cognitive efficiency benefit of automatic processing, often emphasising utilising the resources for processing future plans. (Sauer, 2012, p. 260; Wood et al., 2002, p. 1295) While I agree with this, the tripartite model especially emphasises availing these resources in developmental activity, since it takes this to be the key mode of agency (and since development of character, after which one can rely more on one's characterological behaviour, continues to free up cognitive resources, securing the benefit).

4D: Strategies for character development

Now that we have an idea of why agentive character development may be a crucial mode of agency, the final part of this paper will concern drawing various ways in which character development may be conceived and seeing how such strategies may be empirically substantiated in order to argue for the empirical realism and concrete applicability of the proposal.

As said, character development may occur through one's deliberative and moderative behaviour, as this can feed back into one's cognitive configuration. However, due to all the challenges with exhibiting morally desirable behaviour while having conflicting automatic cognitive structures, this is taken to not suffice as a warrant for even the agency over one's behaviour, let alone warranting the development of one's character. Instead, the tripartite model understands character

development to mainly be achieved offstage, in training situations outside of moral situations. The strategies explored below are a mere first step at cataloguing all the many ways in which agents can engage in developing their own character. By no means this catalogue is exhaustive, and by no means is the empirical substantiation conclusive. Rather, the aim here is to argue for the empirical plausibility of an ample variety of strategies that enable agents to agentively develop their character. Having defended this much, future research on moral behaviour will hopefully focus more on character development and provide us with sturdier empirical backing of a wider variety of strategies, possibly even revealing comparative efficiency of various strategies.[24]

Developmental strategies can be very roughly distinguished as *cognitive strategies* and *experiential strategies*, although these two categories often overlap and interact. The factor that differentiates the two is the focus on theoretical, cognitive learning or experience-based, affective learning, respectively.[25]

4E: Cognitive developmental strategies

One category of developmental strategies can be labelled *cognitive strategies*. There are numerous cognitive strategies, of which I will only discuss several here. An important first step of these strategies is metacognition, the awareness and understanding of cognitive processing, as a start for reflection and training of these. Since biases and other automatic processes operate implicitly, creating awareness is critical for combating them. (Wilson & Brekke, 1994) Already this step is demanding,

[24] One critique here may be that agents who do not already have the right moral character will thereby be unable to select the morally right development for themselves. While this may sometimes be the case, I do not think this is a great worry, since the main problem of automaticity is not that people do have the right *explicit* values and beliefs, but rather that their *implicit* ones are not in line with these. As such, I presume that most people will at least have plenty explicitly endorsed moral attitudes that they can imbue into their character, as a start.

Nonetheless, there are two further interesting question to this. Firstly, concerning the matters where someone may not have the morally right explicit attitudes, it is a question how best to convince such a person, since rational conversation is often thwarted by automaticity and cognitive frailty (and non-rational persuasion may violate someone's agency). Secondly, for any time of developmental effort, motivation to undertake this effortful project is required, but how to stimulate someone's motivation is an equally difficult matter. Both of these issues need more research in order to evaluate the best strategies.

[25] The angle of the strategies explored here may pose an interesting topic to compare with Aristotelian or other virtue ethics. In contrast, Aristotle typically understands moral character development to be guided by another, already morally developed, agent. Secondly, Aristotle typically understands moral character development as coming into being mainly as the result of performing behaviours, copying the behaviours of one's moral guide.

In contrast, the present suggestion may be seen as achieving a similar goal (developing the right moral character), but doing so autonomously, as agent. Secondly, development is seen to be achieved not merely through practice, but through various strategies.

however, since people typically do not easily acknowledge their moral flaws. For example, White people have been found to find it difficult to realise or admit their implicit prejudicial racial beliefs and attitudes, even when confronted with these, because this conflicts with their explicit egalitarian self-image. (Gaertner & Dovidio, 1986)

There are many ways though in which agents may seek awareness. Firstly, one can learn about psychology and automaticity to acquire some more general knowledge on implicit processes. And, more personalised, one can test one's own implicit attitudes, for example through Harvard's freely available online Implicit Association Test. As others have explored, labelling the attitudes one discovers through taking such tests helps to create awareness even more. (Devine et al., 2012) As such, people do not only learn what automaticity is and that they are susceptible to it themselves, but moreover they acquire knowledge about which implicit attitudes they hold that are incongruent with their explicitly endorsed values. Or, not relying on such tools, agents can engage in self-reflection to test oneself; pondering their behaviours, attitudes, beliefs, values, goals, and such, in order to discover one's own moral character, including seeking out deficiencies in one's character (e.g. implicit attitudes that are incongruent with one's explicit values). One of the most sophisticated examples of this can be found in the writings of Roman emperor Marcus Aurelius, who, as a proper Stoic, constantly wrote down daily events and reflected on his behaviours, feelings, and thoughts, unravelling what these all signified. (Hutcheson & Moor, 2008) There are many other such reflective methodologies that are quite freely available to anyone, for example Socratic dialogues, which can be done alone or in publicly organised groups.[26] (Kessels et al., 2002) It is also here, as I argued earlier, that the work on methodologies for moral reasoning by authors such as Musschenga, Blasi, and Hogarth can be best appreciated, as developmental strategies, rather than operational ones. (Blasi, 2009; Hogarth, 2001; Musschenga, 2011) Additionally, one can think of Jennifer Saul's proposal here to challenge deficient generics (e.g. 'Muslims are terrorist') by spelling out the evidence for the claim and reflecting on what that evidence really warrants. (Saul, 2017) Alternatively, one can engage in courses specifically designed to learn about the origin and function of biases. (Sevo & Chubin, 2010) Tying all of this together, there are numerous ways in

[26] In The Netherlands, for example, a network called Socratisch Café Nederland regularly organises public events throughout the country during which participants are guided through various methodologies for moral self-reflection.

which agents can acquire so-called 'bias literacy', knowledge about the origin of biases, how they function, and become aware of one's own biases, through theoretical learning, reflecting on one's behaviour and character, testing it, seeking out deficiencies, analysis, spelling out reasons and arguments, and many other ways.

Bias literacy has two further effects beyond awareness. For one, it helps to increase one's internal motivation to develop one's character and behaviour. (Carnes et al., 2012) For example, a workshop on gender bias for university staff showed that participants become more motivated to be involved with promoting gender equity. (Carnes et al., 2015) Secondly, it also increases 'self-efficacy', knowledge about how to combat one's own biases and biased behaviour, which has been found to be crucial to development. (Bandura, 1991, 1997; Prochaska et al., 1993; Sevo & Chubin, 2010) With that, agents are empowered to organise developmental activities for themselves, knowing how to do that, knowing what to do that about (e.g. concerning certain ethnic rather than sexist biases), and being motivated to do so.

Moreover, reflection and learning is not merely for creating awareness, motivation, and self-efficacy, but can itself already shape one's character. New information, especially when this prompts reinterpretation of the knowledge or beliefs one currently holds, can change one's biases. (Mann & Ferguson, 2015) Such new information can be acquired through general learning (e.g. studying certain sciences), but also through reflection on and testing of one's attitudes. Testing one's biases through 'decoupling' from one's commitments and seeking out evidence for alternative beliefs (or 'considering the opposite'), for example, reduces biases. (Bishop & Trout, 2004; Larrick, 2004; Lord et al., 1984) Moreover, together with awareness, training of one's cognitive capacities and analytical reflection of one's beliefs and values has been found to be an important factor in shaping biases. (Croskerry, 2014; Croskerry et al., 2013; Jenicek, 2010; Mamede et al., 2010; Vohs et al., 2007; Whaley & Geller, 2007) And besides reflection, also expression is important. As a study on controversial political opinions found, reporting one's newly deliberated and reflected explicit opinions (e.g. through writing them down, or in dialogue with someone else) may shape the implicit attitudes one had about the issue over time. (Galdi et al., 2008) Lastly, cognitive factors such as new knowledge, reflection, and changes in explicit attitudes, may play an important role in the hereafter-discussed experiential developmental strategies. Students enrolled in a seminar on prejudice and conflict showed significantly reduced racial stereotyping

and prejudice. (Rudman et al., 2001) Besides affective processes, such as familiarity through exposure to a Black professor and friendships with Black students, which other studies have found effective, the seminar also invoked cognitive processes. The students learned about intergroup conflict, engaged in discussion, and documented instances of bias (including their own), in order to increase awareness of and motivation to counteract biases. While affective, indirect processes most strongly related to changes in implicit attitudes, and cognitive, direct processes to explicit attitudes, the authors concluded from their covariance that the two process types work and develop in concert.

This points to a further way of cognitive development; world knowledge and skill training. Beyond learning specifically about moral psychology, or one's own moral character, people can also acquire knowledge that is often necessary in order to notice and understand the meaning of moral occurrences through learning about social, political, and economic histories and theories. For example, Amir's moral self-education could have included reading up on the history of women's rights and labour participation, which provides a background for him to see the interactional dynamics against. In turn, this may make Amir keener to perceive instances of sexism, comprehend better what exactly is morally problematic about it, and be more motivated to engage with it. And in addition to knowledge, various skills may be important to cultivate. For example, training logical reasoning, tracing cause-effect relations better and recognising the validity of conclusions, through mathematical exercises, debate clubs, or puzzles. Observational skills may be improved through drawing or photography classes, making one more sensitive to noticing important but subtle situational cues. Through reading and writing stories one can develop better language skills that can play a part in making sense of events and experiences. Emotional intelligence may be trained through literature as well, as through other practice, and often plays a role in evaluating the moral valence of some happening.

As such, there are many skills, and there is much knowledge, that can be acquired in order to improve one's moral character, enriching the relation an agent has to their own character and characterological processes. On this picture, developing one's moral character is not merely changing, say, the beliefs that underlie one's implicit biases, but also all other states and processes that are relevant to moral cognition, such as perceptual skills, logical reasoning abilities, and specific knowledge. Tying up all of these cognitive strategies together, it is crucial that agents

develop an agentive attitude towards their own character development. As a fascinating study by Stahl and colleagues shows, a central feature about people who are more inclined to hold attitudes and beliefs that are supported by logic and evidence, is that they moralise morality, meaning that they see evaluating one's attitudes and beliefs by reasoning and data as a moral virtue. (Ståhl et al., 2016) Similarly, for developmental agency to take off, agents have to cultivate an attitude to rationally develop one's own character, which moves one to reflect, test, seek out deficiencies, learn other relevant knowledge and skills, and as such create awareness of one's character configuration and reconfigure one's character.

4F: Experiential developmental strategies

Another category of developmental strategies is *experiential strategies*. The main developmental mechanism of strategies that belong to this category is repeated exposure to certain experiences that shape one's character through mostly affective processes. As argued right above, on the broad notion of moral character employed here, developing one's moral character also involves (and possibly most crucially so) shaping one's emotional attitudes.

The most straightforward types of experiential strategies involve selecting or engineering certain situations to place oneself in that have a developmental influence. Different from moderative agency, no 'real' moral situations that occur in one's life are needed in which one has to exhibit certain moral behaviour, as the purpose is not operant behavioural. Rather, since the purpose is purely developmental, one can use *artificial situations* that merely serve training purposes; 'artificial' in that they are not 'naturally' part of one's life but sought out for development. This makes a vast space of developmental experiences available. Moreover, as argued earlier, since there are no actual moral issues at stake in such training situations, one may be more (emotionally) open to new experiences and alternative perspectives.

Many sorts of contact and interactions, such as meeting, talking, and cooperating with people from a certain group have been found to alter implicit attitudes. Meta-analysis of intergroup contact has shown the effectiveness of such

approaches, which can be extended to various sorts of groups (e.g. ethnicity, sexual orientation, etc.) and generalised over the entire group (not just the individuals one interacted with). (Pettigrew, 1998; Pettigrew & Tropp, 2006) A large body of research has shown, for example, that contact between ethnic groups can positively change implicit attitudes, decrease prejudice, and reduce anxiety. However, not just any intergroup contact has such positive effects, but this depends on conditions such as the context and the affective quality of the contact. (Page-Gould et al., 2008; Pettigrew & Tropp, 2013; Tropp & Page-Gould, 2014) Especially feeling connected decreases bias. (Pettigrew & Tropp, 2008; Voci & Hewstone, 2003) As such, unsurprisingly, intergroup friendship is seen as the most effective way to develop positive attitudes. (Binder et al., 2009; Brown & Hewstone, 2005; Davies et al., 2011)

With that in mind, actively seeking out opportunities for positive interactions with members group a certain group is a robust method of character development. As mentioned above, one way in which this can be done is through participating in a course taught by an African American professor, which enhances familiarity through exposure to the professor, and enhances trust and safety through friendships with Black students. (Rudman et al., 2001) Working together on a project with someone from a different or stigmatised group has also been found to reduce implicit biases. (Blincoe & Harris, 2009) And one could join a sports team with people from another ethnic group to develop more positive attitudes. (Brown et al., 2003) Beyond mere interaction, seeking contact with counter-stereotypical examples of a group can function as a distinct strategy when concerning to certain biases, as has been found in relation to homosexual people and women. (Dasgupta & Asgari, 2004; Dasgupta & Rivera, 2006)

Conversations with or about people from a stigmatised group can significantly reduce prejudice as well. For example, dialogues about transgender people, narrating their personal experiences and engaging in perspective taking (relating to someone else's experience with vulnerability), have been found to positively change people's attitudes towards transgender people and their rights. (Broockman & Kalla, 2016; Paluck, 2016) Perspective taking can also be done through sharing related experiences. And adopting the perspective of a member of a certain group, relating to it on a personal level, has been found to create personal closeness, which ameliorates stereotypic attitudes. (Galinsky & Moskowitz, 2000)

Individuation, detailing specific information about someone as a person, rather than merely a group-member, has also been found to challenge group-based evaluations. (Brewer, 1988; Fiske & Neuberg, 1990) As such, exposure can be sought in a variety of ways, through education, meeting spaces, and many other options, which is a solid method of challenging the validity of biases, making positive exemplars more salient, and developing positive feelings.

However, as argued earlier concerning moderative agency (§3), oftentimes people only have limited opportunity to seek out or create environments in which one is exposed to desirable developmental stimuli. This is even the case, I can conceive, for artificial situations, since not everyone is able to, for example, sign up to some university course, join a diverse sports team, or regularly make time to join some volunteering project. Nevertheless, besides such, say *tangible experiences*, one can also engage in character development through *derivative experiences*. Since operational behaviour goals are not the purpose of developmental agency, agents need not be in actual, tangible situations in which such effects can be achieved. Rather, since only experience through exposure matters, for the purpose of development, an even vaster space of possibilities is opened up when exploring derivative experiences such as other people's experiences, literature, and even one's own imagination.

For one, rather than direct personal contact, people can devise indirect strategies for 'extended contact'. A growing body of research on extended contact shows that even when one lacks interaction opportunities oneself, knowing that others from one's group have positive relations with people from another group changes one's attitudes towards members of that group. (Dovidio et al., 2011; Gómez et al., 2011; Turner et al., 2008; Wright et al., 1997) Moreover, when direct personal contact is only limitedly possible, indirect contact has been found to be the most effective strategy. (Christ et al., 2010; Eller et al., 2012) For example, thinking about one's White friends who have Non-White friends has been found to reduce prejudice to the latter group, positively shifting participants' attitudes and expectations about future contact. This positive enhancement of intergroup attitudes and prejudice especially works through reduction of anxiety. (Mazziotta et al., 2011; Turner et al., 2008; Wright et al., 1997)

Furthermore, derivative experiences can also help developing skills. For example, surgeons have been found to improve their decision-making abilities

through reading stories of more experienced surgeons concerning specific cases they handled. (Abernathy & Hamm, 1995) Besides that not just emotional dispositions, but also skills can be shaped through derivative experiences, this shows that not only orated stories, but also written stories can have developmental value. As such, we can invoke Martha Nussbaum's work on the importance of reading literature to develop one's character, through educating empathy, sensitivities, motivations, and many other traits. (Nussbaum, 1992) Besides literary experiences, virtual experiences through gaming are another type strategy that agents can engage in. Various studies have shown that playing a video game, for example in which one has to navigate a predominantly White university from the perspective of a Black student, can reduce prejudice. (Shih et al., 2013; Todd et al., 2011; Todd & Galinsky, 2014) Cinematic experiences have developmental effects too, as inducing empathy toward an Asian film character was found to decrease implicit biases towards Asian people. (Shih et al., 2013) And one can expose oneself to counter-stereotypical images by hanging photos on the wall of one's office or house, not as much to trigger representations in order to exhibit certain behaviour (in moderative agency), but for developmental purposes. Repeated exposure of photos of admired Black people (e.g. Martin Luther King and Colin Powell) has been found to have long-term effects on one's implicit attitudes. (Dasgupta, 2013; Dasgupta & Greenwald, 2001) Altogether, an important strategy for developmental agency is changing one's media diet, exposing oneself to experiences that representing people in a way that is more in line with the explicit moral beliefs and values one holds, in order to cultivate this in one's character.

Lastly, besides tangible and derivative experiences, one can even engage in character development through *imaginative experiences*. Without relying on other sources that may not be available, an agent can simply use one's own imagination in order to have experiences in one's own mind.

Blair and colleagues tested the effects of mental imagery, e.g. drawing a counter-stereotypical image of a 'powerful woman' in one's mind, which they found had a moderative affect on the expression of stereotypes. (Blair et al., 2001) Moreover, this is the case when one draws from famous examples (e.g. Obama), non-famous ones (e.g. a friend), and abstract, fictional ones. While the study did not address long-term consequences, the authors do speculate on the likeliness of this. Repeated episodes of such imagination increase the weight of connected associations

resulting in stable changes of the form of a representation or stereotype.[27] Gawronski and colleagues have even tested the most effective methods of developing stereotypes, being 'affirmative training', rather than 'negation training' (negation can strengthen stereotypes). (Gawronski et al., 2008) Relating to the gender stereotypes of strong men and weak women, participants responded with an affirmative "yes" to stereotype-inconsistent trait, e.g. 'Angela' plus 'mighty', 'Gloria' plus 'assertive', 'Jason' plus 'dainty', and 'Tony' plus 'tender'. And relating to racial stereotypes, traits were trained through affirmation of stereotype-inconsistent positive or negative traits, e.g. a Black face plus 'intelligent' or 'friendly', and a White face plus 'poor' or 'violent'. The result of such training was not merely the reduction of automatic stereotype activation, but moreover the automatic evaluations of the targets. The reason for these reductions is, the authors argue, a change in the semantic associations that form the content of one's stereotypes, since the training affected not merely activation, but also evaluation (and may affect action-responses).

As such, developmental agency can even be exhibited, say, home alone. Moreover, such strategies allow for structured, repeated exposure as a form of effortful training. As has been shown, such extended practice is an important way of overcoming biases. (Blair, 2002) One can think back, here, about Aristotle's thoughts on habituation through extended exercise. Forming associations between certain beliefs, or between representations and emotional responses, and many other associations, might thus, besides being shaped through cognitive strategies such as reflective deliberation, as well be developed through routine habituation.

4G: Developmental agency in practice

To conclude, innumerable cognitive and experiential strategies may be envisioned, interpersonal or intrapersonal, theoretical or practical, tangible or derivative or even imaginary, through reading or dialogue. *That* agents can actively and intentionally engage in such strategies and thus develop their moral character, seems to be very plausible on the vast variety of evidence pointing in this direction. *How best* to engage in developing one's character will require much more research, comparing various

[27] For the 'connectionist model', see (Smith, 1998)

strategies, and, crucially, relating these strategies to individual agents' capacities (some may be more imaginative than reflective), preferences (some may like films over books), and environmental opportunity (some may have less time and funds to enrol in courses). As such, ways can be found for any agent to develop one's own character. Some have already begun to overview the empirical literature on character development, but if we want to take moral agency seriously, many more resources will have to be directed to this end. (Devine et al., 2012; Godsil et al., 2014)

From this perspective, it is worth looking back at some of the key factors of the automaticity challenge from the developmental perspective we now have. A perfect illustration for this are Haidt and colleagues' infamous studies on moral intuitions and confabulation, most notably in the case of a hypothetical scenario concerning incest, so construed as to have no harms that could justify moral condemnation. Driven by 'moral disgust', people condemn the fictional characters. While pressed to justify their judgement, their moral reasoning functions 'like a lawyer', providing possible arguments. When confronted with the fact that none of these justify the judgement, the participants do not revise their judgement, but rather maintain it while 'morally dumbfounded', from which the authors conclude that reasoning is ineffective in determining the moral behaviour, merely following as post hoc confabulation for the determinant intuition. (Haidt, 2001; Haidt et al., 2000)

Now, this finding only disqualifies the direct operational role of moral reasoning within a moral situation, and thus deliberative or moderative agency. It says nothing about developmental agency. After the affair above, the agent can engage in reflection on one's reasons and consider the lack of evidence, which may be more truth-tracking due to it being detached, allowing for more time, more cognitive resources, and less reasoning frailties, and with awareness of one's automatic processing frailties. Furthermore, the agent can engage in training other responses and exposing oneself to (albeit derivative or fictional) people with other sexual norms. Through such efforts, one may well become aware of how biased one's response is, that it is groundless, cultivate oneself to not feel disgust, and while maybe 'not fancying it' personally, have no moralising intuitive, or characterological, response against it.

Although these reasoning processes may not achieve behavioural change directly, it can very well reshape one's character over a longer period of time, such

that one comes to have different intuitions about the topic in question it should, and thereby form different judgements. As such, it should be very clear by now that in this way the agent most definitely exhibits a mode of moral agency; developmental agency.

Moral Responsibility, Automaticity, and Character

Exploring a Developmentalist Account

Abstract: *Recent empirical research in the psychological sciences has been taken to show that most of people's moral behaviour is driven by 'automaticity', automatic, unconscious, affective processes, rather than moral reasoning. Traditional volitionist and attributionist theories of individual moral responsibility each have significant limitations accounting for this prominent category of behaviour in a satisfactory manner. I explore a 'developmentalist' view as viable alternative account, grounded in the 'tripartite model of moral agency', and focussing on the opportunity for an agent's self-development of one's moral character over time. I argue that developmentalism better appreciates the empirical data while providing a more nuanced evaluation of responsibility.*

Introduction

"A fucked-up childhood, is why the way I am. It's got me in the state where I don't give a, damn.
Hmm, somebody help me? But nah, they don't hear me though. I guess I'll be another victim of the ghetto.

Ain't no escapin', 'cause I'm way too young. Pops is dealin', and on top of that got moms sprung.
Scheamin' off the top, pops never figured. That he'd go down by the hands of another nigga.
Now my pops is gone and that ain't no good. Got to follow in the footsteps of the homies from the hood.

And where's the role model? Niggas puttin' brew in my fuckin' baby bottle.
Damn! And through all the motherfuckin' pain. They done drove my moms insane.
So, I guess I gotta do work, so I ain't finished. I grow up to be a 'streiht' up menace. Gyeah."

- Mc Eiht[28]

Imagine a person, say, an employer, who values gender equality, yet intuitively disqualifies female candidates in a job interview. Or someone else, a judge, who explicitly disapproves of racism, but nevertheless more readily perceives Black people as culpable of alleged crimes. And another, a teacher, who avows opposing classism, while tending to unconsciously evaluate the boys with a strong working-class accent as less talented. Cases such as these are not difficult to imagine, since they routinely occur in our everyday lives. What is more difficult, however, is how to theorise about these agents' moral responsibility in these cases.

For one, such automatic processes typically operate involuntarily, and even resist conscious reasoning. Secondly, as Mc Eiht poetically brings to our attention, the development of one's character is decisively shaped by one's life circumstances, which thus, in turn, their moral behaviour. As such, the difficulty for moral responsibility plays out at two levels; the involuntary expression of automatic processes as well as the inadvertent acquisition of those automatic processes.

[28] Recounting the life of Kaydee 'Caine' Lawson, a fictional character in the Hughes brothers' African American cinematic classic, the coming-of-age drama *Menace II Society*, Mc Eiht discusses how his personality is shaped through his environment; in this case a working-class neighbourhood with high-poverty, broken families, drug-abuse, violence, and lack of guidance. This history may partly mitigate Caine's responsibility for his problematic moral character, as Caine himself is a victim as well. Mc Eiht (1993) *Streiht up menace*, DJ Slip (Prod.), on Menace II Society (CD: Soundtrack), Jive Records.

These issues relate to what is known as *moral automaticity*. Recent empirical research in the psychological sciences has been taken to challenge the moral psychological notion of *moral agency*, at the core of moral philosophy. Traditionally, moral agency is conceptualised as being grounded in conscious, deliberative processes, such that a person's moral behaviour is guided by one's moral reasoning. However, the data purportedly supports the view that moral behaviour is instead driven by *automaticity*; automatic, unconscious, affective processes. As such, people typically lack moral agency over their moral behaviour as reasoning agent, resulting in attitudes, responses, judgements, and actions that adhere to one's automaticity rather than one's rationality. In turn, this implies a challenge for traditional theories of individual moral responsibility as well, since these rely on a notion of rational agency over one's moral behaviour in order to justify responsibility practices; people are responsible for some behaviour in virtue of being the agent of it.

There are many intriguing and still inconclusive issues concerning the existence, nature, and degree of automaticity. However, I am not concerned here with evaluating this, as the topic of this essay is moral responsibility for automaticity. Therefore, while the first part of this essay addresses the empirical plausibility of the moral psychology that the rest of the essay builds on, we can currently just regard this as two base premises; the *significant automaticity thesis* and the *character development thesis*. Thus, starting from the premise that much of people's moral life is issued by automatic cognitive processes, and that one can develop one's automatic cognition, the central question here is what the conditions of moral responsibility are in virtue of which an agent is responsible for moral automaticity, including its acquisition, possession, and expression in moral behaviour. Furthermore, while focussing on this class of behaviour, the account explored here aims to provide general conditions for moral responsibility, encompassing all other modes of behaviour. Similarly, while much of the discussion focuses on implicit biases, it aims to apply to other automatic states and processes as well.

Finally, a proviso concerning the essay's approach and aims. I 'positively' propose a novel account as alternative to the existing discourse. By framing it as such, I aim to explore the strongest version of the concepts that are under discussion, in order to make their central points as clear as possible. To some readers, however, some conceptual distinctions (e.g. between 'moderative agency' and 'developmental

agency', or 'indirect control' and 'developmental control') may not be sufficient to understand the account explored here as an entirely distinct, alternative account. If that is the case, I bid that reader does not get hung up on the framing as more or less distinct, for, even as 'mere' extension of some existing account, the objective of embracing the points of the conceptual distinctions would be achieved, under whatever title that may be. The main points here being that, given automaticity, the moral character development should be appreciated as an important aspect of agency, and with that the opportunities an agent has for developing one's character, in terms of one's rational capacities as well as environmental circumstances.

In §1 I briefly introduce the automaticity challenge and the discourse between 'volitionist' and 'attributionist' accounts of moral responsibility. In §2 I start out with a positive description of a *developmentalist* view as alternative account of moral responsibility, based in a *tripartite view of moral agency*, including a notion of *moral character* and its development over-time. Subsequently, in §3, I analyse a range of moral scenarios in order to explore the explanatory power of developmentalism and further tease out its principles, while exposing the limitations of volitionism and attributionism in comparison. Finally, in §4, I return to detailing the developmentalist principles as considered throughout. I conclude that developmentalism better appreciates our current empirical knowledge of moral psychology including automaticity and development, while simultaneously upholding a normatively useful standard for evaluating moral practices by enabling a more nuanced and accurate understanding of the factors at play.

§1: The automaticity challenge to moral responsibility

Here follows a quite broad albeit eminently rough introduction to *automaticity*. Over the last few decades, the automaticity of human cognition has become one of the most intensely researched phenomena at the intersection of fields such as behavioural, developmental, social, and cognitive neuropsychology. Typically, this research is based in two-system theories of cognition (also often named dual-process theories),

which dichotomously defines 'system 1' processes, which are automatic, affective, and non-conscious (say, intuitions), in opposition to 'system 2' processes, which are conscious, deliberative, and controlled (say, reasoning).[29] (Bargh & Chartrand, 1999; Evans & Stanovich, 2013; Sloman, 1996; Stanovich & West, 2000) Much of this research is taken to evidence that automaticity is so ubiquitous that the majority of human cognition in judgement-formation, decision-making, and action-guidance (hereafter jointly referred to as *behaviour*)[30] is to a large extent driven by automatic processes. As social psychologist Jonathan Bargh and neuropsychologist Tanya Chartrand write, "most of a person's everyday life is determined not by their conscious intentions and deliberate choices, but by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance." (Bargh & Chartrand, 1999, p. 462) Also in philosophy, automaticity is becoming widely discussed, ranging from topics in philosophy of mind, epistemology, and to political philosophy, action theory, and theories of moral responsibility.

Especially in relation to moral behaviour, automaticity constitutes an intriguing and fundamental matter. Traditionally, moral psychology and action theory are grounded in what we may generally call a rationalist paradigm. Rationalism holds that moral behaviour is 'agentive', properly belonging to an actor *as agent*, due to one's conscious moral reasoning being the causal determinant, such that rationality and intentionality are involved. (Kohlberg, 1973; Korsgaard, 2008; Piaget, 1932; Velleman, 2000) In turn, also moral responsibility theories are traditionally grounded in such rationalist notions, holding someone morally responsible for some behaviour due to, for example, one's capacity to consider morally relevant factors and make conscious, deliberative choices. (Wolf, 1990) Consequentially, much of an agent's moral behaviour, the part driven by automaticity, may have to be excluded from the agent's moral responsibility, leaving large gaps in our moral practice. (Levy, 2014)

[29] I do not straightforwardly accept the two-system theory's dichotomous definition of system 1 and 2 processes, or the initial definition of implicit biases provided here, because I believe that system 1 processes, including implicit biases, can be reasons-responsive and controlled rather than merely associative and non-propositional. I return to this later. For now, I employ the terminology as it is traditionally found in the literature, because it helps to clarify the topic and my position.

[30] While some may specifically discuss one specific process, I take discuss judgements, decisions, and actions as a cluster, because of their intimate connections, such that the former two processes ultimately relate to the latter, and especially the latter is most morally relevant. Moreover, as many of the authors mentioned in this essay acknowledge, as a function of their connectedness, the three processes are all susceptible to similar automaticity challenges.

The automaticity literature refutes the rationalist paradigm. As I argued elsewhere, the empirical claims grounding moral automaticity can be seen as two sets of empirical theses concerning moral cognition, and a normative premise involving a standard of agency.

On the one hand, automatic processing, operating through situational factors triggering associated states or processes, are quicker, and its unconscious operation and emotional valence make its operation fairly robust when facing deliberative interference (the *speed*, *unconscious*, *affect*, and *situational* theses). On the other hand, moral reasoning is typically quite frail, since its operation is slow and requires scarce cognitive resources such that it can only be exhibited infrequently, and even when exhibited it often does not rationally track moral truth well due to emotional and cognitive distortions (the *speed*, *resource*, *frequency*, *affect*, *partisan cognition*, and *truth-tracking* theses). As such, reasoning is often not causally determinative of one's moral behaviour, but rather follows afterwards in order to rationalise one's behaviour (the *causality*, *post hoc*, and *confabulation* theses).

This account of moral cognition, then, is evaluated by a normative standard of moral agency in the mind of the rationalist paradigm mentioned above, although sterner, encapsulated in the following set of agency conditions. For an agent to have moral agency, most of one's moral behaviour has to be under deliberative control, meaning that it is causally determined by conscious, rational deliberation in a fairly direct manner (the *frequency*, *control*, *causality*, *consciousness*, and *directness* conditions).

Thus, we have arrived at the *automaticity challenge to moral agency*, holding that; since automatic processes rather than reasoning processes often drive one's moral behaviour, and since reasoning grounds moral agency, people typically lack moral agency. While various authors in psychology and philosophy, under various names for the concepts, and with some content variance, have advanced the automaticity challenge, they all share most of these core commitments outlined above, and instead endorse alternative moral psychological models much in the sentimentalist tradition, emphasising that automatic emotional processes are superior and sufficient for driving moral behaviour. (Baron, 1995; Doris, 2002; Haidt, 2001; Haidt & Bjorklund, 2008; Nichols, 2004; Prinz, 2007)

The Automaticity Challenge to Moral Agency

Empirical premise i	*The primacy of automaticity*: Unconscious, emotional processes are causally determinative of most of people's moral behaviour (judgement-formation, decision-making, and action-guidance).
Empirical premise ii	T*he frailty of reasoning*: Conscious moral reasoning is often rationally deficient and not directly causally determinative of moral behaviour.
Normative premise	*The deliberative standard of moral agency*: Moral agency is marked by deliberative control, the process of conscious, rational reasoning fairly directly causally determining most of one's moral behaviour.
Conclusion	*The lack of moral agency*: People typically lack moral agency over their moral behaviour.

To somewhat further elucidate moral automaticity, let us look more specifically at one form, implicit biases, which form the focus of this essay due to being the most discussed automatic process in the literature.

Implicit biases are automatic attitudes that typically operate without conscious awareness of the agent and are difficult to control even when one is aware of them. Hence, biases influence someone's behaviour (even including perception, evaluation, and emotional responses) so that this often differs from how it would have been, were it determined instead by the agent's consciously endorsed, explicit attitudes. Implicit biases can concern social groups based on class, race, gender, sexual orientation, mental illness, physical ability, religious identity, aesthetic appearance, but also concern many other features, and negatively influence interactions with individuals who belong to these groups due to connecting them to negative properties or stereotypic traits. Studies show that biases can affect an agent's behaviour in relatively minor ways, like blinking more and choosing another seat, or in more significant ways, like job applications and criminal-sentencing decisions. (Blair et al., 2004; Moss-Racusin et al., 2012) Importantly, implicit biases are no rarity but omnipresent; virtually everyone harbours and expresses certain implicit biases, including those who belong to a particular social group themselves and those who explicitly and sincerely avow egalitarian values. (Jost et al., 2009) Ultimately, diverse types of effects, by various agents, all factor in together in systematically continuing and reinforcing patterns of discrimination, marginalisation, and oppression.

In response to the automaticity challenge, a range of authors have criticised the paradigm and defended rational moral agency, often in the rationalist mind. The main dispute concerns the empirical claims, arguing instead that moral reasoning can be truth-tracking and causally determinative much more often, especially through what

we can call *moderating* automatic processes. For example, an agent can become aware of automatic processes operating through consciously reflecting and hence rationally regulate its running further to influence one's moral behaviour or not. (Dreyfus & Dreyfus, 1991; Hogarth, 2001; Holroyd & Kelly, 2016; Kennett & Fine, 2009; Musschenga, 2011; Narvaez, 2011; Pizarro & Bloom, 2003; Sauer, 2012; Snow, 2006) These replies have contributed importantly to further advancing our understanding of moral cognition, and somewhat strengthened the case for moral reasoning, hence ameliorating the automaticity challenge. However, it does not at all seem that agency is successfully defended with that, as, in turn, these accounts face many critiques themselves. (See my doctoral thesis essays 1 and 2) Most crucially, truth-tracking due to emotional and cognitive distortions, and limited frequency due to scarce cognitive resources, seem to be persisting issues. As such, the debate is, at best, at an impasse.

In conclusion, while it is yet an open question exactly how much, how, when, and where people's moral life is automatic, there is now a large body of research and growing consensus that automaticity plays a very significant role in moral behaviour. As such, the automaticity challenge is still very much a heated and pressing issue, which targets shortcomings at the core of moral philosophy at large. This occasion lends itself for a re-evaluation of the fundamental concepts involved, the ground for individual moral responsibility being the concept under discussion in this essay.

