Ontological Critique:

A Philosophical Tool for Advancing Social Psychology

Matthew Lloyd Jenkins

2021

Submitted 3 years, 10 months after commencement

Thesis submitted for the partial fulfilment of the requirement for the degree of

Doctorate of Philosophy

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

South, West & Wales Doctoral Training Partnership

Contents:

Chapter 4

Chapter 5

Chapter 6

List of Figures:

Acknowledgements

I have been unbelievably fortunate throughout my studies to be surrounded by friends and colleagues who have invested their time, effort, and patience in me and my work. Many long conversations with Dr. Liz Irvine, Professor Nick Shackel, Dr. Orestis Palermos, and Professor Alessandra Tanesini have been invaluable for the formation and development of my ideas to the point presented here. Special thanks are due to Professor Jonathan Webber as my lead supervisor who has spent countless hours painstakingly going over my work and being willing to take the time to discuss, with good humour, every tangent and rabbit hole I have explored these last few years.

Continuing this thesis through the past few months of lockdown would have been impossible without the support and friendship of Ollie Florence and Flora McNerney. Their fantastic good humour and words of encouragement, as well as regularly asking when I would be submitting, pushed me to finish when things got hard. The company and camaraderie of Elisa Tabo and Dr. Rossella di Bernardi has also brightened many otherwise dark evenings and kept this project in perspective.

All my parents and grandparents have encouraged and pushed me to pursue a project I was excited for, and knew I would enjoy, and without their support and enthusiasm this thesis would have been far harder to finish. My siblings have tolerated my idiosyncrasies and provided much needed distraction these past few months for which I am forever grateful.

Above all, this project is dedicated to a woman it would not be finished without. Hope has been an inspiration and an impetus to me. Her dynamism, dedication, joy, and fearlessness have been invaluable to me and give me faith that together we can tackle all the challenges to come.

Thesis Abstract:

Replication failures indicate that we have reason to doubt the generalisations made from the original experiment to the original conclusion. This thesis argues that the proposed statistical responses to the replication crisis in social psychology are necessary for long-term progress in the research programme but remain insufficient without a supplementary approach which targets the theoretical frameworks which justify the conclusions drawn.

Beginning by offering an account of the replication crisis in social psychology, the thesis then turns to addressing some extant approaches to parsing the literature into more and less reliable. In contrast to these approaches, the ontological critique is proposed as an approach which explicitly addresses the problems in social psychological theory. This ontological critique is a tool for unpacking the theoretical justifications which underpin the generalisations made when constructing conclusions from data. Proposed desiderata approaches to model choice for implicit cognition are addressed in chapter 3 and presented with problem cases from within the extant literature. These cases give us reason to prefer the ontological critique if it can be shown to be fruitful. The subsequent three chapters are studies in the application of the critique to the trait picture of attitudes, the mainstream model of attitudes, and the Cognitive-Affective Personality System as a model of attitudes, respectively, with the aim of demonstrating this fruitfulness. Each introduces the relevant theoretical framework before applying the ontological critique, illustrating both the directions of clarification and refinement for each account of attitudes and the utility of the critique. The final chapter draws together the conclusions reached throughout the thesis and relates this project to other approaches to the crisis in psychology.

Section 0.0: Introduction:

Psychology has been through a crisis over the past decade, though its roots lie far deeper. Many of the field′s core discoveries have failed to replicate in high-powered, many-lab studies. Understanding why this is the case, what this means, and how to change this in the future has become a core problem for the field. A wide variety of approaches have been proposed, from improved statistical practices to changing the training of psychologists. While psychology has not been unique in facing this problem, some of the challenges it faces are very different to other fields facing replication failures.

This thesis begins by framing the problem of the replication crisis as an epistemic problem for readers of social psychology. While many of the proposals for changes in the field have merit and some combination of them seems essential to the improvement of research in social psychology and its careful interpretation by readers, they are also argued to be systematically inadequate to respond to the problem. The resampling account of replications highlights the importance of the theoretical framework within which we situate our empirical conclusions for the robustness of those conclusions. This theoretical framework is inadequately developed to support many of the generalisations which are made from data to phenomenon.

The ontological critique consists in a series of challenges designed to interrogate the theory surrounding explanations and predictions offered within a research programme. Rather than presenting a formal evaluative tool which offers a clear diagnosis of the theory in question, the ontological critique offers a toolkit which helps unpack the theoretical mechanics of the explanations and predictions offered by a theory. The affordance of doing so lies in its invitation of precise, constructive criticisms of the existing theoretical frameworks in order that they may be gradually improved by being explicitly addressed. This frames the ontological critique as a tool to refine and parse the ongoing dialogue on theory in psychology. This makes it a tool for an ongoing project in social epistemology rather than a straightforward evaluation of the target theory; intended to offer clarity and communicate transparently about the relationships between empirical claims and their theoretical justifications.

This clarity enables more precise judgements about the robustness of explanations or predictions. This has consequences for both readers and practitioners. Readers are equipped with a systematic approach to identifying the theoretical justification for the explanations they are offered which enables judgements about the relative robustness of the conclusions presented in publications. Practitioners are offered a tool which identifies areas of theory that offer relatively weak support for the stated goals and highlights the existing project-internal motivations for targeted improvement in these areas.