Moral responsibility

The notion 'moral responsibility' is taken to mean different things by different philosophers. To be clear, the matter I am concerned with here is the justificatory *conditions* for the judgement of responsibility. I do not wish to enter into any debate regarding different *kinds* of responsibility, and which, if any, is the right or core meaning of responsibility. Concerning this latter debate, the kind of moral responsibility employed here is fairly robust, meaning that it is to treat an agent as the appropriate target for moral appraisal, for certain responses like asking for justification, reactive attitudes like indignation and blame or esteem and praise, and

possibly further actions like punishment or reward.[31] Furthermore, when a judgement of praiseworthiness and blameworthiness is justified, it is yet another, separate matter whether this judgement ought to be expressed, which may not necessarily entail. Finally, the way in which I discuss responsibility rests on the assumption of compatibilism of free will and determinism, or at least on an appreciation of responsibility as a fact of human social life I aim to substantiate, without delving into free will.

Traditional theories of moral responsibility can be roughly divided in two types, *volitionism* and *attributionism*. While individual accounts of one theory may differ significantly from one another, their shared essential commitment to either control or attributability as necessary condition for responsibility can be taken as unifying feature to contrast them from accounts of the other theory.[32] I present the accounts of Neil Levy and Angela Smith as main proponents. This brief introduction is by no means an complete description, but the accounts will be drawn out more when discussed in §3.

Volitionism

One theory of moral responsibility, known as volitionism, focuses on the exercise of the capacity of an agent to consciously choose or reflectively endorse an action or state (or these being causally effective indirectly) as necessary condition, as to relate the action or state to the agent's explicitly considered and endorsed beliefs and evaluative commitments.[33]

Neil Levy's account emphasises control through conscious choice as necessary and central condition for responsibility. "An agent is responsible for

[31] I take this conception of kinds of responsibility to be largely similar to that of Levy (Levy, 2016, p. 6) and Smith (Smith, 2008, p. 370), as to ensure that the discussion here directly address their accounts. Also, to avoid confusion, note that when mentioning 'attributionism' I refer to a theory of *conditions* for moral responsibility, not the *kind* of responsibility known as 'responsibility as attributability' as defined by (Watson, 1996), rather, the kind of responsibility that the attributionist theories of conditions of responsibility discussed here employ is more similar to Watson's 'responsibility as accountability'.

[32] While some accounts may take control or attributability as necessary, but not central condition for moral responsibility, which could make one take these accounts as not properly belonging to the type of theory I categorise them under, I believe we can legitimately do so solely on their acceptance of the necessity of the condition, without including centrality.

[33] Important accounts of volitionism can be found in (Fischer & Ravizza, 1998; Levy, 2005, 2016; Mele, 2006; Rosen, 2004; Vargas, 2013).

something (an act, omission, attitude, and so on) just in case that agent has – directly or indirectly – chosen that thing." (Levy, 2005, p. 2) Levy argues that in cases of actions influenced by implicit attitudes, agents lack 'responsibility-level control' over their actions and or the consequences of their actions, whereby they are not responsible for these, since such control is a necessary condition of responsibility. Firstly, the acquisition, possession, and content configuration of implicit attitudes is not controlled. Levy builds on an account of implicit attitudes as associations between concepts and representations that are the result of their co-occurrence in the learning history of the agent, which entails that "implicit attitudes will display little reasons-responsiveness. They are sensitive to cues with which they have been associated in the agent's learning history, not to (justificatory) reasons." (Levy, 2016, p. 11) Secondly, the expression of implicit attitudes is not controlled either. While in operation, implicit attitudes have a patchy propositional structure, and "since being responsive to reasons requires appropriate sensitivity to the inferential relations entail by semantic content, this patchiness indicates a drastic curtailment of patterned reasons-responsiveness." (Levy, 2016, p. 14)

Attributionism

Another theory of moral responsibility, known as attributionism grew from the work of Peter Strawson on 'quality of will' and Harry Frankfurt on the 'will' as 'higher-order desires'. (Frankfurt, 1971; Strawson, 1962) Attributionism focuses on the necessary condition of the appropriateness of attributing an action or state to an agent as revealing who someone is as moral agent, expressing the agent's 'real self' or reflecting an evaluative judgement or appraisal.[34]

For Angela Smith, an action or attitude can be attributed to an agent if it "reflects her rational judgment in a way that makes it appropriate, in principle, to ask her to defend or justify it." (Smith, 2008, p. 369) Rational or 'evaluative judgements' are not necessarily conscious choices, but rather "tendencies to regard certain things as having evaluative significance." (Smith, 2005, p. 251) More precisely, they are "continuing and relatively stable dispositions to respond in particular ways to

[34] Important accounts of attributionism can be found in (Adams, 1985; Arpaly, 2003; Faraci & Shoemaker, 2010, 2014; Scanlon, 1998; Sher, 2009; Smith, 2005, 2008).

particular situations and not merely one time assessments." (Smith, 2005, pp. 251, footnote twentyseven) An action or attitude reflects these judgements when there is a 'normative link' between them. "If one sincerely holds a particular evaluative judgment, then the mental state in question should (or should not) occur. The 'should' in question here marks the normative ideal of rationality." (Smith, 2005, p. 253) This includes the epistemic assumption that the action or attitude indicates the presence of an underlying judgement.[35] "We take there to be a direct normative connection between the state in question and particular kinds of judgments or evaluative appraisals. Because of this presumed connection, we can make a direct inference from the occurrence of the states to the underlying judgments these responses reflect." (Smith, 2005, p. 254)

Concerning implicit biases Smith argues that while an agent may not be responsible for the acquisition, one is responsible for the possession and expression of it, as this reflects one's evaluative agency. "It is undoubtedly true that the implicit biases most of us harbor are largely a result of cultural influences over which we have little control, this does not change the fact that they are now evaluations we are making that have an influence on our judgments, attitudes, and actions. My claim is not that we are morally responsible for coming to have an implicit bias, but that we are morally responsible for having and manifesting such biases, for the simple reason that we are morally responsible for anything that reflects our evaluative activity." (Smith, unpublished, p. 21) Smith's understanding of how implicit attitudes reflect evaluative activity requires some further unpacking. Unlike Levy's (Levy, 2016, p. 17) suggestion that she supposedly acknowledges his claim that an attitude must in some way cohere with, or be acquired via, an agent's 'deliberative standpoint', in order to be attributable, Smith claims that, "many of our attitudes may be acquired in ways that bypass the deliberative standpoint, but, to the extent that they still involve (subconscious) exercises of evaluative agency, they are things for which we are responsible." (Smith, personal correspondence) Crucially, implicit attitudes are embedded in rich inferential relations with the states of an agent that make up one's practical identity, they are tied up with a process of identifying, evaluating, and accepting putative reasons in favour of the attitudes, which is a rational process,

[35] As the inference rests on an ideal, an actual agent may not always meet that standard, and hence the inference may sometimes be false. However, the inference is fitting unless there is counterevidence available against it. (Smith, 2005, p. 255) But even then, if the agent's mental state fails meeting the standard and can be judged irrational, this is insufficient for the agent not being responsible for it.

despite that those reasons may never be brought to reflective consciousness. (Smith, unpublished, pp. 21-22) Human beings are 'sense-making creatures', and "attitudes themselves rest on this continuing activity of (often subconsciously) taking certain things to count in favor of certain responses." (Smith, unpublished, p. 22) Hence, attitudes and behaviours can be "judgement dependent, but in a way that is not necessarily apparent to us without careful reflection." (Smith, unpublished, p. 18) With this, Smith directly opposes Levy's condition that information is "available for easy and relatively effortless recall." (Levy, 2011, p. 246) Instead, Smith only accepts the general and modest claim that *some* semantic relation between the attitude and the other states of the agent must hold, such that an agent is responsible for one's implicit attitudes, while one may be unaware and not in control of them, because they nevertheless involve exercises of evaluative agency that the agent can be appropriately asked to justify. Smith does, however, hold that agents are less blameworthy for implicit biases compared to explicit biases, not because any excuse condition pertains (e.g. control, awareness, ignorance), but because an implicit bias is only one offensive attitude, without the additional higher-order endorsement attitude (or other second-order attitudes like not caring about the one's bias or feeling guilty about one's bias, when becoming aware of it).[36] (Smith, unpublished, p. 26)

§2: Developmentalism

For an initial grasp of the *character development view*, or *developmentalism* in short, explored here, I will propose a rough definition, followed by a discussion of a hypothetical scenario that can be considered as a paradigm of various modes of moral behaviour and moral agency, the role that automaticity can play in this, how an agent's character can develop, and how moral responsibility relates to all of this. Subsequently I introduce the key condition that developmentalism employs: *opportunity for character development* (including *capacity* and *environment*), a *tripartite model of agency*, and a notion of *moral character*. These notions will be

[36] Smith discusses attributionism's possibility of accounting for degrees of blameworthiness too, which I come to address later.

further elaborated throughout the paper while putting them to work in analysing various moral scenarios.

To shed some light on the nature of this project, developmentalism is unmistakably inspired by Aristotelian virtue ethics. However, while drawing from this rich tradition, the account here is not presented as a neo-Aristotelian or virtue ethical account, mainly because that relationship requires much more elaborate specification and contextualisation than the focus on several seemingly shared basic concepts here allows for. Nevertheless, it is worthwhile to point out the existence of relatedness for three reasons, firstly simply to pay homage when it is due. Secondly, virtue ethics provides to most readers a readily available framework to form an initial understanding of developmentalism, calling to mind ideas of character, character development, and acting from character. And thirdly, in both the focus of the account and the possible challenges, developmentalism has many apparent similarities to virtue ethical theories – even including the overall status of the project, being either a distinct third theory of moral responsibility in contrast to volitionism and attributionism, or merely an account that is best subsumed under the traditional theories, which can be likened to virtue ethics being advanced in modern normative ethics as either a third position in addition to deontology and consequentialism, or merely a particular version of those. Therefore, when beneficial to a specific issue, I will mention similarities with virtue ethics, but otherwise, while I am significantly sympathetic to virtue ethics, much inspired by it, and open to further explore relations in the future, I will not here elaborate on whether developmentalism is a virtue ethical project.

Starting with a rough definition of developmentalist conditions for moral responsibility:

> *An agent is responsible for an action (omission, consequence, or mental state) if it can be appropriately attributed to the agent as reflecting one's moral character, which consists of dispositions and traits rather than choices (but can include choices too), in a way that makes it appropriate to ask the agent to defend and justify one's action and character, the latter regarding the agent having sufficiently had the opportunity*

(depending on capacities and environment) to actively engage with the development of one's own moral character.[37]

Now, consider the following hypothetical scenario.

*Amir grows up in a relatively segregated neighbourhood of a large western European city, and hence his social interactions at home, at his local school, and elsewhere, are for a large part limited to people with a similar immigration, lower socioeconomic class, and/or lower educational background. Initially (*A*), Amir picks up strongly sexist implicit attitudes from his environment, and throughout adolescence he continues to possess these and typically act on them. During early adulthood (*B*), however, Amir becomes increasingly more aware of his attitudes and actions in the light of morality. He begins to actively relate to his character and behaviour more critically, evaluating and reflecting on it, discussing it with others, reading about related matters, practising employing other cognitions and exhibiting other actions for example while around friends, and attempting to moderate his moral behaviours at the meetings. Over time (*C*), Amir's implicit attitudes concerning gender change significantly, becoming increasingly egalitarian, eliciting egalitarian behaviour.*

All throughout this time, Amir is a member of a local community organisation and regularly attends meetings where affairs concerning the neighbourhood and plans for projects are discussed. At the meetings, the other members consist of both men and women, who all have opportunity to speak on the issues at hand.

During time period A*, Amir intuitively disqualifies the contributions voiced by women as less valuable and less important, he automatically listens with only little attention, has an attitude keener on picking out points of criticism, is disposed to wave away worthwhile input, and without a thought interrupts women very easily.*

During time period B*, Amir is aware of his character and effortfully works to re-evaluate, alter, and regulate it, for example through learning about admirable women entrepreneurs, explicitly telling himself to treat women fairly right before the meetings, by holding back on his initial responses during the meetings, and by focusing on the valuable parts of women's contributions, among other strategies, causing him to exhibit automatic sexist behaviour to a lower degree, but nevertheless Amir still has much of the same inclinations and attitudes as during time* A.

During time period C*, Amir engages in the meetings almost completely spontaneously, treating the contributions by men and women equally on their merit, and doing so positively effortlessly, simply acting, so to say, 'from character'.*[38]

[37] There is a question here if 'action by character' and 'action by conscious choice' should be represented in the definition by a disjunction of distinct responsibility conditions for distinct action modes, or, as the current form does, the definition can initially be more general and only subsequently differentiate different action modes with their distinct conditions. I return to this question at the end.

Now, what can we say about Amir's moral agency when we analyse this scenario on a tripartite model? Starting at time period *A*, Amir's behaviour is largely driven by automatic processes. Since these automatic processes are passively acquired through socialisation, they are themselves non-agentive. As Amir is still young, he has only partly obtained a sense of rational capacity such that his deliberation is direct causal or can moderate automatic processes. Additionally, his deliberative and moderative agency are further diminished by the limited environmental opportunity of his so-called 'low opportunity environment'. As for developmental agency, this too is diminished by his capacity and environment, and his age simply temporally restricts his opportunity to have engaged in his character development. As such, we can conclude that overall at this stage of his life Amir typically lacks every type of agency over his moral behaviour.

During time *B*, as an adult, Amir has full rational capacity, albeit possibly somewhat diminished by his environment. Amir exhibits deliberative agency through, for example, consciously choosing certain moral behaviours. And Amir exhibits direct and indirect moderative agency through, for example, cognitive intervention and cognitive preparation, respectively. Amir is also actively engaged with his character development, for which he now has the capacity and time, although still being limited by his environment. However, Amir still possesses the biases he did at time *A*, such that he does not yet exhibit developmental agency.

During time *C*, when Amir exhibits behaviour that is driven by automatic processes, he exhibits developmental agency, since he was agentively involved in developing these processes over-time.

The analysis of the Amir case employs multiple concepts and views that have to be further explained and, of course, argued for, to which we will turn now.

[38] I do not wish to suggest that sexist attitudes in some way *belong* to (Middle Eastern) immigrant populations, as I am aware of the fact that sexism in prevalent among native Westerners just as well. Rather, the scenario aims to stress the importance of intersectionality in the tripartite agency model, through showcasing that opportunity for moral development can be limited by various axes at the same time, including class, race, ethnicity, sex, gender, and housing and schooling segregation, among many other factors.

Rational moral agency

Any account of moral responsibility needs to involve rational agency in some sense, an agent's rational capacities and activity. After all, justification only springs from those things that are suffused with reason. Volitionist accounts include rational agency through rational choice of an action, and attributionist accounts through the rational link of an action with the agent's evaluative judgement. Developmentalism draws on the agent's opportunity for character development, which encompasses *capacity* in interaction with *environment*.

Rational capacities are understood here as being generally in the same mind as concepts such as 'reasons-responsiveness', or 'normative competence', especially as conceived by Susan Wolf as the receptiveness of agents to moral values, moral knowledge, and moral facts, the ability to be reasons-responsive, to be able to be aware of oneself, of one's behaviours, and of one's environment and reflect upon these, to have certain knowledge, ways of obtaining knowledge, and ability to put knowledge to work, and even to have imaginative capacities, among other things. (Wolf, 1990, pp. 121-124) These capacities are understood as being part of an agent's moral character, together with other skills and dispositions (such as implicit biases).

Environmental conditions are constituted by an agent's situational circumstances. Following David Brink and Dana Nelkin, I take an agent's environment is together with an agent's internal capacities to determine the opportunities of an agent, since merely focussing on the latter often does not provide sufficient information to make multi-faceted judgements of an agent's responsibility. (Brink & Nelkin, 2013) For example, jumping on a grenade may be well within a soldier's capacities and as such highly 'controlled' as action, but as a moral situation it also requires extreme sacrifice. Similarly, developing one's racial attitudes may be fully controllable in terms of Amir's adult capacities, but requires immense effort due to his environment.[39] This shows that not merely capacity, but also environment can influence control and difficulty (effort and sacrifice), and with that the degree of responsibility. As such, capacity and environment have to be jointly considered to

[39] Departing from, or possibly merely expanding Brink and Nelkin's account, I apply these two factors primarily to character development, rather than direct moral action.

determine the agent's opportunities, and evaluate what behaviour is morally rational, or reasonably expected *given* those opportunities.

Opportunity for character development as rational agency allows for the evaluation of an agent's involvement as a source in the history of one's own character in two ways. *Initial character development* concerns the acquisition of some character state being due to biological factors, social learning, or other non-agentive sources in contrast to agentive involvement. *Continued character development* concerns the continued possession of some already acquired character state being due to actively embracing and furthering, passively sustaining, or actively disavowing and changing it.

What is peculiar about capacity is that it is a condition of responsibility whereby an agent is a member of the realm of responsible agents, an agent is initially not responsible for having this capacity in the first place. Rather, capacity is largely a matter of biological and social luck. On this point, developmentalism is in harmony with volitionism, for example as found in Levy. (Levy, 2005, p. 7) However, their paths crucially diverge from hereon, since developmentalism is much more cautious and nuanced in excusing agents for their moral character (including capacity), as volitionism more readily does. Typically, on a developmentalist account, an agent is responsible for being the agent one is. This stance, in turn, resonates more with attributionism. However, the background for the developmentalist stance on character responsibility is radically different from attributionism, which simply takes it as a 'bottom-line' of responsibility. For example, Adams writes that "*no matter how he came by them*, his evil beliefs are part of who he is, morally, and make him a fitting object of reproach." (Adams, 1985, pp. 19, my italics) Developmentalism, however, holds an agent responsible for one's character exactly because of the opportunity to develop one's own character, the capacity to work on one's own agency, despite the initial lack of such agency, albeit influenced by the biological and social circumstances of one's capacity. This view of responsibility for one's moral character aims to explain the gradual development of character that we see exemplified in Amir's case, and the gradual development of responsibility that we find in moral practices from youth, to adolescents, to adults, with varying capacities and environments. One of the main challenges is to find a balance between appreciating mitigating factors and sustaining agency. On the one hand, factors that diminish or

enhance the opportunity for character development have to be taken into account. On the other hand, there is reason to be wary of dismissing responsibility, since this can alienate a part of an agent from one's agency, severely damaging a serious perception of an agent *as agent*, fully functioning, with agency over one's own behaviours and character, instead of merely partly agent and partly causal mechanism.

The last major feature to point out at this stage is that developmentalism operates with a particular focus of the object under evaluation. As the phenomenology of *direct moral behaviour* is principally automatic, the locus of rational agency is on *indirect moral behaviour*, or more precisely, offstage character development. As such, continuing Samuel Butler's opening metaphor, being a violin player is not as much an *onstage* affair, concerning a particular performance, as it is an *offstage* matter, practising for hours at home and in other training scenarios with others who provide feedback. Similarly, rational moral agency does not, first and foremost, pertain to *onstage moral behaviour*, which is often driven by automaticity, as it plays out in *offstage moral behaviour*, the opportunity for character development an agent had throughout one's life, which led the agent to become the agent performing now. With this, I do not aim to abolish action evaluation altogether, but rather evaluate an action as belonging to an agent with a history.[40] Lastly, the offstage as locus of moral agency can be thought of as being similar to 'indirect agency', for example as employed in volitionism through 'tracing' an agent's 'indirect control' to a point in time where one had 'direct control'. While indirect control catches much of what developmentalism aims at, such that identifying developmentalism as an indirect control theory is not wholly wrong, it is also not wholly correct due to important differences between offstage and indirect and the conception of tracing and control. All of these issues will be address further throughout the paper.

Tripartite model of moral agency

As seen in the analysis of Amir's case above, developmentalism works with three different *modes* of agency, to which three types of responsibility are tied (see Table

[40] There may be a similarity here with virtue ethics being conceived as focusing on the moral agent rather than the moral action, but expanding on this requires a separate essay. See (Slote, 1995).

1). Any theory of moral responsibility builds on a particular understanding of moral psychology. Here, I draw on a *tripartite model of moral agency*, as argued for elsewhere (doctoral thesis essay 2). As such, I will now only provide a brief sketch of the model.

Agency mode	Objective	Cognitive behavioural process	Responsibility type
Deliberative agency (*prohairesis*)	Action (direct)	Conscious deliberative processing directly causally determining some moral behaviour	Direct responsibility
Moderative agency I (*enkrateia*)	Action (direct)	Cognitive intervention on active automatic processes during the processing of some moral behaviour	Direct responsibility
Moderative agency II (*enkrateia*)	Action (indirect)	Cognitive preparation or environmental regulation just prior to the processing of some moral behavioural to prompt activation of certain automatic processes	Indirect responsibility
Developmental agency (*aretê êthikê)*	Agent (development)	Moral character development over-time and resulting developed automatic processes subsequently determining some moral behaviour	Developmental responsibility

(Table 1: *Tripartite model of moral agency with corresponding types of moral responsibility*)

The first agentive mode is *deliberative agency*.[41] This mode can be seen as the traditionally most clearly appreciated agentive concept, overlapping with a rationalist notion of conscious deliberation, and an Aristotelian notion of 'prohairesis' (conscious choice) resulting from 'boulesis' (rational reflection).[42] Such reasoning is virtually the sole driver of one's behaviour. It aims at exhibiting behaviour (action-focus), and relates to the behaviour in an operationally fairly unmediated manner (direct). In turn, agents have 'direct responsibility' for such behaviour. To illustrate, a teacher can grade assignments by carefully assessing only the relevant factors, or parents can deliberate whether to send their child to a higher-appraised, homogenous White, or a lower-appraised ethnically-mixed school.

The second mode is *moderative agency*. In this mode, deliberative and automatic processes operationally interact. More precisely, deliberative processing in some way regulates the operation of active automatic cognitive processes. In Aristotelian terms, one's moderative agency is marked by 'enkrateia' (mastery) over one's operant 'pathos' (emotions, or, in our discussion, automaticity). Being enkratic, the agent continues to possess their automaticity, but it does not determine one's behaviour. As

[41] To conceptually distinguish three modes of agency in terms of the mechanisms that are involved, I employ the term 'deliberative agency' (and behaviour), rather than 'direct agency' (and behaviour). With this, I hope avoid confusion with concepts such as 'direct control', which typically means 'direct operational impact on behaviour', because such operational directness does not exclusively apply to deliberative agency, but also to 'direct moderative agency'.

[42] This is only one interpretation of Aristotle's concepts, which is merely meant for the purpose of clarifying the agentive modes, without entering into a discussion within Aristotelian scholarship.

such, this mode is also action-focused. Most defences of (a rationalist notion of) agency in the light of the automaticity challenge can be categorised as moderative agency. Based on the way in which deliberative processes work on automatic ones, three sub-modes can be distinguished.

Direct moderative agency can be exhibited through *cognitive intervention*, in which an agent's deliberative processing intervenes on activated, operant, automatic processes, as to regulate whether or how they further influence one's behaviour. This sub-mode shares the directness to the behaviour with deliberative agency, but is distinct from such agency in the operational interplay of deliberative and automatic processes. As such, the ensuing type of responsibility is also direct. To illustrate, Amir can notice and block interruptive inclinations driven by sexism, and a teacher can pay attention to operant biases and adjust the evaluation of papers to them.

Indirect moderative agency can be exhibited through deliberatively regulating the behavioural impact of automatic processes prior to these being activated and operant. As such, this sub-mode is still action-focused, but relates to it somewhat less directly. Therefore, such behaviour falls under indirect responsibility. Indirect moderative agency can itself be distinguished in two broad categories. *Cognitive preparation* involves prepping one's cognition to make the activation of certain automaticity more likely. For example, before marking papers, a teacher can formulate the goal to themselves to do so in an egalitarian manner (so-called 'implementation-intentions'), and Amir resolved to pay extra attention to valuable content in women's discussion contributions. *Environmental regulation*, somewhat differently, involves engineering the situation in which one will exhibit moral behaviour such that it will more likely trigger certain desired automatic processes (or not trigger undesired ones). To illustrate, a teacher can anonymise papers so no personal information elicits biases that influence the evaluation, and Amir can push for an equal gender-representation in meetings causing women to feel more empowered to speak and men more compelled to acknowledge each individual woman's contribution.

The third mode is *developmental agency*. In this mode, an agent consciously engages with the configuration of one's automaticity, or rather, the development of one's own moral character. In turn, one's moral behaviour is mostly driven by one's moral character. As such, the operational phenomenology of the determination of behaviour

is automatic, but the agency is located in the developmental stage earlier on where reasoning is imbued in the configuration of the automaticity. In Aristotelian terms, again, one may think of behaviour from *aretê êthikê* (virtuous character), or with *phronesis* (practical wisdom), without operational prohairesis, boulesis, or enkrateia, but instead against a history of conscious character development. In order to further define developmental agency, it may help doing so by focusing on several differences with other modes of agency (especially indirect moderative agency).

1. Other agentive modes are action-focused (either direct or indirect), having the determination of some moral behaviour as its objective. Developmental agency, instead, is agent-focused, concerned with the formation of the agent's constitution (developmental).

2. Other agentive modes are enkratic, meaning that while one's automaticity does not operationally determine one's behaviour, the configuration of one's automaticity is typically retained. Developmental agency, in contrast, reconfigures one's automaticity.[43] (See my doctoral thesis essay 2)

3. Other agentive modes, even indirect moderative agency, are exhibited 'in-action', either within or just-prior to some moral situation (3a, *temporal demarcation*), through one or several clear behavioural instances (3b, *behavioural demarcation*), that are performed within a limited time-frame of each other and of the ultimate moral behaviour (3c, *temporal demarcation*), in order to ultimately determine one or one or several specific moral behaviours (3d, *foreseeability*). Developmental agency, on the other hand, can be exhibited 'offstage',[44] not within, but rather 'detached' from any moral situation, at virtually any time and place, say, during one's free time, home-alone (3a), and is done through numerous behaviours (3b), which are performed over an extensive period of time (3c), which, in turn, has an effect

[43] Admittedly, indirect moderative agency *can* have an impact on the configuration of one's automaticity, but all modes of agency can. Rather, what matters to distinguish them is what the main aim of each mode is (and, in addition, the effectiveness of achieving that – as the developmental side-effects of deliberative and moderative agency is not very effective, I argue in doctoral thesis essay 2).

[44] Illustrating this point, one can think of a violin player, who develops their skill not as much on-stage, while giving a performance, but rather off-stage, through spending large amounts of time practising, at home, in other training settings, playing, listening, discussing, and even reading about music.

on a wide range of a person's ultimate moral behaviours that are driven by one's character (3d). (This point is elaborated later)

4. Other agentive modes employ 'operant reasoning', reasoning in-action in a moral situation, which is thus vulnerable to all the automaticity challenges. Developmental agency can employ 'detached reasoning', in less hostile environments, framed to be less emotionally heated, with less time-pressure, and less scarcity of cognitive resources. (See my doctoral thesis essay 2)

5. Other agentive modes have a limited 'factor range', only relating to factors that can be clearly specified and comprehensibly determined, such as some specific knowledge, or a certain state-of-mind. Developmental agency targets one's moral character, which comprises a wide variety of factors and their interrelations with one another, such as emotional dispositions, attitudes, values, and skills. (See my doctoral thesis essay 2)

Relating this agentive mode to responsibility, character development does not fall within an agent's indirect responsibility, due to all the differences specified here. Rather, it is useful to think of it as a conceptually distinct type, *developmental responsibility*, since the agency here is achieved through the self-development of one's character.

For example, besides having a responsibility to exhibit morally appropriate behaviour in some moral situation (with which deliberative and moderative agency are concerned), one may also have, opportunities allowing so, a responsibility to in general engage with the development of their beliefs and attitudes and such, outside of any moral situation at hand. A teacher can read about educational performance of students from various class backgrounds, attend workshops on the functioning of biases concerning ethnic groups, and decorate one's wall with images of intellectually and culturally admirable working-class people. Amir can reflect on and evaluate his attitudes, watch a documentary on women entrepreneurs and inventors, and practice other thought-patterns and reactions.

Moral character and character development

Besides the three modes of agency, the notion of *moral character* is very central to the tripartite model (see my doctoral thesis essay 2). The notion of moral character employed here is, for now, very broad general. I employ moral character as an umbrella term for a large family of mental phenomena. As such, on a very loose definition, a person's moral character is the entire set, constituted by wide variety of phenomena, ranging from an agent's beliefs, desires, values, affective attitudes, behavioural dispositions, emotional reactions, spontaneous responses, intuitions, and habits, to what grabs one's attention, perceptions, sensitivities, skills, and one's rational capacities, among other things. And thus, most automatic states and processes, such as implicit biases, unconscious stereotypes, and prejudices, are part of one's character too.[45]

Moral character phenomena, or 'attributes, share four important features. Firstly, character attributes can admit of a high degree of automatic processing. Secondly, character attributes are highly integrated with one another, in their acquisition, possession, and expression. For example, an implicit attitude is tied up with certain beliefs, values, and emotional dispositions. This picture is very similar to what we saw in Smith's attributionism, who sees automatic attitudes as 'embedded in rich inferential relations with the states of an agent that make up his practical identity'. Also Holroyd and Kelly support such a view, arguing that implicit biases interact with one's values. (Holroyd & Kelly, 2016, p. 3) While these three authors take this to claim that character attributes reflect the agent and are thus within one's responsibility, developmentalism rather does so due to the fourth feature (developmental opportunity).

The third feature that character attributes share a functional role in determining behaviour. They can be triggered, influence evaluation, judgement, and determine action (note feature one; they can do all this automatically). Fourthly, character attributes can be developed, meaning that their configuration as to, for example, what triggers them, what emotion to rouse, or what reaction to drive, can be changed. Such change requires many efforts, over a longer period of time. Moreover, the third feature, interconnectedness, implies that development typically concerns not

[45] While some of these phenomena may turn out to not belong to character, this can be discussed later. For now, their grouping serves the modelling of their role in agency.

as much one specific sub-element of one specific sub-attribute, but is rather a project that targets a broader attribute-network or moral character on a whole. The development of character opens up a space for agency, as an agent can engage with determining one's own moral character. Note that with the focus on consciously developing one's character as agency, developmentalism diverges again from attributionism, which takes one's character as a given. Consciously working on one's character development appears more like a volitionist approach, although volitionists such as Levy and Rosen typically oppose the possibility of it.

§3: Explanatory power

In this section I will analyse a range of real and hypothetical moral scenarios, some which are often discussed in the literature, and some new ones. The purpose of this is to exhibit the explanatory power of developmentalism in contrast with volitionism and attributionism. Through looking at several cases together it will become clear that developmentalism allows for more nuanced judgements of degrees of responsibility, whereby it can better differentiate the cases from one another. I will analyse cases relating to initial and continued character development in the light of attributionism (a), and in the light of volitionism (b), and cases relating the extensiveness of character development in the light of volitionism (c).

3A: Attributionism and the opportunity for character development

The first range of cases concern the importance of including an agent's opportunity for involvement with the continuation of one's acquired character as a factor of responsibility judgements. I first analyse three scenarios from a developmentalist account and subsequently discuss what attributionist and volitionist evaluations would amount to, and why the former is preferable.

Phineas Gage, a real case discussed by Levy.

> *Phineas Gage was a railroad construction worker, known as a hard-working and trustworthy person. An accident with an explosion to blast away rock for the roadbed shot a tamping iron through Gage's skull. Miraculously, Gage survived the mishap, but the brain damage he suffered significantly changed his personality, turning him permanently into a dissolute and anti-social person.*[46] (Levy, 2005)

JoJo, a fictional story presented by Susan Wolf.

> *"JoJo is the favourite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given special education and is allowed to accompany his father and observe his daily routine. In the light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or death or to torture chambers on the basis of a whim. He is not* coerced *to do these things, he acts according to his own desires. Moreover, these are desires he wholly* wants *to have. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life expresses a crazy sort of power than forms part of his deepest ideal."* (Wolf, 1987, pp. 53-54)

And, as third case, we can recall the story of Amir, as described above in §2.

When we take these three cases together, judging the moral responsibility of each involves various elements that differ per scenario, but they also have crucial commonalities. I am not interested here with the difference in the badness of each person's behaviours. Instead, I am interested in each agent's possibility for involvement as an agent with becoming the person they are, the possibility of being an agent of their own character, a character from which, in turn, they exhibit the bad behaviour.

In Phineas' case, the origin of his character is brain damage, which was caused by an accident. The first thing we can take from this, is that since this accident is not described as being due to his own lacking, Phineas has no responsibility for the brain

[46] Gage's personality and behavioural changes were probably not nearly as extreme and enduring as Levy's account of them, as documented by those in contact with him and accounts of his later re-socialisation and employment. However, the version presented by Levy serves well for the philosophical point here. See (Macmillan, 2002).

damage, and thus for acquiring his new character. Secondly, as the character change is due to severe brain damage, this is permanent and unchangeable, so Phineas has no opportunity for involvement in whether or not this character will continue to be as it is, and thus no responsibility.

In JoJo's case, his character is formed through extreme conditioning by his father and quite possibly the whole environment of dictatorship. Hence, JoJo has no responsibility for the initial acquisition of his malicious character. Subsequently, living in the dictatorial environment, it is likely that JoJo has little opportunity for continued character development, changing his acquired character. While his rational capacities seem in tact, his environment is severely lacking. For example, fearing repercussions, few people will challenge his evil ways, whereby much of the information that is available to JoJo will be uncritical. As such, JoJo has only a low degree of responsibility for the continuation of his character and the behaviours that are issued by it.[47]

In Amir's case, he initially acquires his character through a common process of social condition, as found all around us. Since Amir is not and cannot be agentively involved with this process while lacking the capacities as a child, he has no responsibility for his character and behaviours then. However, as an adult, Amir does have the capacities, in addition to albeit somewhat diminished environmental possibilities of his often called 'low opportunity environment'. As such, Amir has a degree, probably a higher degree than JoJo, of responsibility for the continuation of his character and the behaviours from it.

Comparing the three cases, developmentalism allows for a range of judgements of responsibility, holding Phineas, JoJo, and Amir responsible to different degrees, by taking into account their opportunity for being involved in the continuation of one's character, including both their capacity and environment. None can be said to have played a role in their initial character acquisition. Further character development is impossible for Phineas, extremely difficult for JoJo, and sub-average for Amir. Hereby we can judge Amir to have sub-average responsibility, JoJo very little, and in the case of Phineas we have reason to employ the unchangeable permanency as a full-fledged excuse.

[47] Wolf herself understands JoJo as 'normatively insane', incapable of recognising badness and goodness. (Wolf, 1987, pp. 379-385) Instead, I interpret JoJo's state as most people do, as someone who *does* have at least sufficient normative competence to be an agent, but maybe somewhat diminished, besides, most crucially, diminished environmental circumstances. (Faraci & Shoemaker, 2010)

From an attributionist perspective, we cannot similarly differentiate these cases. Instead, attributionism judges all three agents as fully responsible, regardless the differences in developmental opportunity. Opportunity for neither initial acquisition nor continued possession matters, only *that* one possesses a certain character. Attributionism's trouble in accounting for these factors has been established by various authors, notably Manual Vargas and Dana Nelkin. I will first discuss continued possession, and then initial acquisition.

The first problem for attributionism is that it is a 'structuralist' account in that it focuses on the current structure of an agent's psychology and disregards history. As Manuel Vargus has extensively argued, structuralist approaches, which focus on "whether the agent at the moment of action had the relevant psychological structure(s) in place", cannot account for the relevance of the history of said structures. (Vargas, 2006, p. 335) Vargas focuses on 'manipulation cases', such as nefarious neurosurgeons implanting certain structures in someone's mind. We can naturally expand cases by including not merely single agents as manipulators, but also other external factors, such as brain damage (Phineas) and social conditioning (JoJo and Amir). As such, it becomes ever more evident that not merely someone's current character, but also one's history, the acquisition and continuation of that character can be an important factor in evaluating responsibility. Attributionism lacks the resources to account for an agent's history. We can see this, for example, in Smith's position as discussed earlier, understands one's real self to involve two features; integration and stability. Firstly, one's real self is integrated, meaning that a particular state does not stand on its own, but is embedded in rich inferential relations with one's other states, jointly making up one's practical identity. Elinor Mason, on an alternative attributionist account that focuses on 'quality of will', describes this idea of state integration well, arguing that implicit biases are tied up with one's 'deep motivations', such as contempt, disgust, or a disposition such as 'openness to accept hierarchies that favour us', and it is because one's states are integrated as such that they express an agent's bad quality of will. (Mason, forthcoming, p. 5) The second feature of the real self is that one's states are stable over time, recalling Smith's note that only continuous, not temporary states can be taken to express one's real self. Similarly, Thomas Scanlon holds that an agent's attitudes and behaviours can be said to reflect one's 'judgement-sensitive attitudes' and character when they are stable; it

does not matter for someone's responsibility *how* one's personality came about (by accident, drugs, hypnosis, or brain stimulation), or whether one can *change* their personality, all that matters is *that* someone, in fact, *has* a certain personality. (Scanlon, 1998, pp. 22, 277-279) Maybe most explicit on this point is Nomy Arpaly. She writes, "HP (the Hapless Patient who is brainwashed by Dr. Nefarious) currently has the psychology of a bad or blameworthy person, except that he acquired that psychology in a strange way." (Arpaly, 2003, p. 166) Or, in the case of kidnapped Patty Hearst, "it matters very little to our judgment if she has indeed been brainwashed deliberately or if she just converted, irrationally, due to the duress she was under (the "Stockholm Syndrome")." (Arpaly, 2003, p. 166) "If we exempt from blame any murderer or terrorist whose convictions or character were acquired irrationally, we would exempt too many murderers and terrorists." (Arpaly, 2003, p. 167) A person may "not (be) to blame for creating his bad character, (…) but that need not reduce from the blame he deserves." (Arpaly, 2003, p. 170) With that, Arpaly embraces the principle of "constitutive moral luck: upbringing, history, and perhaps genetic disposition can contribute to making someone a better or worse person." (Arpaly, 2003, p. 171)

As such, by merely taking an agent's current character as condition of responsibility, attributionism fails to differentiate between cases that differ qua continued history of one's character. Comparing such cases, we can deduct that continued history *does* factor in determining responsibility evaluations, and developmentalism can account for this being a condition of responsibility and therewith differentiate appropriately.

Now, one way out of the deficit of a structuralist focus on one's actually possessed character would be to understand persistence of character as merely a necessary, but not sufficient condition, to which other conditions can be added. Scanlon does as much by introducing 'general capacities' ('reflection', 'self-governance') as extra condition, such that agents who lack these cannot be appropriately blamed, but "must be seen, rather, as simply a force to be dealt with, like an animal." (Scanlon, 1998, p. 280) Somewhat similarly, Arpaly mentions that the amount of 'moral concern' required for some behaviour can be a responsibility condition that mitigates the degree of blameworthiness in the case of agents with mental conditions, because these are 'nonpsychological states', states that do not fit one's other attitudes. (Arpaly,

2003, pp. 167, 171) Faraci and Shoemaker also argue that "the degrees of blameworthiness track the degrees of attributability of actions: actions are more or less attributable to agents in these sorts of cases depending on the degree of difficulty they are judged to have in recognizing various features of their actions about which they remain ignorant." (Faraci & Shoemaker, 2010, p. 331) And also Smith claims that attributionism can allow for 'difficulty' being a responsibility condition that influences degrees of responsibility. Discussing the scenario of Abigail, who grew up in a racist environment, Smith argues that "our understanding of the circumstances in which a person's evaluative tendencies were formed may, however, have a very important influence on the kind or degree of moral criticism we think it appropriate to make. We can appreciate how difficult it might be for Abigail to come to recognize the viciousness of her own evaluative judgments, given their early entrenchment in her psyche, and also how difficult it might be for her to modify these judgments once their viciousness is recognized." (Smith, 2005, p. 268)

I argue that attributionism cannot include responsibility conditions such as capacity, concern, or difficulty to formulate judgements of degrees of responsibility for two reasons. The first problem is that it is very unclear degrees of responsibility are to be conceived on an attributionism account of 'full' kind of responsibility, including accountability, not mere attributability, as discussed earlier. However, on top of the degrees-statement, Smith also writes that "it is very important not to conflate claims about responsibility and claims about blameworthiness." (Smith, 2005, p. 266) This is a highly obscure statement, since her account is generally taken to relate to a robust notion of responsibility that includes blameworthiness, similar to the one employed in this paper, which is reflected in papers by others like Levy 'so that they do not talk past each other', as well in Smith's own papers when writing that responsibility permits responses "well beyond aretaic appraisals." (Smith, 2012, p. 567) And regarding the inclusion of difficult circumstances to develop a certain character as a condition of responsibility for one's current character and resulting behaviours, Smith also writes: "This question of responsibility (namely, the responsibility one has for becoming a certain kind of person) must be distinguished from the question of one's responsibility for the attitudes one in fact holds. In order to regard an attitude as attributable to a person, and as a legitimate basis for moral appraisal, we need not also claim that a person is responsible for becoming the sort of person who holds such an attitude. That is a separate question according to the view I

am putting forward. What matters, according to the rational relations view, is that the attitude is in principle dependent upon and sensitive to the person's evaluative judgments." (Smith, 2005, pp. 267-268) Or, on another occasion: "It will be a complex story, for each and every one of us, how we became the sorts of people we are, with the particular values, interests, cares and concerns that we hold; and very few, if any, of us can plausibly claim to bear full or even substantial responsibility for how we became the particular people we are. Even so, I submit, we cannot help but regard ourselves as responsible and answerable for the particular judgments expressed in our actions and attitudes, regardless of what circumstances may have shaped these assessments." (Smith, 2008, p. 389) Thus, it remains obscure how Smith envisions making a judgement in degrees and a judgement of full responsibility at the same time.

The second problem with extra conditions that ground degree-judgements is even weightier, as it directly opposes the core of attributionism. If Smith wants to include the 'difficulty' of Abigail to change her attitudes, she invokes an epistemic and control condition in the judgement of responsibility. When invoking the difficulty of the situation, stating that an agent lacks opportunity to see the badness and change one's attitudes, due to either, for example, 'entrenchment in one's psyche', or the unavailability of other knowledge, attributionism employs capacity and environmental opportunity conditions. Doing so in effect invokes other conditions that do not merely supplement, but *overrules* attributability, since the presence of these conditions does not make the attitude less attributable. Nelkin extensively argues for the same in relation to Arpaly, elaborating that the will that is 'manifested' by some behaviour comes apart from the will it 'takes'. She writes that, while it may mitigate responsibility, "difficulty does not decrease the expression of ill will." (Nelkin, 2016, p. 363) As such, attributability is only a necessary condition, not a sufficient one. However, following this argumentative strategy further, it logically follows that, since 'low opportunity' can be a mitigating factor in some situations, there is no reason that 'no opportunity' cannot be an excusing factor in other situations (e.g. Phineas). With that, attributability is not even a necessary condition anymore. As a result, taking on conditions such as capacity, concern, or difficulty opens up attributionism for exactly the crucial role of volitionist control conditions that it aims to avoid. In conclusion, it seems that as it stands, attributionist accounts cannot accommodate opportunity conditions, cannot yield judgements of degrees of responsibility, and cannot take

developmental differences into account. Since developmental opportunity strikes us as a significant factor to differentiate cases, attributionism has a crucial shortcoming where developmentalism does not.

Besides in continued development, the initial acquisition is another way in which the opportunity for character development as a condition of moral responsibility can play out. To examine this issue, consider the following scenario.

Felix the reluctant implicit feminist:

> *Felix grew up in an entirely egalitarian society, where, among other egalitarian values, gender equality is strongly endorsed by everyone and embedded in all practices. Accordingly, Felix acquired feminist attitudes through his upbringing and social learning. As an adult, however, his explicit attitudes changed significantly, and Felix became a strongly convinced sexist. No longer feeling at home in his society, he moves to ours, where sexism is more prevalent (and more accepted/pardoned). Nevertheless, his initial socialisation entrenched the egalitarian attitudes so deeply in his psyche that Felix continues to possess implicit feminist attitudes that oppose his explicit sexist attitudes. Subsequently, in his work, conducting job interviews, time after time Felix is swayed by his implicit feminist attitudes, which operate automatically, outside of his conscious awareness and control, and spur him to judge women's suitability for some function by their relevant qualities and merit. Thus, Felix's character causes him to exhibit morally good behaviour by treating women equally, doing so even to a much higher degree than the average person in our society, and with that, possibly even beyond what one could reasonably expect of a moral agent in our society.*

What this scenario, meant as being an inverse case of implicit sexism, is aiming to show is the importance of including the aetiology of character as a condition for responsibility. When we judge the responsibility, and with that the blameworthiness or praiseworthiness, of Felix, we notice that his morally good behaviour is not in line with, and not caused by, his explicit attitudes. From a developmentalist account, Felix's behaviour is seen as caused by his character. His character, however, is formed without the active involvement of Felix himself. As such, he is not part of the aetiology of the character. If anything, his efforts have been to change his character for the worse. With that, while the good actions arise form his character, and are as

such attributable to Felix as something he is responsible for, he is not praiseworthy for the actions, since his praiseworthiness is annulled by the lack of his involvement in the initial history of his character together with his efforts against the good action. The mitigated judgement of a more or less zero degree of praise mirrors the lower degree of blame for bad actions that come from a bad character, the aetiology of which the agent was not involved in.[48]

Alternatively, on an attributionist account, the lack of involvement in the history of the implicit feminist attitudes does not matter for the judgement of responsibility. As for the opposing explicit sexist attitudes, on a strict interpretation of attributionism, these are irrelevant, since they do not influence his behaviour, and do not fit with his other deeply held attitudes, so they do not properly belong to Felix. Or, on a charitable interpretation of attributionism, the higher-order attitudes mitigate the praiseworthiness of the behaviour caused by lower order attitudes. Nevertheless, while mitigated, Felix is still praiseworthy, since his lack of involvement in feminist development is disregarded.[49] Felix has moral luck. Arpaly, for example, writes on moral luck that "a virtuous person may rightly thank her parents for "instilling a sense of duty" in her, but she is still praiseworthy for her dutiful actions, even as her parents are praiseworthy for instilling a sense of duty in her." (Arpaly, 2003, p. 171) Rather than simply taking moral luck as a given, developmentalism can actually account for Felix's lack of involvement in the acquisition of his character in addition, as a factor that mitigates his praiseworthiness, or, in other words, as a condition of moral responsibility – which theories of moral responsibility aim to do as fine-grained as possible. This enables developmentalism to differentiate cases not merely on the basis

[48] A note can be made here that, although mirroring, there may still be small differences in the degree of responsibility, blameworthiness being less mitigated than praiseworthiness. This may relate to what is known as the 'Knobe effect', the asymmetry between responses to bad and good outcomes in relation to agent intentionality. (Knobe, 2003) The point remains valid, however, that developmentalism can account well for the intuition that we would not judge Felix as praiseworthy for his reluctant good behaviour, while we do judge Amir as (albeit mitigated) blameworthy for his reluctant bad behaviour. For discussion on the asymmetry in praise and blame, also see (Wolf, 1980).

[49] Arpaly's discussion of the scenario of Huckleberry Finn may come to mind here, which can be said to have certain resemblances to Felix's scenario, since Huck does the morally right thing by acting against his 'best judgement' and explicit beliefs of duty by not rating out Jim, who is attempting to escape slavery, for which Arpaly argues that Huck is praiseworthy. (Arpaly, 2003, pp. 9-10, 75-17) However, while Huck was not aware of the reasons, Arpaly argues that he still acted for a moral motive. Hence, Huck's action is driven by a different kind of mechanism from the implicit feminism of Felix, since Huck, unlike Felix, can be said to, beyond having an implicit attitude, also have further attitudes, like the realisation that slaves, like Jim, are full-fledged human beings. "Huckleberry's long acquaintance with Jim makes him gradually realize that Jim is a full-fledged human being, a realization that expresses itself, for example, in Huckleberry's finding himself, for the first time in his life, apologizing respectfully to a black man. While Huckleberry does not conceptualize his realization, it is this awareness of Jim's humanity that causes him to become emotionally incapable of turning Jim in." (Arpaly, 2003, p. 10) This difference in motivation and other attitudes shows why the Huck scenario is significantly different to form a counter-example, as it explains why Huck could be more praiseworthy than Felix.

of current character supplemented by current higher-order attitudes, but additionally include aetiological involvement, which Felix lacks, but Amir is praiseworthy for.

3B: Volitionism and the opportunity for character development

There are two main issues I will address concerning volitionist, corresponding to two types of volitionist accounts. First I discuss Neil Levy's structuralist account to discuss the opportunity for character development, similar to the foregoing discussion of attributionism. Subsequently I discuss John Fischer and Mark Ravizza's historical account to discuss the extensiveness of character development.

Contrary to attributionism, on a volitionist account of the three cases above (Amir at time *C*, the still biased adult), the agents are all fully excused and thus not responsible. Similar to attributionism, this is because volitionism (as a structuralist view) also disregards the differences in developmental opportunity, which I will argue is an important shortcoming. For volitionism, all that matters, is that in the moment of action, none of the agents have a sufficient sense of control over their behaviours. Phineas has diminished capacity due to his brain damage, which impaired his reasons-responsiveness. JoJo can be said to lack capacity in the sense of being unable to recognise and respond to moral reasons properly. And Amir lacks the capacity to control his behaviour from being influenced by unconsciously operating implicit biases. Since Amir's case is central to our project, let us elaborate on a volitionist understanding of this.

Neil Levy writes, "an agent may lack direct responsibility for an action caused by their implicit attitudes, because given what her implicit attitudes were at *t*, it would not be reasonable to expect her to control her behavior, or to recognize its moral significance, or what have you." (Levy, 2016, p. 6) Levy argues for the lack of control as follows. In cases where an agent's moral behaviour is partially controlled by one's implicit attitudes, and would have been different were it controlled by one's explicit attitudes, 'personal-level control' is prevented; deliberate and deliberative control, exercised in service of explicit intentions. Personal-level controlled behaviour has very demanding epistemic conditions, which agents typically fail in the case of automatic behaviour. Agents are usually not aware of the automatic processes driving

their behaviour, and even when they are, the way in which these influence one's information processing is opaque, and there is no sound method to inhibit or moderate this influence. Drawing on Fischer and Ravizza's work, Levy holds that behavioural control has to be understood as a mechanism that is 'moderately reasons-responsive'; 'receptive' to recognise reasons *as* reasons, and 'reactive' to appropriately respond to them, exhibiting some minimal pattern in the way it does so. (Fischer & Ravizza, 1998) Besides personal-level control, Levy also puts forth consistency as a condition, by invoking the notion of the 'deliberative standpoint', which is constituted by a relatively coherent set of attitudes. From this, he argues that an attitude only belongs to an agent, and a consideration only counts as a reason, when it is consistent with one's 'web of attitudes', either through the acquisition, elimination, or annexation to the self. With the control and consistency conditions in hand, Levy evaluates implicit attitudes as displaying insufficient reasons-responsiveness, since they are stimulus-representation associations that agents develop through co-occurrence in an agent's learning history, not associations that are sensitive to reasons. As such, they are not 'beliefs', but 'patchy endorsements', because while they do have sufficient propositional structure for truth-conditions, it is insufficient for them to count as beliefs, since they interact only sometimes with only some other propositionally structured representations. For example, disfavouring qualified female job candidates exhibits a failure of 'systematicity' of the interaction with other representations, since the agent is able of recognising job qualifications in other contexts. Or, favouring a sugar jar labelled "table sugar" over one labelled "not poison" exhibits a failure to process negating information properly. On an alternative volitionist account, Jennifer Saul argues that agents are not responsible for automaticity-driven behaviour for the same two main reasons as Levy. Firstly, people attain implicit biases without choice, by living, for example, in a sexist culture. As people attain their biases through enculturation, they are unaware of having them, and hence cannot control them. Secondly, people lack control over expressing implicit biases. Even when people become aware of having them, "they do not instantly become able to control their biases, and so they should not be blamed for them." (Saul, 2013, p. 55) Lacking control, Saul concludes, implicit biases do not properly 'belong' to the agent, since they are not indicative of who someone is; which is rather indicated by their conscious attitudes.[50]

[50] Saul adds a further argument, that judging people as 'sexist' or 'racist' would cause them to be defensive and

As such, based on a demanding view of control, implicit attitudes are not controlled. Moreover, based on a demanding view of agency, uncontrolled attitudes do not properly belong to the agent. Levy understands 'agency' as a deliberative standpoint that is constituted by a relatively coherent set of attitudes. As such, an attitude only belongs to an agent, and a consideration only counts as reason, when it is consistent with one's web of attitudes, either through the acquisition, the elimination, or the annexation of an attitude to the self. Since implicit attitudes are not processed in this way, and are thereby not consistent with one's other attitudes, they cannot be said to belong to the agent.

There are three main weaknesses to this account. Firstly, Levy seems to employ reasons-responsiveness standards much stricter to implicit processes that to deliberative processes. Secondly, Levy has a very uncharitable view of implicit processes. And thirdly, Levy disregards the development of implicit processes.

Full awareness of all on-going processes (personal-level control), flawless receptiveness and reactiveness to reasons ('strong reasons-responsiveness'), and total coherency of all of one's reasons and attitudes (deliberative standpoint) do not hold for implicit biases, but they do not hold for conscious deliberative control either. Since such standards are too strong, Levy, rightfully, posits only moderate requirements for deliberation. However, he does not do so with implicit processes. Most explicitly, we can see this in Levy writing, "on the most optimistic story concerning our capacities for control over implicit attitudes and over their expression, even conscientious and well-informed agents utilizing the best strategies for controlling their implicit attitudes may rarely succeed *entirely* in bringing them under control." (Levy, 2016, pp. 6, my italics) Even in the case of deliberative behaviour, there are very few instances where an agent is *entirely* in control. Concerning deliberative behaviour, however, Levy invokes 'responsibility-level control'. Unless there is a good argument to justify holding different sorts of processes to different standards, we ought to judge all sorts of processes by the same standard, strict or moderate. Let us explore both those options.

---

hostile and hence inhibit acknowledgement of biases and motivation to work on them. (Saul, 2013, p. 55) I agree that *having racist biases* is not identical to *being a racist*, such that the latter judgement may often not be as justified. However, this argument mainly concerns *when we ought to express blame someone*, which is a different, albeit related question from *when someone is blameworthy*. A stance on blaming requires a distinct discussion of the psychology of persuasion and motivation, exploring, for example, the effectiveness of various, positive and negative, forms of expressing blame. I doubt the conclusion from this will be that expressing blame is never an effective strategy.

As for the first option, were volitionism to continue the strict evaluation of automatic processes and employ this standard to deliberative processes too, there would be very few processes left that are sufficiently under control, highly coherent, and strongly reasons-responsive. In effect, much of an agent's states and behaviours have to be discounted as not properly belonging to the agent, which damages the appreciation of the agent *as agent*. Levy may invoke the argument that the agent need not be disqualified as completely irrational when acting badly sometimes, but rather as having some 'islands of irrationality', "discrete beliefs or values which do not cohere well with the other beliefs or which fail to meet evidential standards to which they otherwise adhere, but regarding which they remain committed." (Levy & McKenna, 2009, p. 118) However, appreciating the significance of automaticity, and adding that much deliberative processes are islands of irrationality too, when evaluated by a strict standard, the lion's share of people's states and behaviours is alienated as non-agentive. In other words, under a strict standard of reasons-responsiveness, the agent falls apart into a disconnect group of islands. Note that this is exactly what original formulations of the automaticity challenge hold; agency (as strict reasons-responsive control) is rare.[51] The agent, as such, is no longer a reliable rational agent. Hence, with such limited agency, there is little basis to take someone serious as agent.[52] This implies what Peter Strawson calls the 'objective stance' towards a person, excluding one from the circle of moral agents, regarding one not as someone who can be reasoned with, but merely understood and 'managed'. (Strawson, 1993, p. 59) Also Smith has argued for the danger of such an outcome, writing that, beyond it being patronising and disrespectful, it damages the interaction opportunity of agents, since the other is "not someone with whom it is possible to enter into relationships of mutual respect and recognition." (Smith, 2008, p. 388) To conclude, the severely damaged notion of agency that a strict control condition implies, when applied to both deliberative and automatic behaviour, results in a theory of moral responsibility that champions a normative ideal that is practically

[51] I do not agree with this interpretation of the automaticity challenge, exactly because it employs this flawed and unnecessary strict control standard, among other reasons. See my doctoral thesis, essay 1.

[52] One strategy to meet this challenge is to invoke reasons to 'take on' responsibility, claiming ownership over them on other grounds than control. While I think this idea can be a very useful additional clause to an account, it does not solve its structural flaws and hence does not defend it well. Moreover, this move seems to be available to any type of account, as it is an addition. Mason, for example, argues for taking on responsibility for implicit biases that function without any involvement of attitude or will, as to appropriately respect one's fellow community members. (Mason, forthcoming, p. 12) Alternatively, an agent can take on responsibility for reasons of self-respect and status as agent.

useless to guide our moral practice, since a vast amount of cases in which there are strong intuitions regarding responsibility has to be alienated and excused.

The other option is to apply moderate control standards not only to deliberative behaviours, but also to automatic processes. This strategy becomes especially viable when invoking alternative and more charitable interpretations of automatic processes to Levy's, on which they come out much better that Levy purports. There are various such approaches available, which we can roughly divide in interpretations of current rationality, and developmental rationality. To start with the former, under a moderate standard of reasons-responsiveness, implicit biases can well be understood as having a sufficient propositional structure. For example, the sexist job interview case *does* show systematicity on a richer interpretation: The agent is 'epistemically cautious' regarding treating job-relevant skills of someone's résumé as all-decisive reason for someone's job suitability. Instead, the agent also considers further traits as relevant reasons to respond to, traits that this candidate is likely to possess less, since she is a woman, and (the agent's bias goes), women are typically less professionally competent. Now, there is surely a lot to say against the empirical (and moral) correctness of this process, but *given* the bias, the process *is* responsive to what are the relevant reasons. Similarly, a process of conscious deliberation is not less reasons-responsive when it produces the logical conclusion that follows from some false premises. Even the process of bias acquisition can be understood in this way. People do not just develop any arbitrary bias. Rather, throughout society a clear and consistent pattern can be discerned. This development is typically often reasons-responsive, but often receives incorrect information as data. For example, the predominantly negative media portrayal of women lacking professional capacities and attitudes provides reasons for the development of negative implicit attitudes towards women as professionally incompetent. The falsity of the data, again, does not falsify the way the data is processed. Similarly, a student whose teachers tell her the Dutch were the first to fly to the Moon does not fail to exhibit reasons-responsive learning when she consequentially develops a positive bias for Dutch technical universities. Fischer and Ravizza, on another volitionist account, argue for a similar alternative understanding of automatic behaviour, writing that while not driven by 'practical reasoning', automaticity can be thought of as being issued by some other kind of sensitivity, a mechanism which is nonetheless 'moderately reasons-responsive', since there is no requirement that it need be the deliberative mechanism that is driving it, as

long as it is reasons-responsive. (Fischer & Ravizza, 1998, pp. 85-88) Lastly, we can also recall the interpretation of automatic states by attributionist accounts. For example, Smith sees automatic states as embedded in rich inferential relations with other states, which, crucially, involves evaluative activity through an on-going, albeit unconscious, process of identifying, evaluating, and accepting something as a reason. As last example, there is Mason, who argues how implicit states are integrated with other states, which involves will.

Most importantly, beyond automatic processes typically being reasons-responsive, they can be controlled over-time, through controlling their development. Rather than merely focusing on the *current* control 'at time *t*', in-action in the moral situation, which one may indeed lack, we can also at the potential *developmental* control an agent has had earlier; the opportunity for development. For example, returning to Amir's case (at time *c*), Amir has had various ways in which he could have influenced the state or configuration of his biases, ways in which he has control over developing them, over-time. The fact that he does not have control in-the-moment, does not eliminate his foregoing developmental control. What this amounts to is that while a process may not be properly reasons-responsive in the moment, in so far as there was opportunity for development that was reasons-responsive, to bring a state in accord with others, the agent is responsible for it.

The problem, however, is that Levy's structuralist volitionist account cannot embrace developmental aspects, because of its focus on current control. As such, this strategy is only partly viable for Levy, including only responsibility for some automatic processes that can be interpreted as functioning reasons-responsively. This will nevertheless leave many automatic processes unaccounted for, regardless if development may be a viable possibility (e.g. Amir and Phineas' different developmental opportunity). Moreover, beyond not being able to account for character development, the focus on one's current state may actually incentivise laziness regarding self-development, since, regardless of the opportunity for development one has had, as long as one currently does not have control (maybe with the addition that the agent disavows the state), one is excused. Thus, to conclude, even on this strategy, a structuralist volitionist account is severely limited. However, there are other volitionist accounts that do attempt to open up to this approach, to which we will turn now.

3C: Volitionism and the extensiveness of character development

Now, there is a way for volitionism, as a historical account, to extend 'direct control' beyond the moral situation, by 'tracing' an agent's 'indirect control'. Tracing involves locating the responsibility for some present behaviour, where responsibility conditions fail, at an earlier point in time where the conditions *are* met, where an agent did have the right kind of control. Fischer and Ravizza, on whose work Levy draws, do endorse a historical account that even seems to embrace character development.[53] However, while I am very sympathetic to Fischer and Ravizza's work, and think it greatly improves the volitionist position, I will argue that even the indirect control strategy remains lacking, because the extensiveness of indirect control, on a volitionist account, is necessarily limited. I will focus on the epistemic condition, discussing Vargas' critique concerning 'foreseeability', followed by Fischer and Tognazzini's defence, and then argue that foreseeability is still problematic. Subsequently, I add two additional, related problematic features of the epistemic condition to it; the 'range of factors' one can develop (some specific knowledge, or moral character more broadly), and the 'behavioural and temporal demarcation' of development (one or several actions within a restricted time-period). To start, consider the following scenarios concerning various factors that we will consider as traceable or not.

Luis, a scenario presented by Manuel Vargas.

> *Luis drives home after having drunk alcohol at a bar. While driving in an intoxicated state he runs over a young mother and her two children. Luis cannot be said to be in control of his driving in the direct control sense. However, earlier in the evening, Luis did make an, albeit reckless, but nevertheless conscious and directly controlled choice of drinking at the bar, whilst knowing that he may be tempted to drive home drunk later on.* (Vargas, 2005, pp. 269-270)

Dr Naite's negligence, a scenario inspired by a similar scenario by Gideon Rosen. (Rosen, 2004, p. 303)

[53] Levy also endorses tracing and indirect control, but he sees much less potential in this than Fischer and Ravizza do. Therefore, I treat Levy as a more structuralist account in comparison to Fischer and Ravizza's more historical account.

A neurosurgeon, Dr Naite, who has been in the profession for several years already, regularly fails to read up on new publications in the field, due to no other cause than personal sluggishness, since the doctor possesses all the skills to be able to read relevant scientific literature, and is even provided with print-copies of the most recent relevant articles by the hospital's secretary. As an effect of this negligence, Dr Naite's knowledge is often not up-to-date with the latest research, while attending to that is a clear professional requirement due to the high degree of development in this relatively young branch of science. Subsequently, on an occasion when performing a medical procedure on a patient, it so happens that the neurosurgeon causes brain damage that could have been prevented had the doctor employed the new methods that were proposed in a recent, but already well-known and widely acknowledged article, which was, as usual, provided to the doctor by the secretary.[54]

The cowardly Captaine Le Chuiton.

Captaine Le Chuiton has been a professional soldier in the French army for over 20 years. In her function, Le Chuiton is rarely in direct action on the battlefield, as she prefers the comfort and safety of operations management from a distant encampment. In addition, at earlier stages of her career, she usually managed to get herself assigned to non-combat tasks. However, combat ability and leadership is an official prerequisite of her rank. On one particular mission, Le Chuiton finds herself in an unexpected combat situation that turns chaotic and dangerous. While there are several good strategies to deal with the situation in such a way that the Captaine would aid her subordinates while keeping herself safe, she only sees some of those. But even for those manoeuvres Le Chuiton lacks the calm and courage. Consequently, Le Chuiton resorts to take to her heels, merely bringing herself to safety, deserting the comrades under her command.

And again, as fourth case, we can recall the story of Amir (at time *C* and *D*), as described above in §2.

On a standard volitionist account of indirect control, the cases of Luis and Dr Naite are quite straightforward and uncontroversial. Due to intoxication, Luis may have been insensitive to moral concerns in the moral situation, while driving, but coming to

[54] A question to the reader: What do you imagine the doctor to look like? Note that the scenario makes no mention of the doctor's race or gender, hence, if you find yourself to have created an image of a white, male doctor, this could be an example of implicit biases you possess yourself and express while reading this essay.

be behind the wheel in a drunk state can be traced back to his earlier decision to drink, at which point he did satisfy the conditions of sensitivity to moral considerations, and thus should have taken the possible effects of his drunkenness into account. Therefore, Luis' lack of direct control does not excuse his blameworthiness, since he had indirect control. Fairly similarly, the Dr Naite's ignorance at the time of action diminishes her capacity or control at that time, but at the earlier times during which the neurosurgeon chose to not read the articles, the doctor did have full control such that Dr Naite should have been aware of possible downstream effects of such neglect. With that, the doctor has indirect control, and is thus responsible for incapacity due to culpable negligence. While indirect control is quite clear concerning Luis's inebriated state of mind and Dr Naite's ignorance of specialist knowledge, it is another question how to address the Captaine's strategic insight skills, emotion regulation, and virtues, and Amir's implicit biases. To analyse these cases, we need to further detail tracing indirect control. I will argue a volitionist account is unable to trace responsibility beyond fairly straightforward factors due to its fundamental commitments, which imply three features of tracing; discrete foreseeable effects, of (a) distinct action(s), within a restricted temporal period.

Besides the control condition, volitionism also employs an 'epistemic condition', and it is due to commitment to this condition that tracing is problematic. Fischer and Ravizza loosely propose the epistemic condition as follows: "An agent is responsible only if he both knows the particular facts surrounding his action, and acts with the proper sort of beliefs and intentions." (Fischer & Ravizza, 1998, p. 13) This condition need not be satisfied in the moral situation, or immediately prior to it, but may be traced to some earlier point in time. With a focus on tracing, Vargas more precisely defines the epistemic condition as follows:

> *"For an agent to be responsible for some outcome (whether an action or consequence) the outcome must be reasonably foreseeable for that agent at some suitable prior time."* (Vargas, 2005, p. 274)

The aspect of the epistemic condition that Vargas problematizes is the 'reasonable foreseeability'.[55] He does so by elaborating on cases of automaticity, or 'non-deliberative action', as he calls them; cases in which an agent unexpectedly finds

[55] For another critique on tracing satisfying the epistemic condition, see, for example (FitzPatrick, 2008).

oneself in a moral situation and one's subsequent behaviour is largely caused by one's character (or 'non-deliberative aspects of the agent', including traits, habits, and dispositions). For example, Paulina stands nailed to the ground while an alligator snatches her infant son in a nature reserve, and Jeff fires employees in an unnecessarily rude manner as a result of the girl-attracting jerkiness he embraced. The cases of Naite, Captaine, and Amir can be added to these. As Vargas argues, "the non-deliberative source of behaviour was acquired or retained under conditions where the agent could not have reasonably foreseen the later consequences of having that disposition, habit, or character trait." (Vargas, 2005, p. 275) Tracing responsibly to a moment at which agents can both act freely and at the same time reasonably foresee some particular future outcome is impossible in the case of non-deliberative behavioural sources such as dispositions and traits. At the prior moments of self-formation, people do not have the epistemic powers to predict the full range of downstream effects caused by some form of character. Not satisfying the epistemic condition, these agents are to be excused on a volitionist account. This is problematic, Vargas argues, since in real life there are many "instances of non-deliberative actions for which an agent is intuitively responsible, for which the agent fails to satisfy the knowledge condition." (Vargas, 2005, p. 279)

Fischer and Toganizzini refute Vargas' criticism of tracing, arguing that their historical volitionist account can embrace tracing for the factors he discusses and more, while honouring the epistemic condition.[56] Their main argument is that with a broader interpretation of the 'foreseeability' requirement of the epistemic condition, responsibility can be traced to the cultivation of one's character, which can be a reasons-responsive process. As for character development as reasons-responsive, Fischer and Ravizza already argue for in their original book, even employing virtue ethical language: "An individual might cultivate dispositions to act virtuously in certain circumstances. It might even be the case that when he acts virtuously, his motivation to do so is so strong that the mechanism is not reasons-responsive. But insofar as reasons-responsive sequences issued in his cultivation of the virtue, he can be held morally responsible for his action. It is only when it is true that at no suitable

[56] Fischer and Tognazzini distinguish their 'control account' as a broader account than volitionism, not merely focusing on choice, but also on action and omission. (Fischer & Tognazzini, 2009, p. 249) However, most authors discussed here understand volitionism similarly broad to such a control account. Therefore, I group them all together as volitionist.

point along the path to the action did a reasons-responsive sequence occur that an agent will not properly be held responsible for his action."[57] (Fischer & Ravizza, 1998, p. 50) And later they write, "Trait-actions issue from "thoroughly" nonreflective mechanisms. Nevertheless, our approach to moral responsibility treats them quite naturally in terms of the tracing principle. Recall that the general form of the tracing principle requires that, in order to be morally responsible for (say) an action, there must have been guidance control at *some appropriate point* prior to the action. An agent can be morally responsible for a trait action, then, insofar as there was guidance control in the formation, retention, or expression of the trait." (Fischer & Ravizza, 1998, p. 89) To this, Fischer and Tognazzini add that the epistemic condition "does not tell us just how finely the outcome in question must be specified", which may be very narrowly or more broadly. (Fischer & Tognazzini, 2009, p. 537) The foreseeable outcome need not be some specific action in a specific situation, but may rather be that the agent "treats some people poorly at some point in the future as a result of (the agent's bad) character." (Fischer & Tognazzini, 2009, p. 537) With that, we have an argument for why not merely 'current arrangement of an agents mental ingredients' matters, but also how this configuration came to be instantiated as such, since it is foreseeable that certain types of development will lead to certain broad types of behaviours.

While the view proposed in this paper is obviously highly in agreement with the correctness of this idea, I argue that it does not fit with volitionist conditions of responsibility, and as such, while correct, cannot be embraced within a volitionist account. To start, it is helpful to catalogue what types of factors Fischer and Tognazzini include in the things an agent can be responsible for developing, in order to reveal that the extent of it is limited and inconsistent (or, at best, 'unimaginative'). As a first indication for this, it is notable that when discussing Paulina's alligator fear, they mainly (and uncharitably to the hypothetical's point) focus on her stunned lack of control in the moment and wildlife danger negligence, without exploring an idea of developing a general sense of 'courage' or something such as a 'maternal protective courage'. Similarly, in the case of Ruben who fails catching the baby he throws up, they do not propose responsibility for developing general 'carefulness' or 'carefulness towards vulnerable loved ones. And in the case of Angela Smith forgetting a friend's

[57] More accurate would be not whether a reasons-responsive sequence *did* occur, but whether it *could* have occurred, whether the agent had the opportunity for development.

birthday, they do remain locked within her responsibility for taking precautions (such as a calendar reminder), not raising the option of habituating a more 'attentive attitude as friend' so that her friends are naturally on her mind. The discussion of these cases does not evidence much openness to a broad responsibility for character development. In two other cases, however, Fischer and Tognazzini do go into this. Regarding Jeff, who became a jerk, they argue that "We hold him responsible partly because he freely decided to become a jerk at some point in the past, and it is reasonable to expect Jeff's younger self to have known that becoming a jerk would in all probability lead him to perform jerky actions." (Fischer & Tognazzini, 2009, p. 538) And concerning George Eliot's *Scenes of a Clerical Life* personage Captaine Wybrow, who fails to notice his lover's food preferences, they even go as far as arguing for the possibility of tracing responsibility for becoming a careless lover. They write that Wybrow would be responsive "only if there was something in the past that he did freely (or omitted freely) that led him to be the sort of lover that fails to notice things that he should in fact notice. Perhaps he chose to engross himself in his own affairs rather than think about the needs and wants of Miss Asher, perhaps he omitted asking her certain questions about her emotional life when the opportunity arose, and so on. If there is nothing in the past that he should have done that would have made him much more likely to notice Miss Asher's likes and dislikes, then we would argue that he can't properly be held responsible. But what's more likely is that there were times at which his free actions (or omissions) more or less secured his emotional indifference to her." (Fischer & Tognazzini, 2009, p. 552) Becoming jerky or a careless lover seems a lot like character development in a very broad sense. However, this picture of character development does not satisfy the epistemic condition. Moreover, this picture of character development is too poor, and a more adequate picture fits the epistemic condition even less.

The first problem becomes clear in Fischer and Tognazzini's discussion of Jeff. They write that Jeff's awareness of his development, his self-conception of him becoming 'cool' while blind to the negative aspects is irrelevant, since all that matters is that "the fact that acquiring those characteristics will in fact lead to his treating others poorly and that Jeff should have expected that it might do so." (Fischer & Tognazzini, 2009, pp. 538-539) Now, firstly, just to note, awareness of oneself and of further consequences is a quite stringent capacity demand to make of a 15-year old, which Jeff is, which may well mitigate or even excuse Jeff's responsibility, so Fischer

and Tognazzini's judgement seems quite off here. More importantly, a probably more realistic interpretation of the scenario is that rather than one particular vice, Jeff is simply cultivating 'masculinity', as prescribed by social standards that he probably feels pressure to live up to. This interpretation of including self-conception and awareness is important, because masculinity does not simply translate to 'jerkiness' exclusively, but rather to a large *set* of attitudes and effects, for example including assertiveness, dominance, toughness, and being somewhat emotionally insensitive, which leads to being taken more seriously, getting jobs more easily, being respected by other guys, and, the teenager's main goal, being popular with girls – all besides other harmful effects. Such a broader understanding of character development appreciates how Jeff's development cannot as simply be characterised as concerning one particular vice with one particular type of resulting behaviour that he is responsible for. And this is where satisfying the epistemic condition becomes problematic, because 'foreseeability', then, has to involve connecting an extensive range of factors that an agent develops, many of which are interconnected with one another, to a diffused scope of effects that follow, many of which are influenced by several factors. Even on a 'broad' definition of foreseeability, in very many cases it will be impossible to even somewhat insightfully delimit one particular attitude development that influences one particular type of behaviour. This does not mean, however, that there *are* no such links; it merely means that the epistemic condition for *seeing* and *tracing* such links is too strict. As such, Vargas' critique, albeit somewhat differently detailed, still stands.

What is more, adding to this foreseeability problem, the epistemic condition implies two further problematic features; the focus on one or several distinct acts, and temporal strictness of these acts taking place within a restricted point in time. As can be seen in Fischer and Tognazzini's discussion of Wybrow, volitionist tracing has to find a particular '*something* in the past that he did (or omitted)', such as that *one* opportunity for a personal question, or that specific choice to engross in his own affairs. Or, as we see in the citations of Fischer and Tognazzini above, foreseeability has to be located as '*a* reasons-responsive *sequence* at a suitable *point* along the path to action', and 'guidance control at *some* appropriate *point* prior to the action'. Even on a best-case reading of volitionist tracing, at best there are *several* distinct choices, decisions, or omissions, within a certain temporally confided period have to be marked out. Unfortunately, this is typically not how character development, or even

the cultivation of a certain attitude or emotional disposition happens. Rather, character development involves a plethora of undertakings that take place over an extended period of time. For example in Wybrow's case, reflecting on what is important to his happiness, what other people mean to him, what kinds of relations he wants to have, and what kind of person he wants to be; discussing with friends what other people find important to be attentive to; effortfully focussing on his lover's cares; habituating an attentive attitude by repeatedly being extra attentive; training to become more observant in general, to various affairs of the heart of various people; and cultivating dispositions to continuously work on all of the above. Clearly, such an undertaking is not the type of concrete 'unit' of evaluation, a distinct choice within a distinct period of time with distinct effects that satisfies the epistemic condition. The same tracing limitation to fairly concrete indirect control forms can be seen in the volitionist account proposed by Jules Holroyd and Daniel Kelly, who, while advocating character responsibility, nevertheless focus on onstage direct moderative control (e.g. intervening in operant automaticity) and onstage indirect moderative control (e.g. cognitive preparation and environment selection). (Holroyd & Kelly, 2016) While they do discuss character development, this seems to be a secondary consequence of the behavioural guiding strategies they mainly focus on.[58] Such a limited reach of indirect responsibility, found with Holroyd and Kelly similar to Fischer and Tognazzini, is an unavoidable implication for all volitionist accounts.

It might be possible in some instances to trace responsibility for some particular attitude or a specific implicit bias while satisfying the epistemic condition, but that is hardly a sufficient model for the working of most automatic or characterological behaviour in any interesting way. An accurate account of character includes a wide range of factors that are developed in interaction with one another, the development playing out over an extended period of time, and various characterological factors jointly influencing a diffused range of behaviours. Such development cannot be traced while satisfying the strict type of volitionist epistemic condition, which requires discrete foreseeable effects, of (a) distinct action(s), within a restricted temporal period. With this analysis of tracing in hand, we can now see how the scenarios presented at the beginning of this section present an increasingly difficult feature for volitionist accounts to appreciate. Luis' intoxicated state of mind

[58] Holroyd and Kelly's account seems less deeply committed to volitionist principles however, such that I can imagine them embracing developmentalist ones instead.

may be traced, just as Dr Naite's specialist knowledge. Possibly we can delimit Amir's implicit bias sufficiently, understanding it as a distinct attitude, unconnected to other attitudes, which can be developed with certain distinct efforts within a relatively limited time frame, and also influencing only a specific class of behaviours. This picture is not very plausible already, so this is probably as far as a volitionist account, on a good day, can reach. But when we look at the Captaine's scenario, there are various factors involved (emotional dispositions, skills, and attitudes), which all interact with one another, which are developed through a wide range of different types of developmental efforts (e.g. strategic exercises to improve envisioning behavioural opportunities; becoming familiar with seeing pain, blood, and fatality; intensive combat practice; and psychological training), which occurs over an extended period of time, and which can influence many different behaviours in moral situations (e.g. seeing an opportunity; staying calm; having the courage; being able to communicate plans to comrades; taking leadership; and other such behaviours that Le Chuiton may be held responsible for).[59] With this, we have a case where it seems clear that the Captaine is responsible, while tracing indirect responsibility is impossible while satisfying the epistemic condition. As such, volitionism, with its strict epistemic condition, is unable to account for responsibility concerning a wide range of automaticity, which is a severe limitation, given the significant automaticity thesis. The one solution for volitionism to succeed would be to change the epistemic condition. One strategy is a r*efinement* of the epistemic condition. Through discussing Carl Ginet's proposed 'K' as an example of a more delicately defined epistemic condition (Ginet, 2000, p. 275), Vargas argues that, since it involves a necessity for accurate specification of all the conditionals, such projects seem impossible. (Vargas, 2005, p. 290) The other strategy may be r*elaxing* the epistemic requirement.

[59] Some readers may be unconvinced of this particular scenario, or hung-up over it being a case of professional rather than moral responsibility, but little hangs on this since copious other similarly complex everyday cases exist, such as a mother's courage, or a boy's brusqueness, among other cases discussed here. There is a vast amount of related cases that we could discuss here, for example, besides brain damage cases, we can invoke brain-manipulation by a nefarious neuroscientist (Arpaly, 2003, pp. 165-167), or a spontaneous mental illness that permanently changes an agent's character. Instead of JoJo's dictatorial conditioning, it could be Solomon's secluded sexist community (Arpaly, 2003, pp. 103-104), the real case of Dominic Ongwen who was abducted and indoctrinated by The Sinia Brigade of the Lord's Resistance Army in Uganda, or Patty Hearst who was kidnapped by the Symbionese Liberation Army and brainwashed into supporting the cause and committing terrorist acts (Arpaly, 2003, p. 166), and the traumatised mass-murderer Robbert Harris (Watson, 1987, pp. 268-275). And instead of Amir's social environment, we can think of Abigail who is raised in a racist and religiously intolerant environment (Smith, 2005, p. 267), Herbert Greenleaf in Patricia Highsmith's novel The Talented Mr Ripley replicating the status quo sexism of the 1950s (Fricker, 2007, pp. 82-89; Mason, forthcoming, p. 16), an ancient slaveholder (Rosen, 2003, p. 64), a ruthless capitalist Mr Potter (FitzPatrick, 2008, pp. 599-611), among countless others.

However, this is not an available option for volitionist accounts either, because volitionism is by definition married to a strict epistemic condition. The reason for this is that volitionism, at its core, is committed to the focus on direct control of a moral action, and other forms of behaviour and responsibility may sometimes be included when suitably interpreted within that framework. Here we can return to discussing a developmentalist view as alternative, because in contrast to volitionism, we saw that developmentalism has a very different focus, mainly focusing on developmental opportunity, a process of general moral education, of various character factors, over a long period of time, that takes mainly takes place outside of moral situations, which then affects a range of moral behaviours – with 'direct control' or 'deliberative behaviour' as a less frequent responsibility. Were a volitionist account to shift its focus to a somewhat similar idea of indirect control and general development, this would be such a radical change of the core principles, I do not see why it would still be useful to maintain calling such an account 'volitionist', since it would have no elementary similarity with any of the main accounts that represent volitionism. Similarly, rather than *expanding* indirect control for volitionism, possibly an attributionist account could *incorporate* some notions to embrace character development. However, this attributionist future seems even less viable than the volitionist one, as it involved entirely new notions that it aims to avoid, as discussed earlier. As such, developmentalism can be seen as a distinct third alternative to the two traditional theories of moral responsibility, possibly similar to how virtue ethics is a distinct third view in normative moral philosophy. Not much hangs on this point though, for rather than aspiring towards a distinct alternative, if proponents of other, already existing accounts are able and willing to incorporate and centralise the developmental conditions championed here, the aim of this paper is achieved.

§4: Developmentalism II

To conclude, one remaining issue is the positive specification of an alternative epistemic condition, which a developmentalist view subscribes to in order to trace developmental responsibility for an agent's automatic behaviour. A precise

formulation will have to evolve and mature over the course of future applications of and debate on developmentalist approaches, in collaboration with others. Nevertheless, as an initial and tentative approximation, consider the following *character condition.*

> *For an agent to be responsible for some outcome that is caused by one's automatic processing (whether an action, omission, or consequence), the agent must have had the opportunity (including capacity and environment) to actively engage in the development of their own character (including knowledge, values, skills, traits, and attitudes) to configure it such that one can reasonably foresee one's character influencing good outcomes in a wide range of moral scenarios.*

A first note on this definition is that it concerns the conditions of when an agent is responsible. Nevertheless, as mentioned earlier, a developmentalist view appreciates that the phenomenology of action and the connected initial phenomenology of responsibility judgement is somewhat as attributionism understands it. Initially, we judge an agent as responsible for some behaviour because the behaviour comes forth from one's character, and as such is an expression of one's character. For example, when Amir (at time *C*) bluntly interrupts a female colleague, he is judged as blameworthy for this since this behaviour is issued by his character. Amir is judged as being 'such a type of person who does such types of things'. It is only in the background of this evaluation that developmental opportunity functions. And it is only when the judgement is more important, when it is being scrutinised, for example when Amir is asked to defend his behaviour, that we really have to inquire into the narrative of an agent's developmental opportunity, through which we can discover features that decline the agency that an agent has had over one's character, and thus mitigates (or excuses) the responsibility an agent has over the behaviour.

Hopefully this addition of a phenomenological responsibility practice speaks to a worry one may have about the ability of a developmentalist view to cast quick and clear judgements of an agent as responsible for this action right here. That said, I do picture these initial judgements being nevertheless quite sophisticated already, involving a rich understanding of the agent, for example of Amir being a 'somewhat

harsh and traditional, but well-intending and self-critical lad, who does lovely in his behaviour over-all, especially given his tough background', whereby the initial judgement can already involve a nuanced judgement of him being responsible only to a certain degree, rather than a binary yes or no.

A second note on this character condition definition is that it aims to include the various features of a developmentalist view that have been discussed throughout this paper. Appreciating these features all together, this condition is at the same time looser and stricter in its demands, in comparison to volitionist and attributionist accounts. As a rich notion of character is employed, this greatly expands the range of factors that agents are responsible for. Furthermore, as that behaviour is taken to be mainly produced by one's character, responsibility is primarily related to character development. This makes the account, in traditional terms, an account that centralises *indirect responsibility*. Or better yet, to be more precise, since indirect control is limited (as argue above), and character development is understood as mainly occurring *offstage*, in detached training situations, the account is rather one of *developmental responsibility*.

With that, *moral agency* is the active engagement with one's character development, which indicates intention shown by the agent through effort such as difficulty and sacrifice. However, since such offstage agency requires consciously choosing certain behaviours or selecting certain environments, agents do also have *direct control*, which they can occasionally exhibit *onstage* in moral situations too. Furthermore, *foreseeability* links a wide range of characterological factors to a wide range of behaviours. As such, character development requires broad character development of agents. This development may take place over an extended period of time, including a range of minor and weightier events. Lastly, *opportunity* involves an agent's capacity as well as environmental circumstances, appreciating agency limitations due to one's diminished capacity as well as one's restricted environment as mitigating or even excusing, explaining and sometimes challenging notions of moral luck.

Finally, a third note on this last feature. One of the aims of a developmentalist view is to appreciate not only 'ideal rational agency' factors, but also real world factors that influence moral agency. Crucially, factors such as socioeconomic class, race, sex,

among others, influence not merely how people are treated, but influence people's behavioural opportunity, including available courses of action, but also including one's development, among which one's functional and even structural cognitive development. Embedding theory of moral responsibility in environmental circumstances and moral psychology is not 'merely' an anti-classist, anti-racist, and feminist project, but it just as well a more empirically realistic project. The case of Amir was intended to display how various factors can be appreciated, being more understanding for various forms in which an agent can be limited, and being more critical for various ways in which an agent can be responsible.

To conclude, one last telling example of how a developmentalist view can produce more nuanced and detailed evaluations of an often-discussed agent who typically gets off the hook. Herbert Greenleaf, a personage from Patricia Highsmith's *The Talented Mr. Ripley*, possesses sexist attitudes and exhibits sexist behaviour, through dismissing the (actually correct) suspicions of Marge about who the murderer of his son Dicky is, saying, "Marge, there's female intuition, and then there are facts." Strikingly, even feminist philosophers judge Greenleaf as non-culpably ignorant, and hence not responsible for his behaviour, due to his cultural-historical context, the difficulty of self-awareness, and the difficulty of character development. Miranda Fricker, for example, argues that the "critical concepts he needed were not historically available to him." (Fricker, 2007, p. 101) Elinor Mason similarly argues that Greenleaf's blameworthiness is undermined, firstly by lacking bad will, and secondly by the unavailability of evidence for the falsity of his views, which itself is also not through bad will but through circumstantial isolation. "Greenleaf is just a man of his time, and he is trapped in his time. It seems that it does not make sense to say of Greenleaf that he should take responsibility for his does not make sense to say of Greenleaf that he should take responsibility for his mistakes, because he is not even remotely aware of the possibility that he is making that sort of mistake. (…) "Ideally he would come to see things differently, of course, but if we imagine him in the historical context he is in, he does not have enough distance from his own oppressive actions and tendencies." (Mason, forthcoming, p. 16) Expecting Greenleaf to see and engage with his flaws of character and behaviour is, according to Fricker and Mason, 'hubris', as one cannot be expected to move beyond the orthodoxy of the day. Moral development beyond one's environment, overcoming one's social conditioning, can only be achieved by 'geniuses', they argue, as it requires "exceptional, more

imaginative moves in which existing resources are used in an innovative way that stands as a progressive move in moral consciousness." (Fricker, 2007, p. 104)

In reply, I argue from a developmentalist view that such evaluation is defective for several reasons. Firstly, Greenleaf lives at some point during the 1950s, in New York, is fairly intelligent, educated, cultured, and a wealthy shipping magnate. If there existed any intelligent or otherwise talented women in those days, Greenleaf seems to move in the milieu where one is likely to run into them, which would provide him with counter-evidence to his sexist attitudes. Now, feminist projects of re-appreciating women in history exactly focuses on this fact, that the existence of competent women is not a contemporary novelty, but has occurred throughout history, whereby it is very likely that Greenleaf was exposed plenty of such women. Besides counter-evidence, critical concepts and challenges to the suppression of women were available as well, exemplified by, for example, the publication of Beauvoir's Second Sex in the previous decade, and the nineteenth amendment giving voting rights to every citizen regardless of sex in the USA in 1920. Crucially, these events were not isolated, standing on their own, but originate from a culture of discussion, ideology, and activism regarding women's equality. Unlike the popularised historiographical story that conceptualises the dominant post-war ideology as one that conservatively domesticized women, revisionist research by scholars such as Joanne Meyerowitz evidences that public culture at the time was rife with advocating women's individual striving, public achievement, addressing issues of gender, and supporting political participation: "Just as women's activities were more varied and more complex than is often acknowledged, so (…) was the postwar popular ideology." (Meyerowitz, 1993, p. 1480)

As such, Greenleaf, through his many privileges, was surely in a position that justifies judging him as blameworthy for his lack of awareness, lack of development, and bad behaviour. Not doing so, puts the bar for epistemic, environmental, and capacity so low that even most people today would be blameless for their harmful attitudes, since still today this is deeply entranced in our upbringing and culture, and while there may be some more knowledge available to some, character development will always remain difficult. Moreover, if only extremely exceptional individuals would be able to move beyond current morality, it is difficult to envision how collective moral progress would ever get off the ground, since more than just a few

geniuses are required to create a culture that challenges and changes its moral values and structures.

It is rich, nuanced evaluations of moral responsibility as the one above, taking into account an agent's background, including one's rational capacities and environmental circumstances in relation to the opportunity to develop one's moral character, that merits the developmentalist perspective.

Discrimination in the Bedroom

Sexual Preferences, Character Development, and Moral Responsibility

Abstract: *Sexual preferences, preferences concerning sexual or romantic partners, can target traits such as race ("no Blacks"), sexual orientation ("no bisexuals"), gender identity ("no effeminate"), physique ("no fatties"), and class ("no penniless"), among many others. Such preferences are equally prevalent as they are morally thorny, since similar selection criteria are commonly deemed discriminative in other matters. This essay has two objectives. Firstly, as the topic is underexplored in academic moral philosophy, I expand the debate beyond the focus on racial preferences to address sexual preferences in general, provide a systematic overview of the main arguments defending and criticising sexual preferences, and explore the models of sexual cognition and theories of moral responsibility that are implicit in these arguments. Secondly, I propose an understanding of sexual preferences as attitudes in an agent's character, which are interconnected with other attitudes, and which allow for malleability. From this, I argue that we can best analyse the moral status of sexual preferences by appreciating the various harms they can cause, and evaluate agents' moral responsibility in terms of the developmental control they have over their preferences.*

Introduction

"*And I'm loving all races,*
Hell nah, I don't discriminize."

- Quavo[60]

Recently, a debate has emerged about the morality of certain sexual preferences a person may possess concerning the type of sexual or romantic partner one prefers. In many cases, sexual preferences concern the race, sexual orientation, gender identity, physique, or class of a person, among others traits. Simultaneously, common moral and political theories typically consider selection on the basis of these same traits to be discriminative in other matters, and impose restrictions on selection practices of forms such as "no Blacks", "no bisexuals", "no effeminate", and "no fats". Hence, a shadow of moral suspicion and contentiousness is looming over sexual preferences of exactly such forms, concerning traits that otherwise constitute discrimination. With that, such, what I call *controversial sexual preferences*,[61] are equally prevalent as they are morally thorny, and the question at hand is how such sexual preferences are to be morally evaluated.

Discussion about the morality of controversial sexual preferences has been almost entirely confined to public debate, with numerous articles, vlogs, and now even a television show on the topic. (Brinkhurst-Cuff, 2017; Chu, 2017; Mosbergen, 2016; Tamanna, 2016) And while there has been considerable empirical research on the existence, prevalence, and harms of controversial sexual preferences, it is strikingly underexplored in academic moral philosophy. Moreover, the handful of existing ethics literature on the topic is mostly restricted to discussing racial preferences and assessing the moral wrongness. (Coleman, 2011; Emens, 2009; Halwani, 2017; Thomas, 1999; Zheng, 2016)

[60] Young Thug and Travi$ Scott featuring Quavo (2016) *Pick Up the Phone*, on Travi$ Scott, Birds in the Trap Sing McKnight (CD), Grand Hustle Record, Atlanta, U.S.

[61] Such preferences are, for now, 'merely' morally controversial, as some raise critical moral concerns regarding them. They are not (yet) evaluated as 'bad', 'immoral', or 'discriminative'. Rather, this is the topic of the essay. Some controversial sexual preferences may thus turn out to be *discriminative sexual preferences*, but not necessarily all.

Hence, the first purpose of this paper is to bring together the prevailing arguments and represent them in a detailed and systematic manner, as this is absent in the academic literature. Secondly, I aim to expand and generalise the debate, beyond racial preferences to regard all types of sexual preferences as a conceptual category, and beyond the evaluation of moral wrongness to address moral responsibility in addition. The third way in which I aim to advance this debate is by making explicit the models of sexual cognition and theories of moral responsibility that are implicitly employed as foundation of any moral theorising about sexuality. Ultimately, the main objective is to propose (i) a dynamic model of cognition: an understanding of sexual preferences as attitudes in an agent's character, significantly interconnected with other attitudes and with significant fluidity, and (ii) a developmental control theory of moral responsibility: an understanding of how controversial sexual preferences cause direct and indirect harms, function in perpetuating structural discrimination, for which agents have moral responsibility in virtue of their potential for agentive control over preference development.

After introducing the phenomenon and central concepts (§1), I set out the 'classical liberal' view, which is based in a classical view of the mind, and champions the liberal ideal of personal and sexual freedom, thus defending controversial sexual preferences as morally unproblematic (§2). Subsequently, I discuss sexual cognition and go on to refute classical liberalism by proposing the 'dynamic developmental' view instead, with which I evaluate if various types of preferences are discriminative (§3). Finally, I briefly discuss several further concerns (§4).

1: Sexual preference and discrimination

In this section, I identify in greater detail what the phenomenon of controversial sexual preferences involves. Firstly, I clarify how I understand and use the concepts (a) 'sexual preference' (distinguished from 'sexual orientation' and 'sexual identity'), and (b) 'discrimination' (distinguished as normative evaluation from mere 'differentiation'). Finally, (c) I make the phenomenon more concrete by presenting a

variety of exemplary cases and elaborating on preferences concerning race and sexual orientation in particular.[62]

1A: Sexual preference, sexual orientation, and sexual identity

'Sexual preference', 'sexual orientation' and 'sexual identity' are defined in many different ways by various authors, and often even used interchangeably, with little consensus on any particular definition or distinction. The way that I employ these concepts here is as follows, partly based on the American Psychological Association's Committee on Lesbian and Gay Concerns. (APA, 1991) 'Sexual orientation' concerns the sex or gender that an individual is attracted to, for example heterosexuality or homosexuality. 'Sexual preference' is distinct from sexual orientation in that (a) sexual preference concerns further traits of an individual's partner attraction beyond sex or gender, for example concerning appearance, shared interests, activities, type of humour, or shared beliefs, (b) sexual preference concerns attraction to certain sexual activities, and (c) sexual preference can have a significant degree of intra-individual variability over time, as a person's sexual preferences are fluid and can change (this will be fully addressed later). Furthermore, by sexual preference I do not merely refer to 'sexual' partners, but to a wide range of interactions such as attraction, desire, love, dating, romance, marriage, and other intimacy relationships and partnerships.[63] Finally, 'sexual identity' can be understood as relating to how an individual experiences one's sexuality (inner sexual identity), organises their sexual life (practical sexual identity), and outwardly expresses oneself (expressive sexual identity), for example feeling bisexual, having sex with men and women, but publicly identifying as heterosexual. As we will find, while these three concepts overlap a lot, the distinction between them is pivotal in evaluating sexual discrimination.

[62] Throughout this essay, there is an increased focus on experiences and norms that are common to Western, White-majority societies. This is not because of any evaluation of superior normativity, but merely because of the author's familiarity with the literature, norms, and (own) experiences. As such, while I do attempt to include other perspectives, as seen in the anecdotes and used literature, I am aware that some parts of the essay may well be non-universal. Nevertheless, the championed cognitive model and evaluative methodology is aimed to be applicable to alternative environments that provide other input for it.

[63] While there are undeniable differences between these concepts that makes each distinct, I necessarily discuss them together due to understanding them as jointly forming an inseparable conceptual family of closely related phenomena that have mutually influential interactions with one another, as will become clear from throughout the essay. From hereon after, when writing 'sexual preference' or 'sexual partner', I refer to this whole family of intimacy relations, unless specified otherwise. A possibly better term may be 'intimate partner preference', but as this term is hardly used in the public debate, this seems unfitting.

1B: Preferences, differentiation, and discrimination

Another crucial distinction concerns the normative evaluation of preferences. In the most elementary form of the term, a preference is a form of 'differentiation', the differential treatment of one thing compared to another on the grounds of certain specific traits. This is to be distinguished from a preference being a form of 'discrimination', a morally objectionable form of differential treatment. All preferences involve differential treatment, and are, thus, forms of differentiation. And all discrimination is likewise a form of differentiation. But not all differentiation is discriminative, as discrimination is a subcategory of differentiation, and discriminative preferences are a subcategory of preferences. In order to be able to evaluate which preferences, if any, may be discriminative, it is important to distinguish between different ways in which preferences differentiate, and specify what the conditions of discriminative forms of differentiation are. In short, discrimination is a particular form of immoral differentiation, with on-going history of systemic oppression of a group of people.

A first distinction we can make is between '*moral* differentiation' and '*non-moral* differentiation'. This distinction merely relates to some differentiation being a moral matter at all or not, but not, in addition, to the moral evaluation of it. Differentiation is moral when it relates to the good of other individuals.[64] For example, someone's preference for peeling a banana from the bottom, rather than from the top, because of the enjoyment of feeling rebellious, without any further effects, is a non-moral preference, since it does not relate to the good of other individuals and thus does not have any obvious moral ramifications. Preferences regarding human social affairs do relate to other individuals, and are, thereby, moral. Sexual preferences, being concerned with the attraction to and selection of one's sexual partner, are hence a sub-category of human social affairs, and thus forms of moral differentiation (the distinction between moral and non-moral is not always as sharp, but I trust it is clear enough for our current purpose). (Stember, 1976)

[64] There is, to my knowledge, no one clear and widely-accepted definition of descriptive morality, but I take the crude characterisation made here to be sufficient for the purpose of distinguishing moral from non-moral matters. 'The good' can indicate a wide range of goods, for example justice, equality of opportunity, and pleasure and the absence of harm. 'Other individuals' can be taken narrowly, for example other human beings that belong to one's group, or broadly, indicating all sentient beings, and can, for example on virtue ethical views, include the agent itself.

Within moral differentiation, we can make a further distinction, between '*justified* moral differentiation' and '*unjustified* moral differentiation'. Moral justification relates to defending some differentiation that affects the good of others with reasons in the face of morality as being morally acceptable. For example, a preference for a job candidate because of their job-relevant skills is a justified form of moral differentiation, since, while it does relate to goods of other individuals, it does so on the grounds of features that can defended with reasons (e.g. the inequality of opportunity in this case may be warranted because of the need of the labour being done best as possible, and exclusion does not constitute a more general source of oppression for the ones disfavoured). If a preference is not defended as such, the harm it causes to the good of others is arbitrary and unjustified, and hence it is a morally condemnable, or simply immoral, preference. If, in addition, this harm functions in an on-going history of systemic oppression of a group, the preference is discriminative.

Some traits of individuals are understood by common moral and political views as having a special protected status whereby they require extra rigorous justification in order to be employed as the ground of differentiation, because historically these traits have been, and often still are, employed to systematically oppress groups of people. These 'protected personal traits' include, for example, race,[65] ethnicity, sex, sexual orientation, and gender identity, among others. In some cases, differential treatment based on protected personal traits can be justified. For example, the selection of a security guard partly for his being male, because the job includes frisking people, and this is to be done by someone of the same sex as the subjects. Or, for example, granting admission to a discussion group only to homosexual people with a foreign background, because of the need of a 'safe space' where people with similar lived experiences can be and talk openly. These are cases of moral differentiation, but justified. In contrast, an example of an unjustified, discriminative preference is a teacher disfavouring a student due to the student being Black and thereby awarding lower grades to the student's assignments, since racial background is irrelevant to the evaluation of the quality of the student's worth and work, hence causing unjustified harm to the student, which, in addition, feeds into the systemic oppression of people of colour.

[65] While many (rightly so) consider the concept of 'race' to be intellectually and scientifically void, I employ it as done in everyday language and, as such, typically use race and ethnicity interchangeably, with 'racism' referring to discrimination based on racial and ethnic features. See (Lentin, 2011; Marger, 2012).

We now have the basic concepts to formulate more precisely the two central questions of this essay. (i) Moral wrongness: Are some controversial sexual preferences not mere forms of differentiation, but rather forms of sexual discrimination in virtue of causing unjustified harms that function in the systemic oppression of a group? (ii) Moral responsibility: On what ground can people be held morally responsible for discriminative sexual preferences?

1C: Controversial sexual preferences

To substantiate controversial sexual preferences more as a phenomenon, exhibit their wide range of variety, and reveal some of their intricacies, I now present several exemplary cases and subsequently elaborate in more detail on two of them, concerning race and sexual orientation. Note that the exact description of the following serves only as illustration; nothing in the arguments of this essay hangs on any of these particulars as long as one acknowledges the existence of the general phenomenon. Furthermore, the following is presented, for now, as *controversial* sexual preferences, as they are the topic of moral suspicion and contention, but they are not yet evaluated as *discriminative* sexual preferences, which will be the topic of the rest of the essay.

> *"I experience that, being an Asian man, I have a disadvantage in finding a partner. Often people give me the feeling that they either strongly prefer or dislike Asians. When someone likes Asians, that's fine for getting sex, but not for anything more than that, because only rarely does it feel like someone looks at my personality. They only see Asian… Sometimes I imagine that, if I were White, I would have had a boyfriend by now."*

> *"Several times I have been confronted with very painful reactions to being bisexual. I have had wonderful in-depth chats with people, really connecting to each other. And then when I mention that I am bisexual, they will say "O I don't do bisexuals". They just write me off just on my sexual orientation, and whoever I am as a person or how we connected doesn't matter anymore. Because of that, I now sometimes lie to people, telling women I'm straight and guys I'm gay, so I can get laid. But I feel bad about it, because I'm not honest, not to them, but also not to myself."*

"Over the years, I have become very wary of people and I've had to put up a high wall around myself, to keep people at a distance and really check if they are okay or not. I mean, if they just exotify me, or if they can see me. For example, when chatting online people often just presume I'm into rough sex, telling me they want my 'big black cock', and expecting that I can twerk – as if that's a special magical ability that every black person has. Or when I'm out, people can overdose me in 'compliments', 'charming' me with how soft and black my skin is, or that I'm "actually ok". When I hear that, I think, "there you go", I got the sign I was expecting. It makes me very uncomfortable and stressful to interact with people. And when I call someone out on it, they typically get really aggressive and suddenly call me a nigger."

"On my dating profile, I don't use a clear picture anymore. I have lots of fun chatting with people, and often have very long chats actually. And when people eventually ask for a better picture, I just sent old ones. I know that it's cheating, but it feels like I have to, because otherwise I don't even get a chance to prove to people that I'm attractive in so many ways through my personality. And when we meet up, I just hope that they will like me enough to accept me, and accept that I'm fat."[66]

Preferences concerning race: One trait that sexual preferences can relate to is race. Such preferences are held by all sorts of people. Women exhibit stronger racial preferences than men. (Fisman et al., 2008) But especially among homosexual men this phenomenon is widespread, often very explicit, and well researched, hence I focus on this. (Phua & Kaufman, 2003) Moreover, on online dating platforms for homosexual men, such as *Grindr* and *Hornet*, it is common that a person's profile description, which specifies who the person is and what he is looking for, often explicitly expresses racial preferences. For example, reading "no Blacks, Latinos, Middle-Easterners, or Asians, but I'm not a racist." Such explicit expressions of racial references fit well with reports from people of colour on the large amount of implicitly experienced racial preferences, which are typically difficult to verify, for example feelings of being treated differently in a gay club.

Sexual preferences come in a variety of forms. Firstly, they do not merely relate to attaining sex, but also to being considered sexually attractive or desirable. (Coleman, 2011, p. 13) Secondly, sexual preferences can be manifested with different

[66] These four anecdotes are gathered from informal interviews I had with people who have experienced sexual discrimination in order to explore various perspectives of the phenomenon through the eyes of the targets. They aim to illustrate the multitude of ways in which sexual preferences can be manifested, relating to features like race, sexual orientation, and physique. The anecdotes are not used as data for the essay, personal information is anonymised, and usage is approved by the Norsk Senter for Forskningsdata (NSD) on 2017-06-25, 13:25h, Oslo.

'directions' and 'strengths', which we can categorise as a range of concepts as follows (adapted from (Halwani, 2017)): 'Exclusion' (strongest repellent: precluding a group), 'aversion' (medium repellent: disfavouring a group), 'indifference' (neutral: no consequential attitude towards a group), 'predilection' (medium attraction: favouring a group), and 'exclusivity' (strongest attraction: solely appreciating one group). The dating-profile example right above is a form of exclusion, in which a person is ruled out as partner based on their racial group. In contrast, the exotification described in the third anecdote is an example of favouring someone based on racial traits. Although the latter forms are sometimes called 'inclusive' or 'positive', these terms merely indicates the direction of preference compared to 'negative' ones. The racial proviso targeting is often not experienced positively in a normative sense, but rather as objectifying and stereotyping, and hence also dubbed 'exploitative' or 'fetishistic'.[67]

Racial preferences often involve categorising people as a 'sexual type', a stereotypical construction of someone's sexuality associated with race. (Caluya, 2006) One well-known racial fetish is so-called 'jungle fever', the sexual preference for Black people because of their supposed hyper-sexuality, including having a sizeable penis and being very sexually aggressive (see the anecdote above recounting the sexual roughness that is assumed without any regard to what the person is actually like as an individual). This Black sexual type can explain, for example, why Black men who identify as 'top' (the insertive, rather than receptive partner in anal sex) are much popular than those who identify as 'bottom'. (Robinson, 2007) In contrast, 'yellow fever', the fetishizing of Asian people, often involves stereotypes of being submissive and effeminate for men, or docile and obedient for women. Much has been written about how racial sexual types relate to sexual marginalisation and influence sexual preferences. (Ayala et al., 2012; Han, 2006a; Han, 2006b, 2007; Malebranche et al., 2009; Ridge et al., 1999)

[67] As a third feature, we could add higher-order attitudes, the attitudes a person can have about one's sexual attitudes; acceptance, indifference, rejection, and unawareness. These are not strictly speaking part of the sexual preferences themselves, so I will not discuss this further now, but will come back to it later when evaluating moral responsibility.

Attitude type	Direction	Strength	Description
Exclusion	Repellent	Maximum	Precluding a group
Aversion	Repellent	Medium	Disfavouring a group
Indifference	Neutral	Neutral	No significant attitude
Predilection	Attractive	Medium	Favouring a group
Exclusivity	Attractive	Maximum	Solely appreciating one group

(Table 1: *Categorisation of sexual preferences*)

Preferences concerning sexual orientation (and gender identity): Another trait that sexual preferences can relate to is sexual orientation, and features that can be connected to that like gender identity. While women are found to increasingly exhibit more 'fluid' sexual orientations and take on more fluid gender identities, meaning that they increasingly experience sexual or romantic attraction to other people than merely those of the opposite sex or gender, and increasingly identify and express their gender identity in ways that transgress traditional categories and norms of 'womanhood' or 'femininity', the majority of those same women nonetheless systematically disqualify men that have had homosexual intercourse as non-suitable romantic partners. (Tsoulis-Reay, 2016) A variety of cultural ideas about bisexuality has been found to ground excluding bisexuals as potential partners. (Gorna, 1996; Weinberg et al., 1995) For example, relatedness to homosexuality makes men seen as less masculine. (British Psychological Society (BPS), 2014) And being emotionally transparent men are also disfavoured as partner by women. (Babel et al., 2014; Brown, 2012) Besides that this phenomenon shows that the liberalisation of sexuality and identity seems to be distributed differently between sexes, it constitutes another type of sexual preferences that are morally controversial.

What is interesting for our discussion here is the working of sexual preferences in relation to sexual orientation and gender identity. To conceptualise the phenomenon, we can start with the idea that as sexual orientation selects potential partners based on sex or gender, people can have 'basic sexual orientation compatibility'. For example, a heterosexual woman and a bisexual man have basic compatibility, as they both fall in the other's sex or gender orientation category. In contrast, for example, a bisexual woman and a homosexual man's orientations are incompatible, as the woman does not fall in the man's orientation category. Since bisexual men do fall in the orientation category of heterosexual or bisexual women, their exclusion thus involves something beyond basic orientation compatibility; an

additional sexual preference concerning sexual orientation. Similar to racial preferences, orientation preferences can come in excluding and including forms. Common examples are heterosexual and bisexual women excluding bisexual men, homosexual men excluding bisexual men, and heterosexual men fetishising bisexual or lesbian women. Notably, transsexual and transgender men and women are both often excluded as well as fetishised by cis-gender people (people with a gender identity matching their assigned-at-birth sex) of any sex and orientation. Moreover, these examples portray gender identity on a binary, prohibiting other forms, which may be the corner stone of these types of preferences.[68]

Furthermore, also with these types of preferences, further attitudes concerning sexual orientation and gender identity may be involved, such as the ideas and norms about masculinity and heterosexuality just mentioned. Again, on online gay dating platforms, such preferences are often explicitly communicated. For example, "no femmes" or "only straight acting" indicate exclusion of effeminate men who do not conform to heterosexual masculinity norms. As such, these types of preferences may discriminate based on gender identity. The preferences can also be thought of as forms of biphobia and homophobia, discriminating based on sexual orientation. Crucially, gender identity and sexual orientation seem to be tightly interwoven here, and thus best analysed together.

Concluding, there are endless more forms of sexual preferences that one can imagine, such as preferences relating to socioeconomic class, religion, physical or mental ability, body type, height, and general physical beauty, down to the colour of one's eyes, among many other traits. Some of these may be protected personal traits, while others may not. While I focus on preferences concerning race, sexual orientation, gender identity, and physique, I will address other preferences too. Moreover, beyond any one particular preference, I aim to provide the general conditions whereby any possible sexual preference can be morally evaluated. Lastly, as is already visible, sexual preferences operate in a complex manner, relating to various traits in different ways, and these relations in turn influence each other such that any insightful analysis of any sexual preference requires an intersectional approach that appreciates the various interactions. While I make many comparisons between how certain traits

[68] While I underwrite a wide spectrum of gender identities, I nevertheless mostly employ the binary in this essay, for the sake of clarity and length. As with all preference traits, a specific, elaborate discussion is required regarding sexual preferences concerning diverse gender identities.

function in sexual preferences, I do not assume any kind of 'analogy thesis', since, while being inseparable and having important similarities, the categories are not analogous, such that conceptual analysis of one cannot be straightforwardly applied to the other, as Cressida Heyes meticulously argues. (Heyes, 2006)

2: A classical liberal view of sexual preferences

The essence of what we can call the 'classical liberal view' of sexual preferences, which is based in a classical view of cognition and a liberal ideal of personal freedom. The classical liberal stance on controversial sexual preferences is most candidly captured in the often-heard folk-credo that "it is just a personal preference." I take this position to contain two distinct, albeit related, claims regarding (a) the 'personal'; the domain of sexual preferences, and (b) the 'just'; the agency of sexual preferences. With this, there are three arguments to defend controversial sexual preferences as not morally objectionable; sexual preferences are a non-moral matter (personal freedom), sexual preferences are morally justified (function), and/or sexual preferences meet an excuse condition concerning moral responsibility (agentive control). I now aim to spell-out these claims as fully and structured as possible in order to make the strongest case for this position.

2A: The domain of sexual preferences

The first part of sexual preference being 'just a *personal* preference' addresses the domain of operation. The domain argument holds that controversial sexual preferences do not cause any morally unjustified harm, because their operation is exclusively restricted to the private ('personal') sphere.

Public and private sphere: While sexual preferences concern human social affairs, and hence relate to other individuals, their operation is exclusively restricted to the domain that we can call the 'private sphere', and does not affect the 'public sphere'.[69]

The public sphere is the domain that every individual calls their home. Being home to all, everyone's interests are represented here, and hence behaviour can affect the good of others. This makes the public sphere an obvious domain for morality to function in, in order to protect the just and equal treatment of every individual and group concerning institutions such as businesses, housing, and schools. The private sphere, by contrast, is home only to a select group of people, who are members on a voluntary basis, and are so in virtue of personal relationships to an individual such as family, friends, and sexual partners.

Restricted operation: The influence of sexual preferences is exclusively restricted to sexual partner selection within the private sphere; they do not influence any other behaviour concerning people in the (morally evaluable) public sphere. This is one way of interpreting "what I do in the bedroom is my own business and does not concern anyone else."

A clear example of this widespread view of the domain restriction of sexual preferences is voiced by Jesse Matheson, who states that, when a person excludes, for example, 'Asians' as sexual partners, "they may not be racist to Asians in everyday life, in fact they may have loads of Asian friends. They just aren't sexually attracted to Asians and don't particularly wish to sleep with them." (Matheson, 2012) A more elaborate version of this view is expressed by Raja Halwani, who argues that controversial sexual preferences do not relate to any other, non-sexual attitudes, and, if an agent does not explicitly endorse the controversial attitude as general attitude, they do not reflect an agent's true values. Even if sexual preferences may contain racial stereotypes, "the stereotypes need not be operated outside the sexual context. When it comes to sexual desire, which is often intimately connected with fantasy, people are able to compartmentalize: they can have sexual desires containing weird or immoral beliefs, yet not have these beliefs across the board." (Halwani, 2017, p. 17) Thus, Halwani concludes, "we cannot simply infer from the fact that people with racial stereotypes as part of their sexual desires have these stereotypes across the

[69] An interesting discussion of racial inequality being treated differently in the public and private sphere can be found in (Thomas, 1999, pp. 191-192, 196-197).

board, let alone accept them." (Halwani, 2017, p. 18) Instead of inferring any relation between a person's sexual preferences and other attitudes, we should rely on a "person's higher-level attitudes towards his sexual desires (…) if he ensures that they do not pervade his belief system, he seems to be morally in the clear." (Halwani, 2017, p. 18)

A-moral private sphere: The exclusive nature of membership of the private sphere rules out the representation of the interests of everyone, and thus, on *a priori* grounds alone already, opposes moral notions like justice or equality such that sexual preferences cannot be discriminative. No one has a moral right to be anyone's friend or lover, and no one has a moral obligation to fancy or have sex with anyone. This is another way or reading, "what I do in the bedroom is my own business and does not concern anyone else." In the private sphere, people have a high degree of personal freedom. Similarly, in matters of sexuality, people have sexual freedom, the 'right to sexual choice', whereby it is entirely up to the free choice of those who voluntarily participate with one another. As such, who to sleep with and what to do together is a non-moral matter. (Callander et al., 2012; Matheson, 2012; Watts, 2012)

In summary, controversial sexual preferences are a non-moral matter, because (on a classical view of sexual cognition) they only influence behaviour in the private sphere, in which (on a liberal view of morality) people have personal and sexual freedom.

Sexuality goal-fulfilment: Similar to how selection of employees based on traits that are required to carry out a certain job is morally justified as non-discriminative in virtue of fulfilling the goals of the job, sexual preferences function in fulfilling sexuality goals. Sexual preferences are related to what someone finds sexually attractive. Attraction is typically a necessary feature for fulfilling the goal of sexual pleasure in sexual relations. As such, controversial sexual preferences do not arbitrarily discriminate any group of people, but are rather a morally justified selection in virtue of achieving the function of sexuality.

Halwani draws on this argument concerning racial preferences. A "sexual partner must usually have (or lack) a property related to their race or ethnicity for PRSDs (people with racial sexual desires) to find them attractive. Because racial and ethnic looks are normally part of physical looks, and because physical looks are

normally necessary for satisfying the goal of sexual pleasure or satisfaction, PRSDs choose their sexual partners (partly) on the basis of racial or ethnic looks." (Halwani, 2017, p. 6) Noticeable in this argument is the understanding of controversial sexual preferences as relating only to phenotypic, physical features, such as skin-colour or hair-type, and nothing beyond it (i.e. further beliefs or attitudes). A person with controversial sexual preferences "might have no bad beliefs, feelings, or values towards members of G." (Halwani, 2017, p. 12) Similarly, another philosopher, Charles Mills, also acknowledges the possibility of 'sexual exoticism' concerning race, culture, or class without being connected to any further attitudes, and argues that this is morally unobjectionable. (Mills, 1994, p. 145)

As such, controversial sexual preferences, even if they happen to concern, for example, racial properties, are simply a part of attraction, similar to, for example, fancying freckles; they are not arbitrary and stereotyping, and not part of any oppressive structures, but harmless components of achieving the goal of sexuality. An additional argument that can be added here, is that on the same grounds of attraction and fulfilment, sexual orientations are typically not thought of as being morally wrong and discriminative. While most orientations do exclude based on sex/gender (e.g. homosexual men exclude women), this does not make one sexist. Rather, excluding some sex is justified because it is a necessary requirement for sexual fulfilment. If sexual fulfilment justifies sex differentiation by sexual orientations, then differentiations by sexual preferences are similarly justified. (Matheson, 2012)

In summary, even if the private sphere were to admit to moral evaluation, and controversial sexual preferences were to cause certain moral wrongs, this is not a case of arbitrary and unjustified discrimination, but rather it is morally justified in virtue of the function of sexuality.

2B: The agency of sexual preferences

The second part within the view that sexual preference is '*just* a personal preference' addresses the agency of sexual preferences. The agency argument holds that, since sexual preferences are predetermined and unchangeable, people lack control over

them, and therefore they cannot be held morally responsible for their controversial sexual preferences.

Predetermined preferences: Sexual preferences are predetermined. On a strict version of predetermination, sexual preferences are, like sexual orientation, largely caused by biological factors. While no single cause has been identified as crucial determinant, and the exact interplay of influences is still unknown, there is much scientific consensus that genetic factors and prenatal hormonal exposure in early uterine environment are significantly involved, and biological models of sexuality are generally favoured.[70] (Frankowski & American Academy of Pediatrics Committee on Adolescence, 2004; Lamanna et al., 2014; Långström et al., 2010; Rahman & Wilson, 2003; Vare & Norton, 1998) On this view "there are true and definite forms of sexuality that remain constant (and…) explain sexuality in terms of innate motivational patterns that have evolved." (Baumeister, 2000, p. 347)

A somewhat less strict version of predetermination could allow for sociocultural influences as a determinant of the configuration of people's sexual preferences. Hence, it could be that a person develops, for example, certain racial preferences, due to growing up in a racist society. However, this causal history only evidences that the society is morally pernicious, not the preferences themselves or the person who harbours them. Moreover, this developmental process is significantly arbitrary such that the development of sexual preferences is not necessarily linked to environmental conditions in an interesting way. "The formation of sexual preferences can go in all directions, such that X might find X's-self attracted to members of G, while Y, growing up in the same society as X and belonging to the same race as X, might not." (Halwani, 2017, p. 12)

Unchangeable preferences: Once formed, sexual preferences are unchangeable. An individual cannot exhibit any significant sense of voluntary choice over one's sexual preferences. (Frankowski & American Academy of Pediatrics Committee on Adolescence, 2004; Kersey-Matusiak, 2012; Lamanna et al., 2014) An extensive review of research on 'sexual-orientation change efforts', treatments

[70] Biological explanations have played an important role in opposing and debunking previously popular psychological models, which, for example, understood homosexuality as resulting from distorted development through childhood experiences, like abuse or defective relationships. Such psychological models formed a basis for judging homosexuality to be a mental disorder. The supporting research, however, has been found to be severely flawed, which has played an important role, together with the biological models, in addition to moral arguments, to depathologise and protect sexual orientation. (APA, 2008)

typically aimed at changing same-sex orientations to heterosexuality, through, for example, behavioural, psychoanalytic, medical, and religious 'reparative', 'reorientation', or 'conversion' therapy, typically lack methodological rigour and hence fail to evidence any efficacy. Rather, the studies indicate that there is a 'core' sexuality, that enduring change is uncommon, and that treatments can even be harmful. (APA, 2006; 2009, pp. 42, 82; Beckstead, 2003; Drescher, 2002; Royal College of Psychiatrists (RCoP); Rust, 2006)

Responsibility and control: The fact that sexual preferences are predetermined and unchangeable is crucial when it comes to evaluating moral responsibility. Regardless of whether the effects of controversial sexual preferences are harmful or not, agents cannot be held responsible for their sexual preferences (and their effects), since they lack control through voluntary choice.

This view can be mapped onto (arguably the classical and still most prominent) theories of the conditions for moral responsibility called 'volitionism'. According to volitionism, an agent needs to have a meaningful level of control over one's actions or mental states in order to be responsible for them. Such control involves as a necessary condition the agent exercising conscious, voluntary choice or reflective endorsement, such that the action or state relates to an agent's explicitly considered and endorsed beliefs and evaluative commitments.[71] As one of the foremost proponents of volitionism, Neil Levy, writes, "an agent is responsible for something (an act, omission, attitude, and so on) just in case that agent has – directly or indirectly – chosen that thing." (Levy, 2005, p. 2)

In the case of sexual preferences, there is no meaningful control through voluntary choice, since they are predetermined and unchangeable. Levy's writing on implicit attitudes can be extended to controversial sexual preferences, as in both cases the acquisition is uncontrolled, and subsequently, once possessed, just as sexual preferences are unchangeable, "implicit attitudes will display little reasons-responsiveness. They are sensitive to cues with which they have been associated in the agent's learning history, not to (justificatory) reasons." (Levy, 2016, p. 11)

As additional argument, a similarity to sexual orientations can, again, be made here: Just as sex-exclusion by sexual orientation is not blameworthy sexism, because

[71] While volitionist theories come in various forms, and has many important here undiscussed nuances, the general principles drawn here are fairly universal to most volitionists. For important accounts of volitionism, see (Fischer & Ravizza, 1998; Levy, 2005, 2016; Mele, 2006; Rosen, 2003, 2004; Vargas, 2013).

people lack control over their sexual orientation, so is any exclusion by sexual preferences not part of people's moral responsibility. (Matheson, 2012)

In summary, (on a classical view of sexual cognition) sexual preferences are not a matter of choice, but "*just* the way I am." (Halwani, 2017, pp. 8, my italics) Hence, lacking conscious control, (on a volitionist view of moral responsibility) people meet an excuse condition and are therefore not morally responsible for their controversial sexual preferences.

3: A dynamic developmental view of sexual preferences

Now, I turn to setting out a 'dynamic developmental view' of sexual preferences, aiming to refute the classical liberal view. I argue that some controversial sexual preferences directly and indirectly cause harm, which is discriminative, and for which agents can be held morally responsible. In accord with the previous section, I first discuss (b) the domain, followed by (c) the agency of sexual preferences. Before doing so, however, I take a step back to (a) lay bare the models of human cognition that, albeit implicitly so, form the foundation of any moral theorising about sexuality.

3A: Sexual cognition

As we saw above, there are two crucial properties of the classical liberal view's understanding of sexual preferences; (i) The configuration of sexual preferences is predetermined and unchangeable, developed through biological (and possibly early-life sociocultural) factors, their form is hardwired, permanent, carved in stone. (ii) The operation of sexual preferences is exclusively restricted to sexuality, sexual preferences are 'compartmentalised', functioning within a distinct mental faculty designated to sexuality, isolated from the rest of a person's cognition. Let us call these two properties, respectively, the 'fixed configuration' and 'modular operation' properties of sexual cognition.

There are two theoretical frameworks that can help us to understand what these properties entail, and, subsequently, position us to evaluate their validity. Note that, no classical liberal proponent (to my knowledge) claims any affinity with the following models, but neither does any with any other model. Rather, the model of cognition is only present implicitly, in the claims discussed above, and it is from this that I draw connections to the following models, since I take them to share their essential qualities.

The modular mind: Firstly, both the fixed configuration and modular operation properties can be found in 'strict modularity' model of cognition, with roots in the work of Jerry Fodor. (Fodor, 1983) Fodor conceptualises certain mental systems as 'modular', meaning that, among other features, they are innately specified and hardwired, operate automatically and autonomously, have a specialised and unique role, are domain and content-specific, do not use other aspects of cognition, but are, most crucially, informationally encapsulated from other systems. Nevertheless, although strict modularity may map well onto the classical liberal model of sexual cognition and provide some foundation, in particular in focussing on informational encapsulation, this seems like an unlucky marriage. While a proper discussion of modularity goes beyond the space of this paper, the model seems to be an unfit candidate. Firstly, strict modularity is all but undisputed.[72] (Churchland, 1988; McCauley & Henrich, 2006; Prinz, 2006b) Instead, less stringent, non-informationally encapsulated models have been proposed. (Carruthers, 2006; Jackendoff & Pinker, 2005; Sperber, 2002) The absence of informational encapsulation allows for 'cognitive penetrability', the causal influence on a system's operations by information from other systems, for example, beliefs, desires, and intentions. (Pylyshyn, 1984, 1999; Stokes, 2012, 2013) However, the classical liberal view of sexuality employs a model of strict modularity that *is* informationally encapsulated and which is, thus, heavily contested. Secondly, the modularity thesis principally aims at capturing linguistic and visual cognition, and explaining sexual cognition may well be an entirely different matter, with the weight of evidence in favour of any such applicability being on the proponents of the classical liberal view.

Essentialism: Considering models that are specifically applied to sexuality, we find both fixedness and modularity in so-called 'biological essentialist' theories of

[72] For example visual information may affect language processing. (Carston, 1996; McGurk & MacDonald, 1976) And linguistic cognition may also function in non-linguistic processes. (Heiser et al., 2003; Saygın et al., 2003)

sexuality. (Baumeister, 2000, p. 347) A notable example of essentialism is 'Sexual Strategy Theory' (SST) by Buss and Schmitt. (Buss, 1998; Buss & Schmitt, 1993) SST proposes the existence of countless evolutionarily adaptive psychological mechanisms that are distinct to sexuality, with sexual desire at its core, which can explain sexual preference differences between genders. However, while the theory's success herein is already challenged, it runs into further trouble explaining gender similarities, and especially explaining individual differences. (Allgeier & Wiederman, 1994) Concerning the latter issue, Buss proposed four possible models, but the most plausible ones of these resort to social and experiential learning, which would challenge fixedness and modularity. (Allgeier & Wiederman, 1994; Buss, 1991) Overall, the theory is contested, and the essentialist approach to sexuality seems quite bankrupt. (Bancroft, 2009, pp. 8-10)

What this brief discussion reveals is classical liberalism lacks a compelling theoretical and empirical basis, since the most familiar models, which fit with the classical liberal fundamental claims about sexual cognition, are very frail. On such a shaky foundation, the further claims and judgements that the classical liberal view proposes are in themselves already very shaky. As such, the validity of the classical liberal view is challenged already merely by scrutinising its own theoretical and empirical support. Moreover, as I will now go on to argue, an alternative view seems to have a much more compelling foundation, both enjoying stronger empirical support, and offering more theoretical illumination of the issue's intricacies, in addition to refuting the classical liberal claims.

The dynamic mind: Instead of a fixed and modular view of sexual cognitive functionality and development, I propose to draw on dynamic models, which seem more empirically plausible and have ample potential applicability to sexual preferences. Based in this approach, sexual preferences can be seen as (i) 'pervasive', connected to a person's other attitudes in a mutual causally influential relation, and (ii) 'fluid', having a significant degree of plasticity or malleability, thus being able to develop over time.

There is a range of approaches that capture this view, and as a family of theories I refer to them as 'dynamic' models. Especially 'Connectionism' and 'Dynamic Systems Theory' seem well suited for the current purpose of understanding

sexual preferences. The two are increasingly employed as complementing one another, and interesting efforts are made for an integrated, unified model. (Kloos & Orden van, 2009; Mareschal et al., 2009; McMurray et al., 2009; Smith & Samuelson, 2003; Spencer et al., 2009; Thelen & Bates, 2003; Thomas et al., 2009) Together, they offer a model of interactive functioning and development, with increasing empirical support, and direct applicability to sexuality.[73]

Connectionism understands cognition as functioning through non-modular systems that are connected to one another in a network, sharing information, strengthening associations, and activating other systems; thus, instead of informational encapsulation, systems are taken to operate through interactive processes with other cognitive faculties and transfer information across content domains. (Flusberg & McClelland, 2014; Heberle, 2003; McLaughlin, 1990; Sokolik, 1990) Secondly, such a functional mechanism allows for learning through a complex process of acquiring knowledge through experience. (Schumann, 1990) "For the connectionists, learning takes place through the strengthening and weakening of interconnections in response to examples encountered in the input", which happens across the entire network. (McLaughlin, 1990, p. 624) This approach has been interestingly applied to, for example, the development of attitudes. (Eiser, 1994)

Dynamic Systems Theory (DST) is a model of the functioning and development of complex systems, originating from mathematics and physics and introduced in developmental psychology by Esther Thelen to explain how psychological traits of human beings are shaped through similar mechanisms, as a continuous, mutual-influential interaction with multiple factors, including both internal factors, such as genes, hormones, beliefs, and emotions, and external factors such as social norms, relationships, experiences, and environment. (Lewis, 2000; Lickliter, 2008; Miller, 2002; Spencer et al., 2006; Thelen, 2005; Thelen & Bates, 2003) DST is now being employed to explain a wide range of complex psychological and social phenomena ranging from the development of motor skills (Thelen et al., 1987), language (Christman, 2002; Wolf, 2013), emotion (Izard et al., 2000; Lewis, 1995; Magai & McFadden, 1995), cognition (Magai & McFadden, 1995; Thelen & Smith, 1996), and attachment (Coleman & Watson, 2000), to phenomena more closely related to the current topic such as personality (Read & Miller, 2002), gender

[73] Additionally, one can look at 'Interactionist', or 'Conjoint' theories. (Bancroft, 2002; DeCecco & Elia, 1993; DeLamater & Hyde, 1998; Tolman & Diamond, 2001)

(Fausto-Sterling, 2000), and female same-sex fluidity (Diamond, 2008; Partridge, 2005). Especially within developmental psychology, DST is becoming increasingly dominant. (Granic, 2005; Partridge, 2005) With respect to sexual preferences, an essential aspect of this approach is that it can appreciate and account for 'erotic plasticity', the possibility of people exhibiting substantial intra-individual variability in their sexuality over time.

In line with DST, I do not aspire to conceptualise the full range of determinants or any definite outcomes. According to DST, developmental processes are marked by 'equifinality' and 'multifinality', the possibility of two people from different starting points reaching the same outcome and, vice versa, the possibility of two people reaching different outcomes from the same starting point. Hence, definite prediction of an endpoint or determining a single cause is impossible. Rather, the principal focus of DST is the multi-determined process of continuous change over time, consisting in emergence, stabilisation, change, re-stabilisation, etc., as the essential function of a system, and studying the factors that make up this process can inform us about what makes certain development more *likely*. In this mind, the crucial question for us becomes; What factors influence the change and stabilisation of people's sexual preference, and hence make certain preferences more likely?

Acknowledging the complex of determinants, the purpose of the ensuing emphasis on cultural and personal impact on erotic plasticity should not be taken to suggest that these factors are the *most significant* ones, but rather that the establishment of their *substantially significant* influence enables a meaningful way of approaching the evaluation of moral responsibility in this matter. Hence, the current view expressed here adheres to DST in appreciating a non-hierarchical plurality of influencers while prioritising attention to the developmental potential of erotic plasticity.[74]

[74] Let it be clear that I do not here wish to take position at the other side of the spectrum by endorsing 'social constructivist', 'social learning', or similar theories that depict individuals as 'empty organisms filled and shaped by culture and society, devoid of consciousness and intention'. (DeCecco & Elia, 1993) Instead, I follow Baumeister in denouncing the bifurcated essentialist-constructivist approach and his well-expressed view that human sexuality is "a rich, confusing tangle, in which biological drives, sociocultural meanings, formative individual experiences, and additional unknown factors play powerful roles." (Baumeister, 2000, p. 347)
For example, Franscesca Minerva argues we should "distinguish between characteristics we are hard-wired or socially influenced to consider attractive." (Minerva, 2017, p. 186) Creating such an artificial dichotomy not only incorrectly construes the complex origins of preferences; moreover it unnecessarily restricts possible developmental strategies.

Now, let us return to discussing the evaluation of controversial sexual preferences by advancing the 'dynamic developmental view' of sexual preferences, arguing that sexual preferences are dynamic rather than modular, and plastic rather than fixed.

3B: The domain of sexual preferences

On a 'dynamic developmental view' of sexual preferences, rather than being 'just a *personal* preference', the functionality of sexual preferences is understood not as modular, but dynamic, and hence it can be appreciated that controversial sexual preferences are causally influential both in indirect harms in non-sexual affairs, as in direct harms in sexual affairs, which constitute moral wrongs and discrimination.

Attitude connectedness: Sexual preferences are a type of 'attitudes'; mental states involving beliefs, feelings, values, and motivations in relation to sexual partners (and acts). Understanding sexual preferences as types of attitudes is no contentious claim, as we saw that many others like Halwani and Mills talk about it in a similar manner, albeit without specifying its meaning. Rather, the extra step I propose we take when talking about sexual attitudes, is that we see these as part of a person's entire character, together with one's other attitudes, and that, while they need not all be perfectly in accord with one another, they are tightly intertwined, sharing information and mutually influence each other, thus typically having content-driven relations – as dynamic models conceive. As such, controversial sexual preferences are typically connected to other attitudes, such as implicit (or explicit) biases, unconscious (or conscious) stereotypes, and false beliefs.

To empirically substantiate this view, we have to look at several types of research. Firstly, there are studies directly on the relationship of sexual preferences to other attitudes. These studies generally evidence significant relatedness. However, there they are only scarce in number. For example, a series of studies on racial sexual preferences among homosexual and bisexual men found that practising sexual racism ($R = -0.08$ ($p<0.001$); Callander, personal correspondence) and acceptance of sexual racism strongly associates, more than any other factor such as level of education or place of residence, with intolerant views towards multiculturalism in general, from

which the authors conclude that, rather than a matter of personal preference only related to sexual partners, sexual racism is an expression of racism. (Callander, 2013; Callander et al., 2015) A study on sexual preferences concerning partner dominance found that stronger preference for male dominance was related to sexist attitudes in women (attitudes that undermine women's agency) and increased rape myth acceptance attitudes (beliefs that assume victim consent or responsibility) in men. (Harris et al., 2017) And studies on sexual preferences concerning Asian women, found that stereotypes of Asian women as submissiveness and docility makes them more vulnerable to sexual harassment and sexual violence (besides putting increased pressure on them to fulfil or resist stereotypical expectations. (Cho, 1997; Chou, 2012; Chou et al., 2012, 2015; Nemoto, 2009; Park, 2012; Patel, 2009; Sue et al., 2009)

Such studies can be taken as a first indication that sexual preferences are related to other attitudes. Further support for this can be found in the larger body of research on physical attractiveness and beauty, which are arguably central components of sexual preference. Already in the 1970s the significance of 'halo effects' of attractiveness have been systematically studied as causing all sorts of positive character attributions. (Dion et al., 1972) Attractiveness functions in many other, non-sexual judgements and decisions, causing strong favouritism towards people who meet beauty norms. To mention a few notable examples, on the labour marking, hiring, career advancement, and salary are mediated by attractiveness norms. (Dang, 2017; Little, 2017; Maestripieri et al., 2017; Minerva, 2017) This effect is all but an insignificant side effect, but economically comparable to race and gender gaps. (Hamermesh, 2011) In education, attractive students are evaluated by teachers as more intelligent and having greater academic potential. (Westfall et al., 2016) And, vice versa, attractive professors are evaluated by students as better teachers. (Riniolo et al., 2006) Even in the court of law, sentences are mediated by attractiveness of victims and of the accused. (Mazzella & Feingold, 1994; Wuensch et al., 1993)

Such data supports the claim that typically sexual preferences are causally influential in other, non-sexual judgements and decisions. Hence, we can reject the classical liberal claim that the operation of sexual preferences is typically exclusively restricted to sexual partner selection. As such, there is strong reason to be suspicious

of controversial sexual preferences, and to infer other attitudes from them, since they are typically intertwined with other attitudes.

Indirect harms: From this, we can identify a first type of morally wrong 'indirect harms' that controversial sexual preferences cause: professional, educational, and legal harms. In situations where differential treatment is influenced by irrelevant sexual preferences (for example the selection of a job candidate based on their attractiveness, where attractiveness is an irrelevant trait for job-performance), this is arbitrary and harmful by dismissing relevant qualities and capacities, which is unjust, and thus morally wrong. This point has been argued by various others under the notion 'beautyism' or 'lookism'. (Bartky, 1990; Mahajan, 2007; Rhode, 2010; Soble, 1982; Willard, 1977; Wolf, 2013) Furthermore, in cases of controversial sexual preferences, the indirect harm concerns selection based on traits such as race, sexual orientation, gender identity, and possibly body type, which is, beyond being morally wrong, systemically discriminative.

Proponents of the classical liberal view may reply that it need not *always* be the case that sexual preferences influence other judgements. While this is true, it is empirically plausible that it *usually* will be the case, and in those cases, controversial sexual preferences indirectly cause discrimination. And as it will usually be the case, there is good reason to be suspicious of controversial sexual preferences.

Direct harms: It is a different question if controversial sexual preferences also cause morally wrong 'direct harms' in the private sphere.

Private sphere morality: To start, the supposed a-moral status of the private sphere in virtue of its exclusive nature, which rules out moral notions like equality and justice, can be challenged by having a closer look at other members of the private sphere, besides sexual partners; friends. While I agree that there are no moral obligations to befriend, nor moral rights to a particular person's friendship, but that friendship is governed by mutual voluntary choice, this does not eliminate the possibility of discrimination. To illustrate, take the following scenario. *Susan is a White woman of Western European heritage, who grows up with Abdel in the same neighbourhood. In spite of their mutual appreciation for the other's character, their shared values, many common interests, and typically enjoyable experience of each other's company, Susan does not want to be friends with Abdel, because of his Middle*

Eastern descent, and her preference for only White friends. (Race, in this example, may be replaced by other traits, such as sexual orientation, gender identity, or body type). I take it that there are few people who would argue that, because friendship is a free, voluntary matter, the choice to exclude people based on their race does not have any moral substance. Rather, since such preferences can cause harm, they have to be justified, or else they are arbitrary, morally wrong, and even discriminative.[75] This intuition about friendship shows that the exclusivity of the private sphere does not *a priori* eradicate morality. As such, preferences concerning sexual partners are not a-moral, but, when causing harm, in need of moral justification.

Now, let us turn to assessing direct harms what controversial sexual preferences cause in the private sphere, and whether the arguments presented in section 2 justify this. Note that, drawing on the connectedness of attitudes it is plausible that controversial sexual preferences are usually related to other attitudes, I initially focus on the majority of cases of controversial sexual preferences in which other attitudes *are* involved.

Psychological harms: One type of harms that controversial sexual preferences cause is the 'depersonalisation', 'objectification' and 'otherising' experience of the target. As the anecdotes at the beginning of the essay show, being targeted (both included or excluded) based on stereotypical traits dismisses a person 'as individual', and being perceived and responded to as such. Instead, one is merely treated as instantiation of some group in virtue of several traits (that may not even be present) that one's individuality is reduced to, disregarding other traits as insignificant. Such experiences are very hurtful. As Robin Zheng excellently argues on the shoulders of Martha Nussbaum and Harry Frankfurt, especially in the context of love relationships, it is crucial to be appreciated 'for who you are', and the involvement of controversial sexual preferences ('yellow fever', in Zheng's article) can deteriorate that, making targets feel replaceable, doubtful of their partner's appreciation, insecure about their own worth. (Zheng, 2016) As such, controversial sexual preferences can cause significant psychological harm to targets and sabotage relationships.

[75] There may be justifying reasons for certain groups of people to have friendship preferences concerning protected traits. For example, as Thomas mentions, some argue that for Black college students it is morally permissible, or even obligatory, to befriend other Black college students. (Thomas, 1999, p. 194) I think that in certain cases there are strong arguments for doing so, concerning the hostility of university environments to people of colour and the need to support one another, which at least partially justifies such exceptional positions.

Social harms: Another type of harms that controversial sexual preferences cause is the limitation of targets' opportunities to find sexual partners or other intimate relationships. This harm is highly significant, because sexuality, love, and marriage are important factors for a flourishing human life, both in terms of physical (lifespan and overall health: (Burman & Margolin, 1992; Seldin et al., 2002)) and mental health (depression and life satisfaction: (Diener et al., 2000)), but also socially (the status of marriage), economically (marriage tax benefits), and even for finding meaning in life (happiness: (Blanchflower & Oswald, 2004)). Elizabeth Emens provides a more detailed discussion of this, promptly characterising the matter of the goods of sexuality as a 'market': "Social stigma and structural constraints exclude some people from meaningful participation in the dating, sex, and marriage markets." (Emens, 2009, p. 1374). In the same mind, William Elder et al. conclude on the basis of their various studies on sexual self-understanding, that "being attractive was 'a form of currency' that could be traded not only for the attention of other attractive men for sex or relationships, but for a wide range of opportunities, including occupational advantages (and) social power." (Elder et al., 2015, p. 952) We can thus extend Charles Mills' conclusion about the burden that White beauty norms place on people of colour, writing that "people are socially disadvantaged through not meeting intra-racial standards of attractiveness." (Mills, 1994, p. 146) Not conforming to sexual attraction norms (whether racial, orientation, identity, etc.) diminishes a vast array of people's opportunities, which is a significant harm. Additionally, this can even cause further harms, as people are placed in weak negotiation positions, which pressures them to risky behaviour. For example, as Nathaniel Coleman argues, lessened bartering power due to racial preferences may be a cause of increased HIV-risk of Black homosexuals. (Coleman, 2011)

Discrimination: Ultimately, controversial sexual preferences can cause a further harm; the perpetuation of systemic discrimination. Laurence Thomas argues that, since the private sphere is most important for a 'complete life' (i.e. Aristotelian 'flourishing'), discrimination there outweighs equality measures in the public sphere, and thus undermines any purported egalitarianism. (Thomas, 1999) While I am sympathetic to Thomas' view, I rather not draw too much on a metaphysical hierarchy of life-satisfaction factors. I do, however, endorse the relatedness of the two spheres in fostering discrimination as a crucial insight that can be capitalised on further. While some of the abovementioned harms may be relatively small, others are very

serious. Moreover, all of these harms together, more and less considerable, and private and public, jointly function in continuing and reinforcing structural patterns of discrimination of the targeted groups. Marginalisation and oppression cannot be restricted as supposedly only being relevant in one (public) sphere, but rather happens in all spheres of life. Allowing discrimination in the private sphere essentially undercuts any efforts towards an ideal of egalitarianism in the public sphere. Thomas powerfully points out how this makes our commitments to equality in general disingenuous, rhetorically asking: "How seriously can we be about equality in the public sphere if we believe that it is morally permissible to privilege our own ethnicity as a matter of principle in the private sphere (…)?" (Thomas, 1999, p. 195)

Seeing how equality cannot be restricted to one sphere, we can see how the private sphere harms of controversial sexual preferences are not merely an *expression* of patterns of discrimination elsewhere, but an *inseparable* and *constitutive* part of exactly what systemic discrimination is. And thus, through their systemic effects, some controversial sexual preferences are discriminative preferences, and are forms of sexual discrimination.

In summary, an analysis of the connectedness of attitudes shows how controversial sexual preferences function in harms in the public sphere, which are morally wrong. The classical liberal could acknowledge this much, but still hold that the harms caused in the private sphere, which is the main topic under discussion, are not morally wrong. However, a further analysis of how discrimination operates throughout life's spheres challenges the liberal conception of a private sphere. We cannot so easily distinguish a part of life as non-moral. Rather, if we are to appropriately appreciate the aims of moral values like equality in a serious manner, we have to scrutinise the private sphere with certain moral considerations too, since discrimination is exactly constituted by harms in any sphere. Thus, personal and sexual freedom cannot be wholly unrestricted, or else it undermines the moral project in its entirety.

Justified discrimination: Now that we have a clear view of the harms that controversial sexual preferences cause, we can also make some initial comments on the second classical liberal argument, that this is part of satisfying the goal of sexuality, and thus not arbitrary, but justified. Firstly, the harm effects, in their

totality, encompassing professional, educational, legal, psychological, and social harms, are agonisingly severe, and perpetuate systemic discrimination. Secondly, the harm effects are not side-effects that can be assessed separately, but an integral part of the structure of controversial sexual preferences. As such, certain preferences are by their very nature morally wrong and discriminative. The fact that these preferences fulfil some goal does not erase their wrongness. The classical liberal view, by letting goal-satisfaction so readily trump all the harms, underappreciates the severity of the harms. Instead, I argue that in the case of preferences that are significantly harmful, in a systemic, discriminative manner, this is a harm that does not weigh up to just any individual pleasure satisfaction. As such, pleasure fulfilment does not justify the harms, but, when the harms are severe and systemic, controversial sexual preferences are discriminative preferences, and constitute a form of sexual discrimination.

This may well be insufficient to convince some classical liberalists, however, who might simply beg to differ and insist that individual pleasure does outweigh systemic oppression, especially since how one can satisfy one's sexual needs is not a matter of choice, but fixed, which would thus make pleasure-fulfilment impossible for some people. To address this, there is another feature of sexual preferences; their fixedness, or, as I argue, malleability. As controversial sexual preferences constitute such grave harms, simply accepting them as a given, while there is potential for change, makes that they are not necessary conditions for the satisfaction of sexuality, but that people have avenues available to them to develop other, non-harmful ways sexual fulfilment, which renders controversial sexual preferences to be unnecessary and arbitrary; unjustified discrimination. This will be addressed in 3c.

Minimal preferences: Another argument that classical liberals could make here is that maybe in the cases where sexual preferences are connected to other attitudes, they cause the harms above and are thus morally wrong harms, but, as discussed in 2a, not *all* preferences *need* to be like that; some preferences may only concern a certain phenotypic feature, without any further beliefs, feelings, biases, and so forth. We can call such preferences 'minimal preferences', drawing on the concept of 'minimal race', which refers only to phenotypic traits, such as skin colour, and geographical ancestry, without including stereotypes about other traits, for example psychological or moral characteristics. (Anderson, 2010; Taylor, 2013) Similar to how one can have a minimal preference for milky-white or porcelain skin, one could have the minimal

preference for women with slender bodies, the minimal preference for tall, muscular, and strong-jawed men with particular grooming styles, or a minimal preference for 'straight acting' behaviour without showing emotional vulnerability.

Acknowledging that there is a logical space for minimal preferences, these are very rare, and still constitute harms. The first reply is that in many cases, even minimal preferences can cause severe harms and perpetuate discrimination. For example, while it might be true for some individual to solely prefer pale skin, without having further racial attitudes, this still feeds in to the privilege of White people compared to people of colour, given that culture drives the development of particular preferences more than others. And even 'positive' preferences, like favouring brown skin or high-BMI body types objectify and depersonalise the target. Secondly, as will be discussed in 3c, malleability of preferences implies that they are arbitrary choices, which are discriminative when accepted. Finally, as third reply we can note how the given examples of minimal preference sound increasingly unlikely to be, in fact, minimal. While it may be the case that certain preferences are minimal, such as fancying certain types of eyes, other preferences are inseparable from further attitudes. For example, women having minimal preferences excluding bisexual men, or homosexual men having minimal preferences excluding effeminate men, is impossible to conceptualise without reference to attitudes concerning sexual orientation and gender identity. Here it is important to see the many stigmas around orientation and identity, and their close cultural association, whereby men's deviance from heteronormative masculinity is constrained, and any gender non-conformity and diminishing of masculinity of men is entailed by social pressure to examine their sexual orientation, displaying social homophobia and biphobia. (Hennen, 2008; Steinman, 2011) Men who express same-sex interest are seen as homosexual rather than bisexual. (Flanders & Hatfield, 2014) Even 'one drop of homosexuality' is taken to make a man homosexual, rather than bisexual, and this challenges his masculinity. (Blumstein & Schwartz, 1976; Carrier, 1985) In turn, gay men are seen as non-masculine, and unsuited for masculine careers such as a politician, doctor, or judge. (Taywaditep, 2002) And within the gay community, 'straight acting' is a widely expressed preference, which has been studied as relating to masculinity images, acting in accordance with gender norms. (Eguchi, 2009)

Having clarity of the various harms (educational, professional, legal, psychological, social, and systemic), we can analyse the various sorts of preferences. A full, detailed, and sufficiently nuanced catalogue of all controversial sexual preferences goes beyond the scope of this paper, but an initial exploration should provide a decent ground for some tentative conclusions, pending further research and therewith more elaborate analyses.

Race: Sexual preferences concerning race might be the clearest example of all the harms discussed here. Studies show very evident racial sexual preferences. For example, in dating behaviour White men and women are typically most favoured, while Black women are favoured least. (King, 2013; OkCupid, 2014) Even within racial groups, for example in the African American community, lighter skin-tone is related to greater likelihood of getting married, and doing so to someone with higher income. (Edwards et al., 2004) As such, racial sexual preferences likely cause the psychological and social harms discussed above. In addition, it is a well-established fact that people of colour experience professional, education, and legal harms, and given the existence of racial sexual preferences, and attitude interconnectedness, these public sphere effects are plausibly influenced by sexual preferences to a significant degree. Moreover, all these harms feed into a long, and on-going history of systemic oppression of people of colour.

Sexual orientation and gender identity: Sexual preferences concerning sexual orientation are another clear example of severe and systemic harms, since there is a long, on-going history of discrimination against bisexual and homosexual people. As such, sexual preferences against bisexual men constitute sexual discrimination. Sexual preferences concerning gender identity I take to be a third type of clear severe harm and systemic discrimination. Appreciating how patriarchal social structures and norms construe a particular idea of masculinity, and favour masculinity compared to an idea of femininity, the exclusion homosexual and bisexual who are deemed less masculine feeds into discrimination of anyone who transgresses gender constructions. Similarly, women who transgress femininity norms may be discriminated, but in other ways, since in particular male affiliation with femininity in sexually penalised. (Bailey et al., 1997) As a clear example of harms, portrayal of muscular body ideals, which is tightly connected to heteronormative masculinity norms, causes conformity pressure, body dissatisfaction, lower self-esteem, psychological disorders such as depression, and unhealthy behavioural incentives such as excessive exercising,

whereby men are found to be almost as often unsatisfied with their bodies as women, and adolescent boys are even found to self-criticise and feel distress more than their female peers. (Barlett et al., 2008; Frederick et al., 2016; Mitchison et al., 2017). Finally, also transgender and transsexual people are often faced with much adversity, both in the private sphere, such as being excluded as potential sexual partner, and the public sphere, such as meeting discrimination at work, all of which causes a variety of harms, from psychological to career hardship. (Gerhardstein, 2010; Gerhardstein & Anderson, 2010; Kraemer et al., 2008)

Physique and physical beauty: Sexual preferences can relate to physique in numerous ways. One common preference is thinness, or say, low body mass index (BMI). Sondra Solovay argues that body-size biases have important similarities with other biases, such as sex and racial ones. (Solovay, 2000, p. 237) As discussed earlier, there is significant discrimination on the basis of higher BMI, possibly comparable to racial or sex discrimination, especially so for women. One factor that is different here, however, is the absence of a long history of this discrimination, as will be discussed in 3c. While this may feature into the evaluation somewhat, I do not take this to render BMI preferences non-discriminative. Traits such as height and hair have been found to significantly influence sexual and non-sexual judgements, especially towards men. For example, bald men are judged less attractive and less intelligent, and less likely to be invited for a job interview. (Henss, 2001) And taller men are judged more attractive, and influences earnings as much as weight does for women. (Mautz et al., 2013; Tyrrell et al., 2016) An explicit expression of this can be seen on online dating platforms such as Tinder, where many female profile descriptions contain no more than a height indication such as “1.70”, typically meaning that that men under said height will not be considered as potential partner. Speaking of systemic discrimination of bald or short people, however, sounds odd. Possibly this is merely due to lack of such a tradition of discourse. Alternatively, it may be a too distinct, singular factor to constitute discrimination. In that case, however, the same might be said for BMI. One way of dealing with this is by taking physical attractiveness discrimination as being constituted by not one, but a group of traits together, or not by simply any physical preference, but by overly strict norms that prescribe a very narrow range of acceptable trait-qualifications, and are uphold very sternly, overriding all of someone’s other traits. With this line of thought, I aim to balance appreciating the grave and systematic harms that beautyism can constitute, with the

acknowledgement of the fact that human beings are embodied creatures, whose individuality and relationships cannot be reduced to their minds alone. I thus argue that we should not preclude that people are to have physical preferences, but rather impose some moderation on how strict and how important body norms can be. As such, certain relaxed physical preferences can be justified, while more rigid forms of physical preferences can be judged discriminative.

Class: Painfully absent in this entire debate has been the discussion of sexual preferences concerning socioeconomic status (SES). One of the reasons for this, may be that classism currently gets little attention within a discourse dominated by identity politics, whereby class issues often pass by unnoticed or undiscussed. Additionally, class has become a progressively difficult to address matter, since social status, wealth, income, and social power allow for ever more complex relationships. Nevertheless, it is not difficult at all to imagine examples of people being targeted based on class preferences. One need but browse through almost any 19th century English romantic novel and will find a multitude of examples such as the disqualification of a prospective partner due to lack of social status and/or wealth. While such stories may seem out-dated nowadays, the actual practice is all but dead, albeit not pronounced in the same terms as partners being 'of standing', or often remaining unpronounced entirely, merely operating under the surface. Explicit instances can be found most frankly in hip hop and R&B music. A striking example is Chris Brown's song *Loyal*, elegantly narrating the dynamics of how the affluent are highly sought-after, while both penniless men and women are ruthlessly excluded as potential lovers, which causes the singers to be tormented by distrust of the sincerity of suitors who show them affection that might feasibly aim at their fortune rather than their individuality, thus heedlessly depersonalising them. (Chris Brown et al., 2014) Beyond such psychological harms, other rappers make social harms explicit, portraying how even friendship is reserved for those with deep pockets (for example, see Philphy Rich (2012) *Light It Up*; Thugga Massina (2014) *Comma Sutra*; Lil Duke (2017) *Light My Blunt*). Another telling example is the growing practice of exchanging sexual intimacy (dating, with or without actual sexual intercourse) for financial support (often for university tuition fees), mostly by gay male teens and straight girls in their twenties with older 'sugar daddies', which has become institutionalised through international websites and agencies. Any brief exploration of empirical data on SES and sexual partner preferences shows various significant

effects. For example, especially concerning long-term relationships, women are found to prefer wealthy men to such a degree that it equals or outweighs physical attractiveness. (Greitemeyer, 2005; Townsend & Levy, 1990) Besides functioning in the private sphere, attitudes towards SES cause various harms in the public sphere too. While race, sexual orientation, and especially sex and gender might be more vocally discussed nowadays, class still constitutes, and even increasingly so, the most significant inequalities and injustices, concerning professional opportunity, education opportunity, legal resources, social networks, and many other matters. For example, besides being able to afford better legal representation, high-SES relates favourably to juror judgements. (Mazzella & Feingold, 1994) As such, all of the harm effects that we can identify from preferences concerning race, orientation, identity, or physique, we can find in class preferences too, and note that any racial, sexual, or physique-based income inequality dwarfs compared to class differences. Moreover, the long and on-going history of class discrimination makes that sexual classism may well be a form of sexual discrimination unlike any other.

Other traits: There are countless other traits that sexual preferences can relate to, and going into each goes beyond the scope of this essay. Some important undiscussed traits include age, physical and mental ability, and intelligence. Analysing these, the harms will likely look different for each trait, and some will have more explicit histories of discrimination than others (notably 'ableism'). One consideration that may make a come-back here, however, might be the function-argument of goal-fulfilment, as one could possibly argue how these traits are necessary in sharing certain interests or activities that people could experience as essential to their identity and partnership. Moreover, it is important to note that all of the traits discussed here may interact with one another in a multitude of ways, which implores us to embrace an 'intersectional' approach in any further discussion of the topic.

3C: The agency of sexual preferences

On a dynamic developmental view of sexual preferences, rather than being '*just* a personal preference', the configuration of sexual preferences is understood not as

fixed, but plastic, susceptible to social and personal influences, and thus enabling control over the development of preferences in virtue of which controversial sexual preferences are unjustified forms of discrimination, and agents can be held morally responsible.

Sexual orientation and sexual preference: Firstly, the empirical data discussed in section 2b, if studied more closely, only addresses sexual *orientations*, such as homosexuality, as predetermined and unchangeable.[76] This is where the theoretical distinctions from section 1a become decisive, because these findings do not straightforwardly warrant similar conclusions about sexual *preferences* being equally fixed. Not all of human sexuality can be reduced to the same mechanisms that may govern sexual orientation. For example, both biological and social factors drive the experience of love (DeLamater & Hyde, 1998; Walster & Berscheid, 1974; White et al., 1981), the direction and expression of desire (Tuzin, 1995), cross-cultural variability of sexuality (Parker, 2009), and intra-individual variability such as changing degrees of sexual desire (Kinsey et al., 1953, p. 680)

Preference fluidity: As said, while not endorsing exclusively sociocultural theories in opposition to essentialist perspectives, dynamic approaches do allow for sociocultural factors to play a significant role in shaping sexual preferences. Therefore, let us now explore if there is any empirical substantiation of preference malleability. There is only scarce data on the development of sexual preferences. (Hekma, 1991, p. 11)

[76] Sexual orientation may not be strictly fixed either. Notably, the Kinsey Scale was originally intended to capture same-sex and other-sex orientation on a continuum, including intra-individual variability on this scale over time. (Kinsey et al., 1954) Lisa Diamond provides a solid, contemporary discussion of sexual orientation, finely balancing fluidity around a malleable core with the lack of choice, due to the complex interaction of factors. (Diamond, 2008, pp. 246-253)

This is where Zheng's advocacy for changing people's sexuality is severely flawed and even harmful, since, besides failing to provide any support for various strategies, Zheng lumps together sexual orientation with sexual preference and sexual identity by substantiating her claims through drawing on the work of Joyce Trebilcot, who argues that sexual orientation is a choice. (Trebilcot, 2009; Zheng, 2016, p. 415) Based on exactly such lumping the classical liberal view warrants conclusions of general lack of choice due to 'lumped fixedness', but similarly 'lumped fluidity' warrants conclusions by supports of the earlier mentioned anti-gay conversion therapies (2b).

In order to not make such unwarranted generalisations and avoid unintended support for harmful conversion practices, an adequately nuanced view of sexual development is needed, which is why I focus exclusively on certain sexual preferences in this essay and leave the matter of sexual orientation and sexual identity aside.

One point worth noting is that, while it may turn out that sexual orientation has a degree of fluidity too, since this is probably significantly less controllable than sexual preference fluidity, moral wrongness and responsibility for sexual orientations cannot similarly be argued for on the basis of developmental control. For an interesting discussion of the moral wrongness of sexual orientations, see (Ayres & Brown, 2011).

However, the following examples from sociology and social psychology are meaningful.

Sociologists have documented studied a wide range of topics related to attractiveness in a variety of ways. For example, looking at female beauty icons, preferences concerning body-type have been found to have changed radically over the last 150 years, from the long-time ideal of 'plumpness' to current thinness norms, which even now very vastly per subculture, for example among African American thinness is not nearly as endorsed. (Berry, 2007, p. 5) Another intriguing approach is a study of masculinity ideals through comparing male action figures from different times, which revealed a clear increase in muscularity, at present even far exceeding the largest real bodybuilding, which is taken as a reflection of the evolving standard of male body attractiveness. (Pope et al., 1999) Other interesting historical developments are the views of masculinity from high-heeled, wealthily-adorned elegance to contemporary roughness and minimalistic absence of decoration. Or, as direct example of sexual preferences, the replacement of fur-fetish in times where fur was often represented as wealth, to leather-fetish, with the rise of leather usage by soldiers, pilots, and other new heroes. (Hekma, 1991)

Social psychological studies can provide us with more insight into not merely the changing of phenomena, but the attitudes that people hold towards these. A by now large body of research evidences the significant influence of one's environment, contact with people, socioeconomic status, and level of education on sexual attitudes concerning both who people find attractive and what actions people prefer. For example, higher educated, middle-class people typically have more positive attitudes towards sexual variety such as masturbation, petting, oral sex, compared to lower educated, lower-class people. (Fisher & Byrne, 1981; Kinsey et al., 1954; Kinsey et al., 1948; Schmidt & Sigusch, 1971; Schofield, 1965) Notably, not merely how favourable people view, or how much interest they show, but even their level of arousal, for example concerning novel sexual experiences, is mediated by the level of education. (Michael et al., 1994). Directly concerning the main preferences explored in this essay, there are telling findings too. For example, concerning racial preferences, many studies have shown environmental influences, such as that people growing up more racially diverse environments have sexual preferences that are more positively related to various races. (Fisman et al., 2008) And, beyond racial composition, people growing up in places with more tolerant attitudes towards

interracial marriage show weaker same-race dating preferences. (Fisman et al., 2008) Concerning gender identity, people without internet access were found to prefer more feminine men and more masculine women, compared to people with internet access who prefer more masculine men and feminine women. (Batres & Perrett, 2014) And concerning body type, increased preferences for thinness have been found to associate with internet access too. (Batres & Perrett, 2014) Moreover, body type preferences have been shown to be influenced by acculturation even on an intra-individual level over a relatively short time span. A study on Zulu men found that those living in their country of origin had a stronger preference for heavier women, compared to those who had (even only recently) emigrated to the UK, who preferred slimmer women. (Tovée et al., 2006)

Admittedly, none of this research provides grounds for any conclusive statements about the malleability of sexual preferences, the degree to which they are malleable, what factors play causal roles in shaping them, and how the mechanism involved exactly works. However, it does provide a strong indication for the likelihood of a significant degree malleability and influence by factors such as exposure to people and information. This is exactly what dynamic models hope to achieve, since appreciating the complexity of sexual development rules out any definite causal predictions. Plausibility of developmental hypotheses is what dynamic we aim for, with which we can interpret found patterns of development, and propose methods that are likely to cause different development. Moreover, while there is only scarce data on sexual development, whereby a reader may be still unconvinced of the fluidity of sexual preferences, I direct that reader to work on the fluidity of other, non-sexual attitudes, including implicit biases, unconscious stereotypes, non-sexual preferences, and others, of which there is an immense body of research that evidences malleability. In essay 2 of my doctoral thesis I elaborately explore a general model for the development of moral automaticity, which the model in the current essay is derivative of. Many others have explored somewhat similar projects. (Fricker, 2007; Hogarth, 2001; Holroyd & Kelly, 2016; Sauer, 2012; Saul, 2013) Appreciating this, the challenge I pose this sceptical reader is to refute the meaningful likeness of sexual attitudes and other attitudes that is argued throughout this essay.[77]

[77] As counterargument, one could raise the intuition that sexual preferences are not as fluid as other attitudes due to their centrality in evolutionary selection. It is unclear to me, however, what sort of sexual preferences would be subject to this evolutionary fixedness. For example, from an evolutionary perspective it could be thought that in-group preferences would prevail, but studies on partner selection show this not to be the case. Alternatively, body

The first point that we can make based on data such as the above, is that sexual preferences are fluid; their configuration is not predetermined, nor unchangeable, which refutes the fixedness model of the classical liberal view. Secondly, experiential (e.g. contact with people and exposure to media) and informational (e.g. prevailing attitudes of others) stimuli are among the factors that can influence the configuration of sexual preferences. What we can draw from this is that selected exposure to certain stimuli is likely to cause changes in people's sexual preferences. The most fascinating example of this is a study by Emily Harris and colleagues on sexual preferences concerning partner dominance, as I take the idea that women prefer dominant men and men prefer submissive women 'as a function of natural evolution' to be one of the most mainstream essentialist claims about sexual preferences. Contrary to this dogma, however, and in line with a dynamic view that appreciates fluidity, the study showed that erotica stories in which a woman is dominant caused similar arousal as male-dominance ones, and, moreover, reading the female-dominance story eradicated the typical dominance preferences and caused both men and women to prefer equal partner dominance. (Harris et al., 2017)

Intentionally directed preference fluidity: The fluidity through experiential and informational stimuli brings us to one of the crucial claims of this essay, which is that fluidity implies the potential for intentionally directed fluidity. This claim is fairly straightforward, merely inferring that, since sexual preferences are partly configured by formative stimuli, how people's preferences develop can be intentionally directed through selecting what type of formative stimuli people are exposed to.

An interesting background for this view can be found in the work of John Money, one of the most influential sexologists of the last century, who advanced the model of sexual preferences, or 'turn-on patterns', as he referred to, as 'lovemaps'; templates in the brain that connect various objects, practices, and situations that people acquire through experience.[78] (Money, 1986) Recently, Esben Esther Benestad and colleagues revived Money's lovemaps model, with an addition exactly in the mind of the argument advanced here. Based on their clinical practice as sexologists, they argue that, rather than merely through accidental and unconscious exposure to

type preference could be thought of as evolutionarily important, but this too greatly varies through time and culture. Also partner dominance preferences, as discussed above, is commonly advanced as evolutionary central, but seems fluid.

[78] While this is a too radical social constructivist approach, we need not take experience as *only* or *chief* factor, but among a complex of factors.

certain stimuli, through time, people are able to expand their sexual preferences as to include other objects, practices, or situations, by consciously choosing stimuli exposure. As such, it is possible for people to "cultivate their particular sexual turn-on patterns." (Benestad et al., 2015, p. 27)

While there may well be much variety in how much time and effort certain developments take and how well they turn out, varying per person, per preference, and per stimuli strategy, and while much more research on this is needed, we can nevertheless plausibly conclude from the model of development argued for in this entire section that there is a meaningful possibility to intentionally direct sexual preference fluidity, meaning that, through selecting stimuli that have (albeit not fully-determinative, but) significant effects, there are (albeit no certain, but) likely outcomes of particular preference configurations.

Formative stimuli: An important step for preference development is to analyse what stimuli currently influence the development of people's sexual preferences. A now large body of research examining the development of implicit biases has reached a general consensus that the origins of such associations lie in the direct and indirect information that people receive throughout life, notably from personal experiences, media exposure, and cultural ideas. (Castelli et al., 2009; Dasgupta, 2013; Kirwan Institute, 2016; Rudman, 2004a, 2004b) Similarly, for sexual preferences, while not fully determinative, and while there are no certain outcomes, the patterns of clear preferences that societies develop can be illuminated by looking at the stimuli that people are exposed to.

Concerning race, it is remarkable how in major Western films, most central characters with a multidimensional, individualised, intelligent, successful, and desirable representation are White. The Hollywood Report found that almost 90% of Hollywood films had a White lead actor, and over half had a minority cast of 10% or less. (HDR, 2014, p. 6) Also looking at pornography, studies of patterns in homosexual pornography reveal clear portrayal of White beauty standards. (Kendall, 1997; Morrison, 2004; Morrison et al., 2007) In relation to sexual orientation, there are copious cultural ideas that stigmatise bisexuals as partner. For example, bisexuals are seen as having bad personality traits, over-sexualised promiscuous natures, incapable of monogamy, and being untrustworthy partners, although research shows this is not the case. (Spalding & Peplau, 1997) Also, in particular bisexual men are

stereotyped as a heightened risk for HIV transmission. (Hansen & Evans, 1985; Klesse, 2011) Moreover, bisexuality is often seen as a phrase, rather than a proper orientation. Regarding gender identity, heteronormative masculinity ideals are culturally importantly tied up with, for example, muscularity, and unrealistic male body ideals are increasingly depicted in media. (Frederick et al., 2016; Lanzieri & Hildebrandt, 2011) And women are typically displayed passively in pornography. (Cabrera & Ménard, 2013) In relation to physique, media, in movies, television shows, advertising, etc., generally employ a very narrow range of beauty norms, not displaying people of colour, heavier, older, or disabled people as sexual. (Berry, 2007)

Developmental strategies: A further step in preference development is mapping, designing, and organising developmental strategies, importantly through alternative stimuli that people can expose themselves to. A growing body of research supports various possible ways in which all sorts of attitudes are malleable, which can be combined in multifaceted strategies for attitude development, reducing prejudices, biases, and stereotypes. (Blair, 2002; Dasgupta, 2013; Dasgupta & Asgari, 2004; Dasgupta & Greenwald, 2001; Joy-Gaba & Nosek, 2010; Lai et al., 2014; Rudman et al., 2001) Similar efforts have to be organised concerning sexual attitudes, which can be inspired by existing literature.

One group of strategies can be labelled ‘cognitive strategies’. Metacognition, the awareness of, understanding of, analytical reflection on, and training of one’s own cognitive processes, has been found to be an important factor in judgements, which can counteract biases. (Croskerry, 2014; Croskerry et al., 2013; Jenicek, 2010; Mamede et al., 2010; Vohs et al., 2007; Whaley & Geller, 2007) For example, awareness can be created through taking implicit association tests, and labelling the attitudes that one discovers. (Devine et al., 2012) This is crucial, because much of the preferences that people hold are implicit, so awareness of having preferences, and detecting them, is a crucial step. Furthermore, new information, especially when this prompts reinterpretation of knowledge, changes biases. (Mann & Ferguson, 2015) And decoupling, for example through ‘considering the opposite’ of one’s judgements, and hence seeking out evidence in support of this, can reduce bias. (Bishop & Trout, 2004; Larrick, 2004; Lord et al., 1984)

Another set is ‘experiential strategies’, through increased contact, individuating people, dialogue, perspective-taking, and counter-stereotyping. Actively

seeking out opportunities for positive interactions with members of a certain group challenges evaluations of the group and changes cognitive representations, which ameliorates biases. (Pettigrew, 1998; Pettigrew & Tropp, 2006) Detailing specific information about particular people as a person prevents stereotypic group-based inferences. (Brewer, 1988; Fiske & Neuberg, 1990) Dialogue and perspective taking: Adopting the perspective of a member of a group, and relating to it on a personal level, creates personal closeness, which ameliorates stereotypic group-based attitudes. (Galinsky & Moskowitz, 2000) Conversations about transgender rights, involving active perspective-taking (relating to participants' personal experience with vulnerability) significantly reduced prejudice. (Broockman & Kalla, 2016; Paluck, 2016) And even playing a video game in which participants had to navigate a predominantly White university from the perspective of a Black student can reduce prejudice. (Shih et al., 2013; Todd et al., 2011; Todd & Galinsky, 2014) And exposure to individuals who counter certain stereotypes, and thus challenge the validity, reduce biases and make positive exemplars more salient as alternative cognition. (Blair et al., 2001; Dasgupta & Greenwald, 2001) Finally, an important strategy can be changing one's media diet, consuming stimuli that enforce other images of people by representing them in a different light (or by representing them at all). For example, regarding developing sexual and romantic partner preferences, self-developing through exposure to alternative representations could involve a wide range of strategies. For example, watching certain films like *Moonlight*, directed by Barry Jenkins, about a Black homosexual boy, or *Hiroshima Mon Amour*, by Alain Resnais, and *In the Mood for Love*, by Wong Kar-Wai, both about romantic affairs involving Asian characters; or *Hidden Figures* by Theodore Melfi depicting African American women as intelligent scientists and reliable partners, or watching series like *Dear White People,* reading particular novels, like *The Lover*, by Marguerite Duras, about a romance between a French girl and a Chinese man, or even possibly viewing other types of pornography.

Moral responsibility: The classical liberal view employs a theory of moral responsibility that focuses on control through conscious voluntary choice or reflective endorsement. There are two responses to this, proposing alternative theories or extending volitionist theories. While a proper discussion of responsibility is beyond

the scope of this essay, I take the following to provide a sufficient argument to refute the classical liberalist view of volitionism.

There is a range of alternative approaches to moral responsibility, most notably 'Attributionism'. (Adams, 1985; Arpaly, 2003; Scanlon, 1998; Sher, 2009; Smith, 2005, 2008) Attributionism is rapidly gaining support, especially in relation to debates concerning attitudes of the kind as unconscious stereotypes and implicit biases, which applies well to sexual preferences given important similarities: Acquisition through formative stimuli (a person develops an implicit bias/sexual preference partly through acculturation), a structure that is interconnected with other attitudes (an implicit bias/sexual preference usually relates to a person's other fitting attitudes), and operation with typically little direct conscious control (even though a person may not endorse the bias/sexual preference it will often still influence their judgements). Rather than control as the condition for moral responsibility, attributionism focuses on the appropriateness of attributing an action or mental state to a person as expressing someone's 'real self', their deeply held evaluative appraisals. As Angela Smith, plausibly the most prominent attributionist, writes: "It is undoubtedly true that the implicit biases most of us harbor are largely a result of cultural influences over which we have little control, this does not change the fact that they are now evaluations we are making that have an influence on our judgments, attitudes, and actions. My claim is not that we are morally responsible for coming to have an implicit bias, but that we are morally responsible for having and manifesting such biases, for the simple reason that we are morally responsible for anything that reflects our evaluative activity." (Smith, unpublished, p. 21) Translating this to sexual preferences, since controversial sexual preferences usually associate with further attitudes (such as negative attitudes towards multiculturalism, beliefs about gender norms, or dispositions on body norms), and are "relatively stable dispositions to respond in particular ways to particular situations" (Smith, 2005, pp. 251, footnote twenty-seven), they can well be taken to reflect one's evaluative activity, such that the person can be held morally responsible for it. In the case where a person explicitly disavows endorsing their preference, they are still responsible and blameworthy, but just not blameworthy for an additional higher-order attitude such as not caring about it or approving of it. (Smith, unpublished, p. 26)

A related approach to moral responsibility that may be worthwhile in this context comes from Elinor Mason, who argues that people are responsible for implicit

biases since in most cases they are way of having a bad quality of will, as they are tied up one's deep motivations such as disgust or dispositions such as accepting privileging hierarchies. (Mason, forthcoming, p. 5) Additionally, even when such quality of will is not involved (say, sexual preferences that are entirely unconnected to other attitudes), one can nevertheless 'take on' responsibility, as to appropriately respect one's fellow community members. (Mason, forthcoming, p. 12) Yet another related approach can be found in the work of Jules Holroyd and Daniel Kelly, who argue that a person's unconscious attitudes can be taken as belonging to the moral character of a person, and since we typically hold people responsible for their character, they are responsible for their unconscious attitudes as well. (Holroyd & Kelly, 2016)

An entirely different like of argument to refute the classical liberal argument that people are not responsible is by somewhat buying into volitionism, but disputing that people lack a meaningful sense of control over their sexual preferences. I argue for such an account elsewhere (my doctoral thesis, essay 3). This requires an extension of what volitionists typically take as being under one's control. Besides for 'direct control', such as a conscious choice in the moment to help a person in need, volitionism allows for responsibility through 'indirect control'. (Levy, 2005, p. 2) For example, a doctor who neglects reading up on new research may lack direct control during a subsequent procedure due to lacking relevant knowledge, but this does not excuse responsibility since the doctor had the opportunity to acquire the knowledge at an earlier stage. (Rosen, 2004, p. 303) What is crucial, for volitionists, is that control can be 'traced back' to a point where certain outcomes of one's actions are 'foreseeable'. (Vargas, 2005, p. 274) Now, since we have ample evidence that sexual preferences can be developed with likely outcomes, I argue that exposing oneself to formative stimuli makes it foreseeable that one will develop certain sexual preferences over time. As such, the possibility of intentionally directed preference fluidity enables a meaningful sense of control over one's controversial sexual preferences, in virtue of which one has moral responsibility for them; *developmental control*.[79]

[79] Volitionists may be unwilling to stretch indirect control as far as to include attitude development. While I do not see any principle within the theory against it, I recognise that it both significantly shifts the theory's central notion, control, as to focus on indirect strategies, as well radically change the action-model with automatic attitudes as focal point. In virtue of this, this approach could possibly be better presented as a distinct theory, a third alternative, next to volitionism and attributionism, called *developmentalism*.

In conclusion, I have argued that, while there is no decisive empirical evidence, as far as there is data available, this indicates that the classical liberal view of fixed and modular preferences in incorrect. Rather, a dynamic model of sexual cognition seems true, on which controversial sexual preferences are interconnected with a person's other attitudes and admit to a significant degree of fluidity in which experiential and informational stimuli have a significant influence. Hereby, people have a meaningful sense of control over developing their preferences through intentionally exposing themselves to stimuli that are likely to reconfigure their preferences in a certain way, in virtue of which they have moral responsibility for their preferences.

Given developmental control, controversial sexual preferences that cause grave harms and function in the perpetuation of systemic discrimination cannot be argued to be justified in virtue of being necessary to sexual fulfilment, since they can be developed as to be of such configuration that their fulfilment does not cause said harms. Moreover, it is because of the gravity of the harms involved, direct harms as well as indirect ones, that the classical liberal cannot simply waive the harms away as being irrelevant, to make room for sexual freedom. Rather, people have a moral responsibility to engage in changing their preferences, challenging the dominant, discriminative norms typically presented in our mainstream culture, and aim to develop more inclusive, respectful sexual preferences.

4: Conclusion

In this final section I address some further topics that have not come up throughout the essay. The main objective of this essay is to propose an understanding of what sexual preferences are from a dynamic model of cognition, and proposing an understanding of moral responsibility for controversial sexual preferences from a theory of developmental control. Crucially, much more research is needed to fully understand sexual preferences. This requires an effort somewhat in the spirit of John Money, who utilised his model of lovemap development to compile a remarkably extensive catalogue of various sexual preferences and their links with cultural influences. (Money, 1986) In the same mind, but based on dynamic models and

including developmental responsibility, we should catalogue controversial sexual preferences, categorising the various sorts, studying their functional mechanisms, analysing their impact (harms), surveying developmental strategy possibilities, evaluating moral responsibility for them, and organising opportunities for people to work on their preferences.

Individual and collective responsibility: The focus of this essay has been on individual moral responsibility. However, this is no way implies that this is not equally, or potentially even more, a matter to be addressed under collective responsibility at a structural, institutional level. For example, many of the discussed influential factors are not under the control of individuals, but rather collective or political, such as housing and schooling segregation to create more interaction, and legislating institutional procedures such as blind résumé reviewing and school grading, stimulating more diverse representation in public functions and media, and organising educational programmes. Elizabeth Emens and Deborah Rhode provide interesting discussions of collective responsibility and legal changes that are required concerning sexual racism and lookism. (Emens, 2009; Rhode, 2009, 2010) The dynamic developmental approach argued for in this essay may provide a good basis to address collective responsibility as well, providing an understanding of what sexual preferences are.

Public campaigns are one example where collective and individual responsibility meet. While the National Association to Advance Fat Acceptance has been campaigning to change the public image of a variety of body types since the 1970s, similar movements advancing equal sexual treatment concerning other traits still have much needed growth ahead. Some noteworthy examples do exist, such as the project 'Be Switched On', in which homosexual and lesbian people of colour share personal experiences in order to raise awareness about how racial sexual stereotypes affect them and how each is an individual that deviates from stereotypes.[80] Another example is 'Look Different', an online project concerning racial, gender, and LGBT+ biases, which employs celebrities like Rick Ross, Ashanti, and Ja Rule, to inform people about biases and even a seven-day programme that aims at creating

[80] Be Switched On is an initiative by Big Up Together, a UK charity. Some interviews can be viewed on www.youtube.com/watch?v=GukuwC9x6eE, www.youtube.com/watch?v=iT6al3lhBrc, and www.youtube.com/watch?v=EB3g5hFEFxQ.

awareness of people's own biases and challenging these.[81] And a last example is a website exclusive dedicated to sexual racism called 'Sexual Racism Sux', explaining the harms and proposing less offensive communicative strategies.[82]

Degrees of responsibility and blame: Another set of considerations concerns how, and how much to hold individuals responsible. These questions require metaethical discussion that does cannot be detailed here, so I only briefly mention some. For one, on a developmental account, I conjecture that responsibility can come in degrees, depending on the opportunity an individual has to engage in self-development. Besides possible variation in developmental possibility per preference-type, individuals have different possibilities themselves to engage in developmental efforts. For example, differences in environment, available time, means, and related abilities may mediate the degree of responsibility. As such, socioeconomic conditions may matter (financial hardship may hinder personality change (Anger et al., 2017; Soto, 2015; Specht et al., 2013)), or even biological sex may matter (sexual preference fluidity may be different between sexes (Baumeister, 2000; Diamond, 2008)). Moreover, an upshot of general resource limitation that any person is subject to may be that only so much developmental effort can be reasonably required of people. One implication of this could be that for reasons of impact, people should focus attention on the most harmful preferences. While such selection does not underwrite any strict hierarchy of the phenomenology of harms, it is likely to entail that some people take their lived experience to be underappreciated. As such, pragmatic considerations such as resource allocation may require choices that need careful deliberation and much more input from respective communities.

Another issue is the difference between *being* responsible and blame*worthy*, and whether we also should *hold* someone responsible and *blame* a person. The second question requires further considerations, for example concerning effectiveness of change rather than culpability of wrongs. Emens, for example, does not address whether people *are* responsible, but merely argues they should not be *held* responsible for their preferences, although these do constitute harms and discrimination. As reason, she posits that holding people responsible may have 'perverse consequence', with which she alludes to imposing psychiatric conversion therapy on people of the

[81] Look Different is developed by MTV, in cooperation with the Kirwan Institute for the Study of Race and Ethnicity. See www.lookdifferent.com.
[82] See www.sexualracismsux.com.

sort imposed on homosexuals. (Emens, 2009, pp. 1356-1357, 1359) In the same mind, she emphasises that, while people can undertake ethical measures against biases, they should not be morally judged for how they do so. (Emens, 2009, p. 1361) Now, while such measures can hardly be called 'ethical' as such, it is important to notice that this stance is based entirely on mere philosophical intuitions about the effects of holding people responsible, and ignores significant metaethical nuances in the practice of blame. In contrast, I think certain ways of holding people responsible are not merely justified, but also necessary for change. Firstly, we can distinguish between blaming someone for *having* certain controversial preferences, and fully *being,* say, a racist, sexist, biphobe, transphobe, or fatphobe, because these latter can be said to include additional higher-order attitudes of explicit endorsement of the initial sexual attitude. Since such labels are imbued with strong social condemnation, which people typically want to avoid, I agree that these types of blaming may cause unwillingness to acknowledge and work on one's preferences. (Rapley, 2001; Saul, 2013) More nuanced responsibility, however, for possessing preferences and working on them, may well be beneficial, for example to induce engagement, generate concern, and stimulate motivation. (Croskerry et al., 2013; Larrick, 2004) Also here, more research is required to inform our best practice.

To conclude, two points addressed to potential readers that might be shocked by this essay into thinking they lost their entire sexual freedom. Firstly, I in no way propose that anyone has a moral obligation or responsibility to have sexual intercourse with anyone that do not want to. Rather, I argued that people have a moral responsibility to challenge attitudes that may have concerning who is sexually desirable. Secondly, I in no way argue that people are 'not allowed to have any preferences any more' and should 'just find everyone equally attractive'. Rather, I argued that certain preferences that concern generic traits that typically have little to do with someone's individuality as a human being can be objectionable. There are literally countless other, more personal traits left whereby people can select their partners. What we can call 'essential partner qualities', concern traits that accommodate individuation, connection with one another, understanding one another, and sharing experiences with one another. For example, shared ideas, or fascinating interests, mutual activities, admirable character traits, certain capacities, someone's actions, and even

particular physical features (for we are, after all, embodied beings[83]), and all of these can vary vastly along with each people's particular personality and lifestyle.

[83] For an amazing sci-fi story on entirely abolishing the perception and influence of facial physical beauty, see Chiang's *Liking What You See: A Documentary* (Chiang, 2002).

Bibliography

Abernathy, C., & Hamm, R. M. (1995). *Surgical intuition: What it is and how to get it*: Hanley & Belfus.

Adams, R. M. (1985). Involuntary sins. *The Philosophical Review, 94*(1), 3-31.

Allgeier, E. R., & Wiederman, M. W. (1994). How useful is evolutionary psychology for understanding contemporary human sexual behavior? *Annual review of sex research, 5*(1), 218-256.

Anderson, E. (2010). *The imperative of integration*: Princeton University Press.

Anger, S., Camehl, G., & Peter, F. (2017). Involuntary job loss and changes in personality traits. *Journal of Economic Psychology, 60*, 71-91.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy, 33 (124)*(124), 1-19.

APA. (1991). American Psychological Association: Committee on Lesbian and Gay Concerns. Avoiding Heterosexual Bias in Language. *American psychologist, 46*(9), 973-974.

APA, A. P. A. (2006). *Statement of the American Psychological Association*. Retrieved from Retrieved at June 1 2017 from https://web.archive.org/web/20110514111323/http://www.apa.org/pi/lgbt/resources/policy/ex-gay.pdf:

APA, A. P. A. (2008). Sexual orientation and homosexuality.

APA, A. P. A. (2009). *Report of the American Psychological Association Task Force on Appropriate Therapeutic Responses to Sexual Orientation*. Retrieved from Retrieved from http://www.apa.org/pi/lgbc/publications/therapeutic-resp.html:

Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*: Oxford University Press.

Atkins, R., Hart, D., & Donnelly, T. M. (2004). Moral Identity Development and School Attachment. In D. K. Lapsley & D. Narvaez (Eds.), *Moral Development, Self, and Identity* (pp. 65-82).

Auwarter, A. E., & Aruguete, M. S. (2008). Effects of student gender and socioeconomic status on teacher perceptions. *The Journal of Educational Research, 101*(4), 242-246.

Ayala, G., Bingham, T., Kim, J., Wheeler, D. P., & Millett, G. A. (2012). Modeling the impact of social discrimination and financial hardship on the sexual risk of HIV among Latino and Black men who have sex with men. *American Journal of Public Health, 102*(S2), S242-S249.

Ayres, I., & Brown, J. G. (2011). *Straightforward: How to mobilize heterosexual support for gay rights*: Princeton University Press.

Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PloS one, 9*(2), e88616.

Bailey, J. M., Kim, P. Y., Hills, A., & Linsenmeier, J. A. (1997). Butch, femme, or straight acting? Partner preferences of gay men and lesbians. *Journal of personality and social psychology, 73*(5), 960.

Banaji, M., & Greenwald, G. (2013). Blindspot: Hidden biases of good people [Kindle iPad version].

Bancroft, J. (2002). Biological factors in human sexuality. *Journal of Sex Research, 39*(1), 15-21.

Bancroft, J. (2009). *Human sexuality and its problems*: Elsevier Health Sciences.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational behavior and human decision processes, 50*(2), 248-287.

Bandura, A. (1997). *Self-efficacy: The exercise of control*: Macmillan.
Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. *Handbook of social cognition, 1*, 1-40.
Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American psychologist, 54*(7), 462.
Barlett, C. P., Vowels, C. L., & Saucier, D. A. (2008). Meta-analyses of the effects of media images on men's body-image concerns. *Journal of Social and Clinical Psychology, 27*(3), 279-310.
Baron, J. (1993). *Morality and Rational Choice*. Dordrecht, The Netherlands: Kluwer.
Baron, J. (1995). A psychological view of moral intuition. *The Harvard Review of Philosophy, 5*(1), 36-40.
Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological bulletin, 130*(4), 553.
Bartky, S. L. (1990). *Femininity and domination: Studies in the phenomenology of oppression*: Psychology Press.
Batres, C., & Perrett, D. I. (2014). The influence of the digital divide on face preferences in El Salvador: People without internet access prefer more feminine men, more masculine women, and women with higher adiposity. *PloS one, 9*(7), e100966.
Batson, C. D. (1998). Altruism and prosocial behavior. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (Vol. 2, pp. 282-316). Boston: McGraw-Hill.
Battle, J., & Lewis, M. (2002). The increasing significance of class: The relative effects of race and socioeconomic status on academic achievement. *Journal of poverty, 6*(2), 21-35.
Baumeister, R. F. (2000). Gender differences in erotic plasticity: the female sex drive as socially flexible and responsive. *Psychological bulletin, 126*(3), 347.
Beckstead, A. L. (2003). Understanding the self-reports of reparative therapy "successes.". *Archives of Sexual Behavior, 32*(5), 421-423.
Benestad, E. E. P., Almas, E., & Weingarten, K. (2015). Sex positive ways of perceiving sexual turn-on patterns: part I-understanding. *International Journal of Narrative Therapy & Community Work*(1), 26.
Berry, B. (2007). *Beauty bias: Discrimination and social power*: Greenwood Publishing Group.
Berthoz, S., Armony, J., Blair, R., & Dolan, R. (2002). An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain, 125*(8), 1696-1708.
Besser-Jones, L. (2008). Social psychology, moral character, and moral fallibility. *Philosophy and Phenomenological Research, 76*(2), 310-332.
Binder, J., Zagefka, H., Brown, R., Funke, F., Kessler, T., Mummendey, A., . . . Leyens, J.-P. (2009). Does contact reduce prejudice or does prejudice reduce contact? A longitudinal test of the contact hypothesis among majority and minority groups in three European countries. *Journal of personality and social psychology, 96*(4), 843.
Bishop, M. A., & Trout, J. D. (2004). *Epistemology and the psychology of human judgment*: Oxford University Press.
Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*(3), 242-261.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological science, 15*(10), 674-679.
Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of personality and social psychology, 83*(1), 5.
Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of personality and social psychology, 81*(5), 828.
Blanchflower, D. G., & Oswald, A. J. (2004). Money, sex and happiness: An empirical study. *The Scandinavian Journal of Economics, 106*(3), 393-415.
Blasi, A. (2009). The Moral Functioning of Mature Adults and the Possibility of Fair Moral Reasoning. In D. Narvaez & D. K. Lapsley (Eds.), *Personality, Identity, and Character: Explorations in Moral Psychology* (pp. 196-440). New York: Cambridge University Press.
Blincoe, S., & Harris, M. J. (2009). Prejudice reduction in white students: Comparing three conceptual approaches. *Journal of Diversity in Higher Education, 2*(4), 232.
Blumstein, P. W., & Schwartz, P. (1976). Bisexuality in men. *Urban Life, 5*(3), 339-358.
Brewer, M. B. (1988). A dual process model of impression formation. *Advances in social cognition, 1*.
Brink, D. O., & Nelkin, D. (2013). Fairness and the Architecture of Responsibility. *Oxford Studies in Agency and Responsibility, 1*, 284-313.
Brinkhurst-Cuff, C. (2017). Is Love Racist? The TV show laying our biases bare *The Guardian*.
British Psychological Society (BPS). (2014). Masculinity still viewed as tied to sexuality.
Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *science, 352*(6282), 220-224.
Brosnan, K. (2011). Do the evolutionary origins of our moral beliefs undermine moral knowledge? *Biology & Philosophy, 26*(1), 51-64.
Brown, B. (2012). *Daring greatly: How the courage to be vulnerable transforms the way we live, love, parent, and lead*: Gotham.
Brown, K. T., Brown, T. N., Jackson, J. S., Sellers, R. M., & Manuel, W. J. (2003). Teammates on and off the field? Contact with black teammates and the racial attitudes of white student athletes. *Journal of Applied Social Psychology, 33*(7), 1379-1403.
Brown, R., & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in experimental social psychology, 37*(37), 255-343.
Brownstein, M. (2016). Context and the ethics of implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy 2: Moral Responsibility, Structural Injustice, and Ethics* (pp. 215-234). Oxford: Oxford University Press.
Brownstein, M., & Madva, A. (2012a). Ethical automaticity. *Philosophy of the social sciences, 42*(1), 68-98.
Brownstein, M., & Madva, A. (2012b). The normativity of automaticity. *Mind & Language, 27*(4), 410-434.
Burman, B., & Margolin, G. (1992). Analysis of the association between marital relationships and health problems: an interactional perspective. *Psychological bulletin, 112*(1), 39.

Buss, D. M. (1991). Evolutionary personality psychology. *Annual review of psychology, 42*(1), 459-491.
Buss, D. M. (1998). Sexual strategies theory: Historical origins and current status. *Journal of Sex Research, 35*(1), 19-31.
Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: an evolutionary perspective on human mating. *Psychological review, 100*(2), 204.
Byrne, G. (2015). Coming to Terms With Being a Working Class Academic. Retrieved from https://www.sociologylens.net/article-types/opinion/coming-to-terms-with-being-working-a-working-class-academic/14799
Cabrera, C., & Ménard, A. D. (2013). "She exploded into a million pieces": a qualitative and quantitative analysis of orgasms in contemporary romance novels. *Sexuality & Culture, 17*(2), 193-212.
Callander, D. (2013). *Just a preference : exploring concepts of race among gay men looking for sex or dates online.* (Ph.D.), The University of New South Wales, Australia, Kensington, Australia. Retrieved from http://handle.unsw.edu.au/1959.4/53044
Callander, D., Holt, M., & Newman, C. E. (2012). Just a preference: Racialised language in the sex-seeking profiles of gay and bisexual men. *Culture, health & sexuality, 14*(9), 1049-1063.
Callander, D., Newman, C. E., & Holt, M. (2015). Is sexual racism really racism? Distinguishing attitudes toward sexual racism and generic racism among gay and bisexual men. *Archives of Sexual Behavior, 44*(7), 1991-2000.
Caluya, G. (2006). The (gay) scene of racism: Face, shame and gay Asian males. *Australian Critical Race and Whiteness Studies Association e-Journal, 2*(2), 1-14.
Carnes, M., Devine, P. G., Isaac, C., Manwell, L. B., Ford, C. E., Byars-Winston, A., . . . Sheridan, J. (2012). Promoting institutional change through bias literacy. *Journal of Diversity in Higher Education, 5*(2), 63.
Carnes, M., Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., . . . Magua, W. (2015). Effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. *Academic medicine: journal of the Association of American Medical Colleges, 90*(2), 221.
Carrier, J. M. (1985). Mexican male bisexuality. *Journal of Homosexuality, 11*(1-2), 75-86.
Carruthers, P. (2006). *The architecture of the mind*: Oxford University Press.
Carston, R. (1996). The architecture of the mind: modularity and modularization. In D. Green (Ed.), *Cognitive Science: An Introduction* (pp. 53-83). Cambridge: Blackwell.
Castelli, L., Zogmaister, C., & Tomelleri, S. (2009). The transmission of racial attitudes within the family. *Developmental psychology, 45*(2), 586.
Chaiken, S. (1987). *The heuristic model of persuasion.* Paper presented at the Social influence: the ontario symposium.
Chaiken, S., Giner-Sorolla, R., & Chen, S. (1996). Beyond accuracy: Defense and impression motives in heuristic and systematic information processing.
Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73-96). New York: Guilford Press.
Chiang, T. (2002). Liking What You See: A Documentary *Stories of Your Life and Others*: Tor Books, US.

Cho, S. K. (1997). Converging stereotypes in racialized sexual harassment: Where the model minority meets Suzie Wong. *J. Gender Race & Just., 1*, 177.
Chou, R. S. (2012). *Asian American sexual politics: The construction of race, gender, and sexuality*: Rowman & Littlefield.
Chou, R. S., Lee, K., & Ho, S. (2012). The White Habitus and Hegemonic Masculinity at the Elite Southern University: Asian Americans and the Need for. *Sociation Today, 10*(2).
Chou, R. S., Lee, K., & Ho, S. (2015). Love Is (Color) Blind: Asian Americans and White Institutional Space at the Elite University. *Sociology of Race and Ethnicity, 1*(2), 302-316.
Chris Brown, Lil Wayne, & Tyga. (2014). Loyal. On *X*. NYC: RCA Records.
Christ, O., Hewstone, M., Tausch, N., Wagner, U., Voci, A., Hughes, J., & Cairns, E. (2010). Direct contact as a moderator of extended contact effects: Cross-sectional and longitudinal impact on outgroup attitudes, behavioral intentions, and attitude certainty. *Personality and Social Psychology Bulletin, 36*(12), 1662-1674.
Christman, S. S. (2002). Dynamic Systems Theory: Application to language development and acquired aphasia. In R. G. Daniloff (Ed.), *Connectionist approaches to clinical problems in speech and language: Therapeutic and scientific applications* (pp. 111-146). Hillsdale, NJ: Lawrence Erlbaum Associates.
Chu, K. C. (2017, Accessed 2017-06-06). How Queer People of Color Are Combating Sexual Racism. *VICE*.
Churchland, P. M. (1988). Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor. *Philosophy of Science, 55*(2), 167-187.
CJay. (2016). Sauce Drip. On *Sauce Drip*. Independent Artist.
Clark, A. (2010). Memento's Revenge: The Extended Mind. In R. Menary (Ed.), *The Extended Mind* (pp. 43-67). Cambridge, MA: MIT Press.
Colby, A., & Damon, W. (1992). Some do care: Contemporary lives of moral commitment: New York: Free Press.
Coleman, N. A. T. (2011). What? What? In the (black) butt. *The Newsletter on Philosophy and Lesbian, Gay, Bisexual, and Transgender Issues, 11*(1), 12-15.
Coleman, P., & Watson, A. (2000). Infant Attachment as a Dynamic System1. *Human Development, 43*(6), 295-313.
Cooper, J. M., & Cooper, J. M. (1975). *Reason and human good in Aristotle*: Hackett Publishing.
Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of personality and social psychology, 83*(6), 1314.
Cosmides, L., & Tooby, J. (2004). Knowing thyself: The evolutionary psychology of moral reasoning and moral sentiments. In R. E. Freeman & P. Werhane (Eds.), *Business, Science, and Ethics: The Ruffin Series of the Society for Business Ethics* (Vol. 4, pp. 93-128). Charlottesville: Society for Business Ethics.
Croskerry, P. (2014). ED cognition: any decision by anyone at any time. *CJEM, 16*(1), 13-19.
Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 2: impediments to and strategies for change. *BMJ Qual Saf*, bmjqs-2012-001713.
Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain*. London: Penguin Book.

Damon, W., & Colby, A. (1992). Some do care: Contemporary lives of moral commitment: New York: Free Press.
Dang, J. (2017). Is there an alternative explanation to the evolutionary account for financial and prosocial biases in favor of attractive individuals? *Behavioral and brain sciences, 40*.
Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in experimental social psychology, 47*, 233-279.
Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of experimental social psychology, 40*(5), 642-658.
Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of personality and social psychology, 81*(5), 800.
Dasgupta, N., & Rivera, L. M. (2006). From automatic antigay prejudice to behavior: The moderating role of conscious beliefs about gender and behavioral control. *Journal of personality and social psychology, 91*(2), 268.
Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes: A meta-analytic review. *Personality and Social Psychology Review, 15*(4), 332-351.
DeCecco, J. P., & Elia, J. P. (1993). A critique and synthesis of biological essentialism and social constructionist views of sexuality and gender. *Journal of Homosexuality, 24*(3-4), 1-26.
Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review, 95*(2), 158-165.
DeLamater, J. D., & Hyde, J. S. (1998). Essentialism vs. social constructionism in the study of human sexuality. *Journal of Sex Research, 35*(1), 10-18.
Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology, 48*(6), 1267-1278.
Diamond, L. M. (2008). *Sexual fluidity*: Harvard University Press.
Diener, E., Gohm, C. L., Suh, E., & Oishi, S. (2000). Similarity of the relations between marital status and subjective well-being across cultures. *Journal of cross-cultural psychology, 31*(4), 419-436.
Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of personality and social psychology, 24*(3), 285.
Doris, D., & Stich, S. (2005). As a Matter of Fact: Empirical Perspectives on Ethics. In F. Jackson & M. Smith (Eds.), *The Oxford Handbook of Contemporary Philosophy* (pp. 114-152): Oxford University Press.
Doris, J. M. (1998). Persons, situations, and virtue ethics. *Noûs, 32*(4), 504-530.
Doris, J. M. (2002). *Lack of character: Personality and moral behavior*: Cambridge University Press.
Dovidio, J. F., Eller, A., & Hewstone, M. (2011). Improving intergroup relations through direct, extended and other forms of indirect contact. *Group Processes & Intergroup Relations, 14*(2), 147-160.
Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education, 77*(4), 267-282.
Drescher, J. (2002). Sexual conversion ("reparative") therapies: History and update. In B. E. J. M. J. Hill (Ed.), *Mental health issues in lesbian, gay, bisexual, and*

transgender communities (pp. 71-91). Arlington, VA: American Psychiatric Publishing.
Dreyfus, H. L., & Dreyfus, S. E. (1991). Towards a phenomenology of ethical expertise. *Human Studies, 14*(4), 229-250.
Druckman, J. N. (2004). Political preference formation: Competition, deliberation, and the (ir) relevance of framing effects. *American Political Science Review, 98*(4), 671-686.
Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing Black: Race, crime, and visual processing. *Journal of personality and social psychology, 87*(6), 876.
Edwards, K., Carter-Tellison, K. M., & Herring, C. (2004). For Richer, For Poorer: Whether Dark or Light: Skin Tone, Marital Status, and Spouse's Earnings. In C. Herring, V. M. Keith, & H. D. Horton (Eds.), *Skin/Deep: How Race and Complexion Matter in the "Color-Blind" Era* (pp. 65-81). Chicago, IL: University of Illinois.
Eguchi, S. (2009). Negotiating hegemonic masculinity: The rhetorical strategy of "straight-acting" among gay men. *Journal of Intercultural Communication Research, 38*(3), 193-209.
Eisenberg, D., Golberstein, E., & Hunt, J. B. (2009). Mental health and academic success in college. *The BE Journal of Economic Analysis & Policy, 9*(1).
Eiser, J. R. (1994). *Attitudes, chaos and the connectionist mind*: Blackwell Publishing.
Elder, W. B., Morrow, S. L., & Brooks, G. R. (2015). Sexual Self-Schemas of Gay Men: A Qualitative Investigation Ψ. *The Counseling Psychologist, 43*(7), 942-969.
Eller, A., Abrams, D., & Gomez, A. (2012). When the direct route is blocked: The extended contact pathway to improving intergroup relations. *International Journal of Intercultural Relations, 36*(5), 637-646.
Emens, E. F. (2009). Intimate Discrimination: The State's Role in the Accidents of Sex and Love. *Harvard Law Review*, 1307-1402.
Europe, R. (2017). *Understanding mental health in the research environment*. Retrieved from https://royalsociety.org/%7E/media/policy/topics/diversity-in-science/understanding-mental-health-in-the-research-environment.pdf
Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science, 8*(3), 223-241.
Faraci, D., & Shoemaker, D. (2010). Insanity, deep selves, and moral responsibility: The case of JoJo. *Review of philosophy and psychology, 1*(3), 319-332.
Faraci, D., & Shoemaker, D. (2014). Huck vs. JoJo: moral ignorance and the (a) symmetry of praise and blame. *Oxford Studies in Experimental Philosophy, 1*, 7-27.
Farmer, J. J., Jr., & Zoubek, P. H. (1999). *Final Report of the State Police Review Team*. Retrieved from http://www.state.nj.us/lps/Rpt_ii.pdf:
Fausto-Sterling, A. (2000). *Sexing the body: Gender politics and the construction of sexuality*: Basic Books.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*: Cambridge university press.
Fischer, J. M., & Tognazzini, N. A. (2009). The truth about tracing. *Noûs, 43*(3), 531-556.
Fisher, W., & Byrne, D. (1981). Social background, attitudes, and sexual attraction. *The bases of sexual attraction*, 23-63.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in experimental social psychology, 23*, 1-74.
Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2008). Racial preferences in dating. *The Review of Economic Studies, 75*(1), 117-132.
FitzPatrick, W. J. (2008). Moral responsibility and normative ignorance: Answering a new skeptical challenge. *Ethics, 118*(4), 589-613.
Flanders, C. E., & Hatfield, E. (2014). Social perception of bisexuality. *Psychology & Sexuality, 5*(3), 232-246.
Flusberg, S. J., & McClelland, J. L. (2014). Connectionism and the emergence of mind. *The Oxford handbook of cognitive science, 1*.
Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*: MIT press.
Foundation, N. S. (2015). *Doctorate Recipients from U.S. Universities: 2014*. Retrieved from https://www.nsf.gov/statistics/2016/nsf16300/digest/nsf16300.pdf
Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy LXVIII*, 5-20.
Frankowski, B. L., & American Academy of Pediatrics Committee on Adolescence. (2004). Sexual orientation and adolescents. *Pediatrics, 133*(6), 1827-1832.
Frederick, D. A., Sandhu, G., Morse, P. J., & Swami, V. (2016). Correlates of appearance and weight satisfaction in a US national sample: Personality, attachment style, television viewing, self-esteem, and life satisfaction. *Body image, 17*, 191-203.
Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*: Oxford University Press.
Fridland, E. (2017). Automatically minded. *Synthese, 194*(11), 4337-4363.
Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological science, 7*(4), 237-241.
Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61-90). Orland, FL: Academic Press.
Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *science, 321*(5892), 1100-1102.
Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of personality and social psychology, 78*(4), 708.
Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of experimental social psychology, 44*(2), 370-377.
Gendler, T. S. (2008). Alief and belief. *The Journal of Philosophy, 105*(10), 634-663.
Gerhardstein, K. R. (2010). *Attitudes toward transsexual people: Effects of gender and appearance.* Indiana State University.
Gerhardstein, K. R., & Anderson, V. N. (2010). There's more than meets the eye: Facial appearance and evaluations of transsexual people. *Sex roles, 62*(5-6), 361-373.

Gigerenzer, G. (2008). Moral Intuition = Fast and Frugal Heuristics? In A. Sinnott, W. (Ed.), *Moral Psychology. Vol 2: The cognitive sciene of morality: Intuition and diversity* (pp. 1-26). Cambridge, M.A.: MIT Press.

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia, 50*(14), 3600-3611.

Ginet, C. (2000). The Epistemic Requirements for Moral Responsibility. *Noûs, 34*(s14), 267-277.

Godsil, R. D., Tropp, L. R., Goff, P. A., & Powell, J. A. (2014). Addressing implicit bias, racial anxiety, and stereotype threat in education and health care. *The Science of Equality, 1*.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American psychologist, 54*(7), 493.

Gollwitzer, P. M., Gallo, I. S., Keil, A., McCulloch, K. C., & Rockstroh, B. (2009). Strategic automation of emotion regulation. *Journal of personality and social psychology, 96*(1), 11.

Gómez, A., Tropp, L. R., & Fernández, S. (2011). When extended contact opens the door to future contact: Testing the effects of extended contact on attitudes and intergroup expectancies in majority and minority groups. *Group Processes & Intergroup Relations, 14*(2), 161-173.

Gorna, R. (1996). Vamps, Virgins and Victims: How Women Can Fight AIDS: London: Cassell.

Govorun, O., & Payne, B. K. (2006). Ego—depletion and prejudice: separating automatic and controlled components. *Social Cognition, 24*(2), 111-136.

Graham, S., & Lowery, B. S. (2004). Priming unconscious racial stereotypes about adolescent offenders. *Law and human behavior, 28*(5), 483.

Granic, I. (2005). Timing is everything: Developmental psychopathology from a dynamic systems perspective. *Developmental Review, 25*(3), 386-407.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *science, 293*(5537), 2105-2108.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review, 102*(1), 4.

Greitemeyer, T. (2005). Receptivity to sexual offers as a function of sex, socioeconomic status, physical attractiveness, and intimacy of the offer. *Personal Relationships, 12*(3), 373-386.

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

Haidt, J. (2003). The emotional dog does learn new tricks: A reply to Pizarro and Bloom (2003).

Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about morality. In A. Sinnott, W. (Ed.), *Moral psychology, Vol. 2: The cognitive science of morality: Intuition and diversity, p181-217*. Cambridge, U.S.: MIT Press.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*.

Halwani, R. (2017). Racial Sexual Desires. In A. Noble, R. Halwani, S. Hoffman, & J. M. Held (Eds.), *The Philosophy of Sex: Contemporary Readings, 7th ed.* Lanham, Md.: Rowman & Littlefield.

Hamermesh, D. S. (2011). *Beauty pays: Why attractive people are more successful*: Princeton University Press.
Han, A. (2006a). I think you're the smartest race I've ever met: Racialised economies of queer male desire. *Australian Critical Race and Whiteness Studies Association ejournal, 2*(2), 1-14.
Han, C.-s. (2006b). Geisha of a different kind: Gay Asian men and the gendering of sexual identity. *Sexuality & Culture, 10*(3), 3-28.
Han, C.-s. (2007). They don't want to cruise your type: Gay men of color and the racial politics of exclusion. *Social Identities, 13*(1), 51-67.
Hansen, C. E., & Evans, A. (1985). Bisexuality reconsidered: An idea in pursuit of a definition. In F. A. Klein & T. Wolf (Eds.), *Bisexualities: Theory and research* (pp. 1-6). New York, NY: Haworth Press.
Hardy, S. A., & Carlo, G. (2011). Moral identity: What is it, how does it develop, and is it linked to moral action? *Child Development Perspectives, 5*(3), 212-218.
Harman, G. (1999). *XIV—Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error.* Paper presented at the Proceedings of the Aristotelian Society.
Harris, E. A., Thai, M., & Barlow, F. K. (2017). Fifty shades flipped: Effects of reading erotica depicting a sexually dominant woman compared to a sexually dominant man. *The Journal of Sex Research, 54*(3), 386-397.
HDR, H. D. R. (2014). *2014 Hollywood Diversity Report: Making Sense of the Disconnect*. Retrieved from
Heberle, V. (2003). Modularity versus Connectionism: two different views on the architecture of the mind. *Fragmentos: Revista de Língua e Literatura Estrangeiras, 25*.
Heekeren, H. R., Wartenburger, I., Schmidt, H., Schwintowski, H.-P., & Villringer, A. (2003). An fMRI study of simple ethical decision-making. *Neuroreport, 14*(9), 1215-1219.
Heiser, M., Iacoboni, M., Maeda, F., Marcus, J., & Mazziotta, J. C. (2003). The essential role of Broca's area in imitation. *European Journal of Neuroscience, 17*(5), 1123-1128.
Hekma, G. (1991). *A history of sexology: Social and historical aspects of sexuality*.
Hennen, P. (2008). *Faeries, bears, and leathermen: Men in community queering the masculine*: University of Chicago Press.
Henss, R. (2001). Social perceptions of male pattern baldness. A review. *Dermatology and Psychosomatics/Dermatologie und Psychosomatik, 2*(2), 63-71.
Heyes, C. J. (2006). Changing Race, Changing Sex: The Ethics of Self‐Transformation. *Journal of Social Philosophy, 37*(2), 266-282.
Hogarth, R. M. (2001). *Educating intuition*: University of Chicago Press.
Hogarth, R. M. (2003). *Educating intuition: a challenge for the 21st century*: CREI, Centre de Recerca en Economia Internacional.
Hogarth, R. M. (2006). Is Confidence in Decisions Related to Feedback? Evidence from Random Samples of Real-World Behaviour. In K. Fiedler & P. Jusline (Eds.), *Information Sampling and Adaptive Cognition* (pp. 456-479): Cambridge University Press.
Holroyd, J., & Kelly, D. (2016). Implicit Bias, Character and Control. In A. Masala & J. Webber (Eds.), *From Personality to Virtue: Essays on the Philosophy of Character*: Oxford University Press.

Horgan, T., & Timmons, M. (2007). Morphological rationalism and the psychology of moral judgment. *Ethical theory and moral practice, 10*(3), 279-295.
Horton, K. (2004). Aid and bias. *Inquiry, 47*(6), 545-561.
Hovdhaugen, E. (2013). Widening participation in Norwegian higher education. *Oslo: NIFU*.
Huebner, B. (2009). Troubles with stereotypes for spinozan minds. *Philosophy of the social sciences, 39*(1), 63-92.
Hutcheson, F., & Moor, J. (2008). *The Meditations of the Emperor Marcus Aurelius Antoninus*: Liberty Fund.
Hysenbegasi, A., Hass, S. L., & Rowland, C. R. (2005). The impact of depression on the academic productivity of university students. *Journal of Mental Health Policy and Economics, 8*(3), 145.
Ibrahim, A. K., Kelly, S. J., Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of psychiatric research, 47*(3), 391-400.
Izard, C. E., Ackerman, B. P., Schoff, K. M., & Fine, S. E. (2000). Self-organization of discrete emotions, emotion patterns, and emotion-cognition relations. In M. D. Lewis & I. Granic (Eds.), *Emotion, development, and self-organization: Dynamic systems approaches to emotional development* (pp. 15-36). New York: Cambridge University Press.
Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition, 97*(2), 211-225.
Jenicek, M. (2010). *Medical error and harm: Understanding, prevention, and control*: CRC Press.
Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *science, 310*(5745), 116-119.
Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in organizational behavior, 29*, 39-69.
Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology*.
Kahn, K. B., & Davies, P. G. (2011). Differentially dangerous? Phenotypic racial stereotypicality increases implicit bias among ingroup and outgroup members. *Group Processes & Intergroup Relations, 14*(4), 569-580.
Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.
Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of personality and social psychology, 78*(5), 871.
Kendall, C. N. (1997). Gay male pornography after little sisters book and art emporium: A call for gay male cooperation in the struggle for sex equality. *Wis. Women's LJ, 12*, 21.
Kennett, J., & Fine, C. (2009). Will the real moral judgment please stand up? *Ethical theory and moral practice, 12*(1), 77-96.
Kersey-Matusiak, G. (2012). *Delivering culturally competent nursing care*: Springer Publishing Company.

Kessels, J., Boers, E., & Mostert, P. (2002). *Vrije ruimte: filosoferen in organisaties: klassieke scholing voor de hedendaagse praktijk*: Boom.
King, R. (2013). The uncomfortable racial preferences revealed by online dating. *Quartz*.
Kinsey, A. C., Pomeroy, W. B., Martin, C. E., & Gebhard, P. H. (1953). Sexual behavior in the human female.
Kinsey, A. C., Pomeroy, W. B., Martin, C. E., & Gebhard, P. H. (1954). Sexual behavior in the human female: JSTOR.
Kinsey, A. C., Pomeroy, W. B., Martin, C. E., & Sloan, S. (1948). Sexual behavior in the human male.
Kirwan Institute. (2016). *State of the Science: Implicit Bias Review 2016*. Retrieved from
Klesse, C. (2011). Shady characters, untrustworthy partners, and promiscuous sluts: Creating bisexual intimacies in the face of heteronormativity and biphobia. *Journal of Bisexuality, 11*(2-3), 227-244.
Kloos, H., & Orden van, G. C. (2009). Heidi Kloos, and Guy C. Van Orden. In J. Spencer (Ed.), *Toward a Unified Theory of Development Connectionism and Dynamic System Theory Re-Consider*. Oxford: Oxford University Press.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*(279), 190-194.
Kohlberg, L. (1973). The claim to moral adequacy of a highest stage of moral judgment. *The Journal of Philosophy, 70*(18), 630-646.
Korsgaard, C. (2008). *The Constitution of Agency: Essays on Practical Reason and Moral Philosophy*. Oxford: Oxford University Press.
Kraemer, B., Delsignore, A., Schnyder, U., & Hepp, U. (2008). Body image and transsexualism. *Psychopathology, 41*(2), 96-100.
Krueger, J., & Funder, D. (2004). Towards a balanced social psychology: causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *The Behavioral and Brain Sciences, 27*(3), 313-327.
Kühn, S., Haggard, P., & Brass, M. (2014). Differences between endogenous and exogenous emotion inhibition in the human brain. *Brain structure and function, 219*(3), 1129-1138.
Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., . . . Koleva, S. P. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765.
Lamanna, M. A., Riedmann, A., & Stewart, S. D. (2014). *Marriages, families, and relationships: Making choices in a diverse society*: Cengage Learning.
Långström, N., Rahman, Q., Carlström, E., & Lichtenstein, P. (2010). Genetic and environmental effects on same-sex sexual behavior: A population study of twins in Sweden. *Archives of Sexual Behavior, 39*(1), 75-80.
Lanzieri, N., & Hildebrandt, T. (2011). Using hegemonic masculinity to explain gay male attraction to muscular and athletic men. *Journal of Homosexuality, 58*(2), 275-293.
Lapsley, D. K., & Narvaez, D. (2004). A social-cognitive approach to the moral personality. In D. K. Lapsley & D. Narvaez (Eds.), *Moral development, self and identity* (pp. 189-206): Fribaum, NJ.
Larrick, R. P. (2004). Debiasing. *Blackwell handbook of judgment and decision making*, 316-338.
Lentin, A. (2011). *Racism and ethnic discrimination*: The Rosen Publishing Group.

Levy, N. (2005). The good, the bad and the blameworthy. *J. Ethics & Soc. Phil., 1*, 1.
Levy, N. (2011). Expressing who we are: Moral responsibility and awareness of our reasons for action. *Analytic Philosophy, 52*(4), 243-261.
Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. *Noûs, 48*(1), 21-40.
Levy, N. (2016). Implicit Bias and Moral Responsibility: Probing the Data. *Philosophy and Phenomenological Research*.
Levy, N., & McKenna, M. (2009). Recent work on free will and moral responsibility. *Philosophy Compass, 4*(1), 96-133.
Lewis, M. D. (1995). Cognition-emotion feedback and the self-organization of developmental paths. *Human Development, 38*(2), 71-102.
Lewis, M. D. (2000). The promise of dynamic systems approaches for an integrated account of human development. *Child Development, 71*(1), 36-43.
Lickliter, R. (2008). The growth of developmental thought: Implications for a new evolutionary psychology. *New Ideas in Psychology, 26*(3), 353-369.
Little, A. C. (2017). Evolutionary explanations for financial and prosocial biases: Beyond mating motivation. *Behavioral and brain sciences, 40*.
Lombardi, W. J., Higgins, E. T., & Bargh, J. A. (1987). The role of consciousness in priming effects on categorization: Assimilation versus contrast as a function of awareness of the priming task. *Personality and Social Psychology Bulletin, 13*(3), 411-429.
Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology, 47*(6), 1231.
Macmillan, M. (2002). *An odd kind of fame: Stories of Phineas Gage*: MIT Press.
Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research, 84*(1), 24-62.
Maestripieri, D., Henry, A., & Nickels, N. (2017). Explaining financial and prosocial biases in favor of attractive people: Interdisciplinary perspectives from economics, social psychology, and evolutionary psychology. *Behavioral and brain sciences, 40*.
Magai, C., & McFadden, S. H. (1995). *The role of emotions in social and personality development: History, theory and research* (Vol. 1): Springer Science & Business Media.
Mahajan, R. (2007). The naked truth: Appearance discrimination, employment, and the law. *Asian Am. LJ, 14*, 165.
Malebranche, D. J., Fields, E. L., Bryant, L. O., & Harper, S. R. (2009). Masculine socialization and sexual risk behaviors among Black men who have sex with men: A qualitative exploration. *Men and Masculinities, 12*(1), 90-112.
Mamede, S., van Gog, T., van den Berge, K., Rikers, R. M., van Saase, J. L., van Guldener, C., & Schmidt, H. G. (2010). Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *Jama, 304*(11), 1198-1203.
Mandelbaum, E. (2013). Against alief. *Philosophical Studies, 165*(1), 197-211.
Mandelbaum, E. (2014). Thinking is believing. *Inquiry, 57*(1), 55-96.
Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of personality and social psychology, 108*(6), 823.

Mareschal, D., Leech, R., & Cooper, R. P. (2009). Combining Connectionist and Dynamic Systems Principles in Models of Development: The Case of Analogical Completion. In J. Spencer (Ed.), *Toward a Unified Theory of Development Connectionism and Dynamic System Theory Re-Consider*. Oxford: Oxford University Press.
Marger, M. (2012). *Race and ethnic relations: American and global perspectives*: Wadsworth/Thomson Learning Belmont, CA.
Martin, L. L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of personality and social psychology, 51*(3), 493.
Martin, L. L., Seta, J. J., & Crelia, R. A. (1990). Assimilation and contrast as a function of people's willingness and ability to expend effort in forming an impression. *Journal of personality and social psychology, 59*(1), 27.
Mason, E. (forthcoming). Taking Responsibility for our Biases.
Matheson, J. (2012). I'm a sexual racist. *Sydney Star Observer*.
Mautz, B. S., Wong, B. B., Peters, R. A., & Jennions, M. D. (2013). Penis size interacts with body shape and height to influence male attractiveness. *Proceedings of the National Academy of Sciences, 110*(17), 6925-6930.
Mazzella, R., & Feingold, A. (1994). The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims on judgments of mock jurors: A meta‐analysis. *Journal of Applied Social Psychology, 24*(15), 1315-1338.
Mazziotta, A., Mummendey, A., & Wright, S. C. (2011). Vicarious intergroup contact effects: Applying social-cognitive theory to intergroup contact research. *Group Processes & Intergroup Relations, 14*(2), 255-274.
McCauley, R. N., & Henrich, J. (2006). Susceptibility to the Müller-Lyer illusion, theory-neutral observation, and the diachronic penetrability of the visual input system. *Philosophical Psychology, 19*(1), 79-101.
McDowell, J. (1996). *Mind and world*: Harvard University Press.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.
McLaughlin, B. (1990). “Conscious” versus “unconscious” learning. *TESOL quarterly, 24*(4), 617-634.
McMurray, B., Horst, J. S., C., T. J., & Samuelson, L. K. (2009). Integrating Connectionist Learning and Dynamical Systems Processing: Case Studies in Speech and Lexical Development. In J. Spencer (Ed.), *Toward a Unified Theory of Development Connectionism and Dynamic System Theory Re-Consider*. Oxford: Oxford University Press.
Mele, A. R. (2006). *Free will and luck*: Oxford University Press.
Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin, 36*(4), 512-523.
Merritt, M. (2000). Virtue ethics and situationist personality psychology. *Ethical theory and moral practice, 3*(4), 365-383.
Meyerowitz, J. (1993). Beyond the feminine mystique: A reassessment of postwar mass culture, 1946-1958. *The Journal of American History, 79*(4), 1455-1482.
Michael, R. T., Gagnon, J. H., Laumann, E. O., & Kolata, G. (1994). Sex in America: A definitive survey.
Miller, C. B. (2013). *Moral character: An empirical theory*: Oxford University Press.
Miller, P. H. (2002). *Theories of Developmental Psychology* (4th ed.). New York: Worth Publishers.

Mills, C. W. (1994). Do Black Men Have a Moral Duty to Marry Black Women? *Journal of Social Philosophy, 25*(s1), 131-153.
Minerva, F. (2017). The Invisible Discrimination Before Our Eyes: A Bioethical Analysis. *Bioethics, 31*(3), 180-189.
Mitchison, D., Hay, P., Griffiths, S., Murray, S. B., Bentley, C., Gratwick‐Sarll, K., . . . Mond, J. (2017). Disentangling body image: the relative associations of overvaluation, dissatisfaction, and preoccupation with psychological distress and eating disorder behaviors in male and female adolescents. *International Journal of Eating Disorders, 50*(2), 118-126.
Money, J. (1986). *Lovemaps: Clinical concepts of sexual/erotic health and pathology, paraphilia, and gender transposition of childhood, adolescence, and maturity*: Ardent Media.
Monteith, M. J., Sherman, J. W., & Devine, P. G. (1998). Suppression as a stereotype control strategy. *Personality and Social Psychology Review, 2*(1), 63-82.
Monteith, M. J., & Voils, C. I. (1998). Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies. *Journal of personality and social psychology, 75*(4), 901.
Morrison, T. G. (2004). "He was treating me like trash, and I was loving it…" Perspectives on gay male pornography. *Journal of Homosexuality, 47*(3-4), 167-183.
Morrison, T. G., Morrison, M. A., & Bradley, B. A. (2007). Correlates of gay men's self-reported exposure to pornography. *International Journal of Sexual Health, 19*(2), 33-43.
Mosbergen, D. (2016, Accessed 2017-06-06). Online Dating Is Rife With Sexual Racism, 'The Daily Show' Discovers. *Huffington Post*.
Moshman, D. (2005). *Adolescent psychological development: Rationality, morality, and identity*: Psychology Press.
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474-16479.
Musschenga, B. (2008). Moral Judgement and Moral Reasoning. In M. Düwell, C. Rehmann-Sitter, & D. Mieth (Eds.), *The Contingent Nature of Life: Bioethics and the Limits of Human Existence* (pp. 131-146): Springer.
Musschenga, B. (2011). The Relevance of Conscious Moral Reasoning. In D. J. De Ruyter & S. Miedema (Eds.), *Moral Education and Development* (pp. 71-83). Rotterdam: Sense Publishers.
Naess, A. (1966). *Communication and argument: Elements of applied semantics* (T. f. N. b. A. Hannay;, Trans.). London: Allen & Unwin.
Narvaez, D. (2011). Neurobiology, moral education and moral self-authorship. In D. J. De Ruyter & S. Miedema (Eds.), *Moral Education and Development* (pp. 31-44). Rotterdam: Sense Publishers.
Nelkin, D. K. (2005). Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy, 29*(1), 181-206.
Nelkin, D. K. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs, 50*(2), 356-378.
Nemoto, K. (2009). *Racing romance: Love, power, and desire among Asian American/White couples*: Rutgers University Press.
Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*: Oxford University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review, 84*(3), 231.
Nucci, L. P., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development*, 400-407.
Nussbaum, M. C. (1992). *Love's knowledge: Essays on philosophy and literature*: OUP USA.
Ochsner, K. N., & Gross, J. J. (2008). Cognitive emotion regulation: Insights from social cognitive and affective neuroscience. *Current Directions in Psychological Science, 17*(2), 153-158.
OkCupid. (2014). *Race and Attraction 2009-2014: What's changed in five years?* Retrieved from https://theblog.okcupid.com/race-and-attraction-2009-2014-107dcbb4f060:
Page-Gould, E., Mendoza-Denton, R., & Tropp, L. R. (2008). With a little help from my cross-group friend: Reducing anxiety in intergroup contexts through cross-group friendship. *Journal of personality and social psychology, 95*(5), 1080.
Pain, E. (2014). Breaking the Glass Ceiling. Retrieved from http://www.sciencemag.org/careers/2014/05/breaking-class-ceiling
Paluck, E. L. (2016). How to overcome prejudice. *science, 352*(6282), 147-147.
Park, H. (2012). Interracial violence, Western racialized masculinities, and the geopolitics of violence against women. *Social & Legal Studies, 21*(4), 491-509.
Parker, R. (2009). Sexuality, culture and society: shifting paradigms in sexuality research. *Culture, health & sexuality, 11*(3), 251-266.
Partridge, T. (2005). Are genetically informed designs genetically informative? Comment on McGue, Elkins, Walden, and Iacono (2005) and quantitative behavioral genetics.
Patel, N. (2009). Racialized Sexism in the Lives of Asian American Women. In C. Raghavan, A. E. Edwards, & K. M. Vaz (Eds.), *Benefiting by Design: Women of Color in Feminist Psychological Research* (pp. 116-128): Cambridge Scholars Publisher.
Payne, B. K. (2005). Conceptualizing control in social cognition: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality Social Psychology, 81*(181), 92.
Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education*. NJ: Erlbaum.
Pettigrew, T. F. (1998). Intergroup contact theory. *Annual review of psychology, 49*(1), 65-85.
Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology, 90*(5), 751.
Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta‐analytic tests of three mediators. *European Journal of Social Psychology, 38*(6), 922-934.
Pettigrew, T. F., & Tropp, L. R. (2013). *When groups meet: The dynamics of intergroup contact*: Psychology Press.
Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage, 16*(2), 331-348.
Phua, V. C., & Kaufman, G. (2003). The crossroads of race and sexuality: Date selection among men in internet "personal" ads. *Journal of Family Issues, 24*(8), 981-994.

Piaget, J. (1932). The moral development of the child. *Kegan Paul, London*.
Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001).
Plant, E. A., & Devine, P. G. (2003). The antecedents and implications of interracial anxiety. *Personality and Social Psychology Bulletin, 29*(6), 790-801.
Pollard, B. (2003). Can virtuous actions be both habitual and rational? *Ethical theory and moral practice, 6*(4), 411-425.
Pope, H. G., Olivardia, R., Gruber, A., & Borowiecki, J. (1999). Evolving ideals of male body image as seen through action toys. *International Journal of Eating Disorders, 26*(1), 65-72.
Powell, J. A. (2012). *Racing to justice: Transforming our conceptions of self and other to build an inclusive society*: Indiana University Press.
Prinz, J. (2006a). The emotional basis of moral judgments. *Philosophical Explorations, 9*(1), 29-43.
Prinz, J. (2007). *The emotional construction of morals*: Oxford University Press.
Prinz, J. J. (2006b). Is the mind really modular? In R. Stainton (Ed.), *Contemporary Debates in Cognitive Science* (pp. 22-36). Oxford: Blackwell.
Prochaska, J. O., DiClemente, C. C., & Norcross, J. C. (1993). In search of how people change: Applications to addictive behaviors. *Addictions Nursing Network, 5*(1), 2-16.
Pylyshyn, Z. (1984). Computation and cognition: Toward a theory for cognitive science: MIT Press.
Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and brain sciences, 22*(3), 341-365.
Rahman, Q., & Wilson, G. D. (2003). Born gay? The psychobiology of human sexual orientation. *Personality and Individual Differences, 34*(8), 1337-1382.
Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics, 124*(4), 813-859.
Rapley, M. (2001). How to do X without doing Y: Accomplishing discrimination without 'being racist'—'doing equity'. In M. Augoustinos & K. J. Reynolds (Eds.), *Understanding prejudice, racism and social conflict* (pp. 231-250). London: SAGE Publications.
Read, S. J., & Miller, L. C. (2002). Virtual personalities: A neural network model of personality. *Personality and Social Psychology Review, 6*(4), 357-369.
Rest, J., Narvaez, D., Bebeau, M., & Thoma, S. (1999). A neo-Kohlbergian approach: The DIT and schema theory. *Educational Psychology Review, 11*(4), 291-324.
Rhode, D. L. (2009). The injustice of appearance. *Stanford Law Review*, 1033-1101.
Rhode, D. L. (2010). *The beauty bias: The injustice of appearance in life and law*: Oxford University Press.
Ridge, D., Hee, A., & Minichiello, V. (1999). Asian" men on the scene: Challenges to "gay communities. *Journal of Homosexuality, 36*(3-4), 43-68.
Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of general psychology, 133*(1), 19-35.
Rist, R. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard educational review, 40*(3), 411-451.
Robinson, R. K. (2007). Structural dimensions of romantic preferences. *Fordham L. Rev., 76*, 2787.

Rosen, G. (2003). *Culpability and Ignorance.* Paper presented at the Proceedings of the Aristotelian Society.
Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives, 18*(1), 295-313.
Ross, L., & Nisbett, R. (1991). *The Person and the Situation*: Pinter & Martin Ltd
Royal College of Psychiatrists (RCoP). *Royal College of Psychiatrists' Position Statement on Sexual Orientation*. Retrieved from Retrieved June 1 2017 from http://www.rcpsych.ac.uk/pdf/RCPsychposstatementsexorientation.pdf:
Rudman, L. A. (2004a). Social justice in our minds, homes, and society: The nature, causes, and consequences of implicit bias. *Social Justice Research, 17*(2), 129-142.
Rudman, L. A. (2004b). Sources of implicit attitudes. *Current Directions in Psychological Science, 13*(2), 79-82.
Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). "Unlearning" automatic biases: the malleability of implicit prejudice and stereotypes. *Journal of personality and social psychology, 81*(5), 856.
Rust, P. C. R. (2006). Reparative science and social responsibility: The concept of a malleable core as theoretical challenge and psychological comfort. In J. D. K. J. Zucker (Ed.), *Ex-gay research: Analyzing the Spitzer study and its relation to science, religion, politics, and culture* (pp. 171-177). Binghamton, NY: Haworth Press.
Sabini, J., Siepmann, M., & Stein, J. (2001). The Really Fundamental Attribution Error in Social Psychological Research. *Psychological inquiry, 12*(1), 1-15.
Sauer, H. (2012). Educated intuitions. Automaticity and rationality in moral judgement. *Philosophical Explorations, 15*(3), 255-275.
Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: What needs to change?* (pp. 39-60).
Saul, J. (2017). Are generics especially pernicious? *Inquiry*, 1-18.
Saygın, A. P., Dick, F., W. Wilson, S., F. Dronkers, N., & Bates, E. (2003). Neural resources for processing language and environmental sounds: evidence from aphasia. *Brain, 126*(4), 928-945.
Scanlon, T. (1998). *What we owe to each other*: Harvard University Press.
Schelling, T. C. (1984). *Choice and consequence*: Harvard University Press.
Schmidt, G., & Sigusch, V. (1971). Patterns of sexual behavior in West German workers and students∗. *Journal of Sex Research, 7*(2), 89-106.
Schneider, W., Dumais, S. T., & Shiffrin, R. M. (1984). Automatic and control processing and attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 12-17). San Diego: Academic Press.
Schofield, M. G. (1965). *The sexual behaviour of young people*: Little, Brown.
Schumann, J. H. (1990). Extending the scope of the acculturation/pidginization model to include cognition. *TESOL quarterly, 24*(4), 667-684.
Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology, 45*(3), 513.
Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly, 91*(4), 531-553.
Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box *New essays on belief* (pp. 75-99): Springer.

Seldin, D. R., Friedman, H. S., & Martin, L. R. (2002). Sexual activity as a predictor of life-span mortality risk. *Personality and Individual Differences, 33*(3), 409-425.
Sevo, R., & Chubin, D. E. (2010). Bias Literacy: A Review of Concepts in Research. *Women in Engineering, Science and Technology: Education and Career Challenges: Education and Career Challenges*, 21.
Sher, G. (2009). *Who knew?: Responsibility without awareness*: Oxford University Press.
Sherman, N. (1989). The fabric of character: Aristotle's theory of virtue.
Shiffrin, R. M. (1988). Attention. In R. C. Atkinson, R. J. Hernstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (Vol. 2, pp. 739-811). Hoboken: Wiley.
Shih, M. J., Stotzer, R., & Gutiérrez, A. S. (2013). Perspective-taking and empathy: Generalizing the reduction of group bias towards Asian Americans to general outgroups. *Asian American Journal of Psychology, 4*(2), 79.
Sie, M. (2009). Moral agency, conscious control, and deliberative awareness. *Inquiry, 52*(5), 516-531.
Singer, T., Seymour, B., O'doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439*(7075), 466.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin, 119*(1), 3.
Slote, M. (1995). Agent‐based virtue ethics. *Midwest Studies in Philosophy, 20*(1), 83-101.
Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics, 115*(2), 236-271.
Smith, A. M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies, 138*(3), 367-392.
Smith, A. M. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics, 122*(3), 575-589.
Smith, A. M. (unpublished). Implicit Biases, Moral Agency, and Moral Responsibility. *Workshop on Accountability for Attitudes, Harvard University, May 1 2016*, 11-32.
Smith, E. R. (1998). Mental representation and memory. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 4th ed., Vol. 1, pp. 391-445). New York: McGrew-Hill.
Smith, L. B., & Samuelson, L. K. (2003). Different is good: connectionism and dynamic systems theory are complementary emergentist approaches to development. *Developmental Science, 6*(4), 434-439.
Sneddon, A. (2007). A social model of moral dumbfounding: Implications for studying moral reasoning and moral judgment. *Philosophical Psychology, 20*(6), 731-748.
Snow, N. E. (2006). Habitual virtuous actions and automaticity. *Ethical theory and moral practice, 9*(5), 545-561.
Soble, A. (1982). Physical Attractiveness and Unfair Discrimination. *International Journal of Applied Philosophy, 1*(1), 37-64.
Sokolik, M. (1990). Learning without rules: PDP and a resolution of the adult language learning paradox. *TESOL quarterly, 24*(4), 685-696.
Solovay, S. (2000). *Tipping the scales of justice: Fighting weight based discrimination*: Prometheus Books.

Soto, C. J. (2015). Is happiness good for your personality? Concurrent and prospective relations of the big five with subjective well‐being. *Journal of Personality, 83*(1), 45-55.
Spalding, L. R., & Peplau, L. A. (1997). The unfaithful lover: Heterosexuals' perceptions of bisexuals and their relationships. *Psychology of Women Quarterly, 21*(4), 611-625.
Specht, J., Egloff, B., & Schmukle, S. C. (2013). Examining mechanisms of personality maturation: The impact of life satisfaction on the development of the Big Five personality traits. *Social Psychological and Personality Science, 4*(2), 181-189.
Spencer, J. P., Clearfield, M., Corbetta, D., Ulrich, B., Buchanan, P., & Schöner, G. (2006). Moving toward a grand theory of development: In memory of Esther Thelen. *Child Development, 77*(6), 1521-1538.
Spencer, J. P., Dineva, E., & Schöner, G. (2009). Moving toward a Unified Theory While Valuing the Importance of the Initial Conditions. In J. Spencer (Ed.), *Toward a Unified Theory of Development Connectionism and Dynamic System Theory Re-Consider*. Oxford: Oxford University Press.
Sperber, B. (2002). In defense of massive modularity. In I. Dupoux (Ed.), *Language, Brain, and Cognitive Development* (pp. 47-57). Cambridge, M.A.: MIT Press.
Sperber, D., & Mercier, H. (2012). Reasoning as a social competence. *Collective wisdom: Principles and mechanisms*, 368-392.
Sreenivasan, G. (2002). Errors about errors: Virtue theory and trait attribution. *Mind, 111*(441), 47-68.
Ståhl, T., Zaal, M. P., & Skitka, L. J. (2016). Moralized rationality: Relying on logic and evidence in the formation and evaluation of belief can be seen as a moral issue. *PloS one, 11*(11), e0166332.
Stanovich, K. E., & West, R. F. (2000). Individual Differences in Reasoning: Implications for the Rationality Debate? *Behavioural and Brain Sciences, 23*, 645-726.
Steinman, E. (2011). Revisiting the invisibility of (male) bisexuality: Grounding (queer) theory, centering bisexual absences and examining masculinities. *Journal of Bisexuality, 11*(4), 399-411.
Stember, C. H. (1976). *Sexual racism: The emotional barrier to an integrated society*: New York: Harper & Row.
Stokes, D. (2012). Perceiving and desiring: A new look at the cognitive penetrability of experience. *Philosophical Studies, 158*(3), 477-492.
Stokes, D. (2013). Cognitive penetrability of perception. *Philosophy Compass, 8*(7), 646-663.
Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy 48*, 1-25.
Strawson, P. (1993). Freedom and resentment. In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility* (pp. 45-66): Ithaca: Cornell University Press.
Sue, D. W., Bucceri, J., Lin, A. I., Nadal, K. L., & Torino, G. C. (2009). Racial microaggressions and the Asian American experience.
Sunstein, C. R. (2008). Fast, frugal, and (sometimes) wrong. In A. Sinnott, W. (Ed.), *Moral Psychology. Vol. 2: The cognitive science of morality: Intuition and diversity* (pp. 27.30). Cambridge, M.A.: MIT Press.

Sutton Trust. (2013). *Advancing Access and Admissions: The Sutton Trust Summit, November 2013*. Retrieved from https://www.suttontrust.com/wp-content/uploads/2014/11/Universities-Summit-Report.pdf
Tamanna, Y. (2016, Accessed 2017-06-06). 'No Rice, No Curry And No Blacks' - The sexual racism running rampant within the LGBT community. *SBS*.
Taylor, P. C. (2013). *Race: A philosophical introduction*: Polity.
Taywaditep, K. J. (2002). Marginalization among the marginalized: Gay men's anti-effeminacy attitudes. *Journal of Homosexuality, 42*(1), 1-28.
The Higher Education Academy. (2013). *Transition to higher degrees across the UK: an analysis of national, institutional and individual differences*. Retrieved from https://www.heacademy.ac.uk/system/files/transition_to_higher_degree_across_the_uk_0.pdf
Thelen, E. (2005). Dynamic systems theory and the complexity of change. *Psychoanalytic Dialogues, 15*(2), 255-283.
Thelen, E., & Bates, E. (2003). Connectionism and dynamic systems: Are they really different? *Developmental Science, 6*(4), 378-391.
Thelen, E., Kelso, J. S., & Fogel, A. (1987). Self-organizing systems and infant motor development. *Developmental Review, 7*(1), 39-65.
Thelen, E., & Smith, L. B. (1996). *A dynamic systems approach to the development of cognition and action*: MIT press.
Thomas, L. M. (1999). Split-level Equality: Mixing Love and Equality. In S. E. Babbitt & S. Campbell (Eds.), *Racism and Philosophy* (pp. 189-201). Ithaca: Cornell University Press.
Thomas, M. S. C., McClelland, J. L., Richardson, F. M., Schapiro, A. C., & Baughman, F. D. (2009). Dynamic and Connectionist Approaches to Development: Toward a Future of Mutually Beneficial Coevolution. In J. Spencer (Ed.), *Toward a Unified Theory of Development Connectionism and Dynamic System Theory Re-Consider*. Oxford: Oxford University Press.
Thompson, E. P., Roman, R. J., Moskowitz, G. B., Chaiken, S., & Bargh, J. A. (1994). Accuracy motivation attenuates covert priming: The systematic reprocessing of social information. *Journal of personality and social psychology, 66*(3), 474.
Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of personality and social psychology, 100*(6), 1027.
Todd, A. R., & Galinsky, A. D. (2014). Perspective‐Taking as a Strategy for Improving Intergroup Relations: Evidence, Mechanisms, and Qualifications. *Social and Personality Psychology Compass, 8*(7), 374-387.
Tolman, D. L., & Diamond, L. M. (2001). Desegregating sexuality research: Cultural and biological perspectives on gender and desire. *Annual review of sex research, 12*(1), 33-74.
Tovée, M. J., Swami, V., Furnham, A., & Mangalparsad, R. (2006). Changing perceptions of attractiveness as observers are exposed to a different culture. *Evolution and Human behavior, 27*(6), 443-456.
Townsend, J. M., & Levy, G. D. (1990). Effects of potential partners' physical attractiveness and socioeconomic status on sexuality and partner selection. *Archives of Sexual Behavior, 19*(2), 149-164.
Translee. (2016). Lost in the Sauce. On *Album M.A.O.T.P., Pt. 1*. Digital Nativ3 Culture.

Trebilcot, J. (2009). Taking Responsibility for Sexuality. In R. Baker, K. Wininger, & F. Elliston (Eds.), *Philosophy and Sex, 4th ed.* (pp. 337-345). Amherst, New York: Prometheus Books.
Tropp, L. R., & Page-Gould, E. (2014). Intergroup Contact. In J. Dovidio & J. Simpson (Eds.), *APA Handbook of Personality and Social Psychology* (Vol. Vol 2.: Group Processes, pp. 535-560). Washington: American Psychological Association.
Tsoulis-Reay, A. (2016). Are You Straight, Gay, or Just… You? *Glamour*.
Turner, R. N., Hewstone, M., Voci, A., & Vonofakou, C. (2008). A test of the extended intergroup contact hypothesis: The mediating role of intergroup anxiety, perceived ingroup and outgroup norms, and inclusion of the outgroup in the self. *Journal of personality and social psychology, 95*(4), 843.
Tuzin, D. (1995). Discourse, intercourse, and the excluded middle: Anthropology and the problem of sexual experience. In P. R. A. S. D. Pinkerton (Ed.), *Sexual nature, sexual culture* (pp. 257-275). University of Chicago Press, Chicago.
Tyrrell, J., Jones, S. E., Beaumont, R., Astley, C. M., Lovell, R., Yaghootkar, H., . . . Hirschhorn, J. N. (2016). Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *bmj, 352*, i582.
Vare, J. W., & Norton, T. L. (1998). Understanding gay and lesbian youth: Sticks, stones, and silence. *The Clearing House, 71*(6), 327-331.
Vargas, M. (2005). The trouble with tracing. *Midwest Studies in Philosophy, 29*(1), 269-291.
Vargas, M. (2006). On the importance of history for responsible agency. *Philosophical Studies, 127*(3), 351-382.
Vargas, M. (2013). *Building better beings: A theory of moral responsibility*: OUP Oxford.
Velleman, D. (2000). *The Possibility of Practical Reason*. New York: Oxford University Press.
Voci, A., & Hewstone, M. (2003). Intergroup contact and prejudice toward immigrants in Italy: The mediational role of anxiety and the moderational role of group salience. *Group Processes & Intergroup Relations, 6*(1), 37-54.
Vohs, K. D., Baumeister, R. F., & Loewenstein, G. (2007). *Do Emotions Help or Hurt Decisionmaking?: A Hedgefoxian Perspective*: Russell Sage Foundation.
Walster, E., & Berscheid, E. (1974). A little bit about love. *Foundations of interpersonal attraction*, 355-381.
Watson, G. (1987). Responsibility and the Limits of Evil: Variations on a Strawsonian theme. In F. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 256-286). NY: Cambridge University Press.
Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*(2), 227-248.
Watts, J., & Robertson, N. (2011). Burnout in university teaching staff: a systematic literature review. *Educational Research, 53*(1), 33-50.
Watts, L. (2012). Gay Men and Women are Not More Racist. *The Huffington Post*.
Wegner, D. (2002). The illusion of conscious will. Cambridge. MA: The MIT Press.
Weinberg, M. S., Williams, C. J., & Pryor, D. W. (1995). *Dual attraction: Understanding bisexuality*: Oxford University Press.
West, R. (2017). Virtue Ethics is Empirically Adequate: A Defense of the Caps Response to Situationism. *Pacific Philosophical Quarterly*.
Westfall, R., Millar, M., & Walsh, M. (2016). Effects of instructor attractiveness on learning. *The Journal of general psychology, 143*(3), 161-171.

Wetzel, C. G., Wilson, T. D., & Kort, J. (1981). The halo effect revisited: Forewarned is not forearmed. *Journal of experimental social psychology, 17*(4), 427-439.
Whaley, A. L., & Geller, P. A. (2007). Toward a cognitive process model of ethnic/racial biases in clinical judgment. *Review of General Psychology, 11*(1), 75.
Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological science, 16*(10), 780-784.
White, G. L., Fishbein, S., & Rutsein, J. (1981). Passionate love and the misattribution of arousal. *Journal of personality and social psychology, 41*(1), 56.
Willard, L. D. (1977). Aesthetic discrimination against persons. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie, 16*(4), 676-692.
Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin, 116*(1), 117.
Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of personality and social psychology, 60*(2), 181.
Wolf, N. (2013). *The beauty myth: How images of beauty are used against women*: Random House.
Wolf, S. (1980). Asymmetrical freedom. *The Journal of Philosophy, 77*(3), 151-166.
Wolf, S. (1987). Sanity and the Metaphysics of Responsibility. In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 46-62): Cambridge University Press.
Wolf, S. (1990). *Freedom within reason*: Oxford University Press.
Wood, W., Quinn, J. M., & Kashy, D. A. (2002). Habits in everyday life: Thought, emotion, and action. *Journal of personality and social psychology, 83*(6), 1281.
Wright, S. C., Aron, A., McLaughlin-Volpe, T., & Ropp, S. A. (1997). The extended contact effect: Knowledge of cross-group friendships and prejudice. *Journal of personality and social psychology, 73*(1), 73.
Wuensch, K. L., Chia, R. C., Castellow, W. A., Chuang, C.-J., & Cheng, B.-S. (1993). Effects of physical attractiveness, sex, and type of crime on mock juror decisions: A replication with Chinese students. *Journal of cross-cultural psychology, 24*(4), 414-427.
Zheng, R. (2016). Why Yellow Fever Isn't Flattering: A Case Against Racial Fetishes. *Journal of the American Philosophical Association, 2*(3), 400-419.