The ontological critique also offers a systematic approach to clarifying the generalisations made in constructing an explanation and the theoretical warrant for those generalisations for a given population. In conjunction with a resampling account of replications, the critique is a tool which fills a specific need in the literature responding to the replication crisis in psychology.

The thesis further offers critical readings of three key models of attitudes: the trait picture, the mainstream model and the cognitive-affective personality system. These readings present clear problems and directions for fruitful improvement to each model and demonstrates the efficacy of the ontological critique as a tool for advancing theory in social psychology.

Section 0.1: Structure:

Chapter 1 introduces the replication crisis and presents the resampling account of replications. On this account a replication resamples one of the experimental components of the original experiment from the population to which the original conclusion was, or could be, generalised. The problem of a given replication failure is therefore a problem of the unreliability of a given conclusion − one of the components involved in the construction of that conclusion is unreliable. By adopting a Lakatosian approach in the philosophy of science, this problem for a given finding is generalised to a problem of a research programme whose findings tend to fail to replicate. The chapter concludes with the in-credibility problem: the publication of a conclusion in the social psychology literature is not good enough reason to rely on that conclusion.

Chapter 2 explores how we might address this problem of in-credibility. Set against the backdrop of the plausibility of rejecting social psychology altogether, and the considerable social and epistemic cost of such an approach, the chapter begins by focussing on statistical heuristics for ensuring reliability. Selecting for $p<0.005$ and calculating PPV/prior curves are considered as heuristics for readers to exclude unreliable conclusions. The chapter goes on to highlight that these approaches and others like them, while necessary to ensure the reliability of the underlying statistical inferences, do not ensure the overall reliability of the conclusion. The reliability of the conclusion further depends on the theoretical framework within which the conclusion is situated and from which its explanations and predictions are resourced. The chapter closes by outlining the challenges of the ontological critique and guides for robust answers to each challenge. The five challenges are the axiology challenge, the objects challenge, the properties challenge, the explanation challenge, and the prediction challenge.

Chapter 3 introduces the desiderata approach to model evaluation, focussing on desiderata for a model of implicit cognition. The desiderata approach is explored as a means of identifying the explanatory and predictive power of a model and is presented as an alternative approach to evaluating and unpacking social psychological models to the ontological critique. Some problem cases for the desiderata approach to model evaluation are raised and discussed: the good case of comparing the mainstream model with the CAPS model and the bad case of the trait picture of attitudes. The desiderata approach is shown to face fundamental challenges when evaluating these cases, which are already present in the literature. As such, insofar as another approach can resolve such cases it is to be preferred. The subsequent three chapters aim to demonstrate that the ontological critique is such an approach by applying the critique to the identified problem cases.

Chapter 4 introduces the Freudian picture of attitudes and Machery′s argument against the adoption of that picture. By addressing each argument against the picture in turn only one of these arguments is shown to be unanswerable by the Freudian picture. This analysis clarifies that what is at stake in this argument is not the Freudian picture as a whole, but rather a subset of its theoretical commitments which continues to have widespread acceptance in both attitude psychology itself and the philosophical literature on attitudes. By framing the negative programme in this way, Machery′s positive programme, arguing for the trait picture, is recast and the theoretical advantage of the approach is clarified as avoiding committing to the problematic elements of the Freudian picture. By applying the ontological critique to the trait picture, criticisms of the trait picture are made more precise and avenues for future research are indicated. This illustrates both the advantage of the trait picture over some of the available alternatives and the advantage of the ontological critique over a desiderata approach to model choice.

Chapter 5 presents the mainstream model of attitudes employed in attitude psychology. By presenting some *prima facie* challenges to the model, the necessity of situating this mainstream model within an overarching MODE model is demonstrated in order that the mainstream model can offer explanations of its target phenomena. This composite MODE-mainstream model is then analysed using the ontological critique, drawing out several key strengths and weaknesses of the model and highlighting fruitful directions for further theorising and experimental research. This highlights several affordances of the critique, including its advantage over the enumeration of features and properties of the model.

Chapter 6 presents the cognitive-affective personality system (CAPS) model of attitudes. Breaking down the structure of the model in its own terms highlights its plausibility as an explanation of its key phenomena. By applying the ontological critique to the model this *prima facie* plausibility is shown to be glossing over key questions about the grounding of some of the model's core claims. Exploring candidates for providing these grounds, dynamical systems theory and connectionism, occupies the second part of the chapter. This exploration highlights that there is no clear candidate for grounding the claims of the model that simultaneously preserves the various functions. Given this challenge, some fruitful directions for further research are identified. This demonstrates the utility of the critique in unpacking the content of a model, beyond simply presenting the model on its own terms.

Chapter 7 draws together many of the findings of the thesis toward the overarching conclusions. First, that psychology generally and social psychology in particular faces a serious challenge in the replication crisis which will require an improvement in its underlying theoretical framework to adequately address. Second, that we have good theoretical reason to think the ontological critique is a tool which will offer some traction in addressing this problem. Third, that the practical implementations of the critique have shown it to be both useful as a critical reading tool and fruitful as a way of offering constructive criticism to the theoretical framework surrounding our explanations. The chapter briefly introduces several other ongoing projects in the literature and situates the ontological critique in relation to each. The chapter also raises the concern that there is currently inadequate theoretical framework to design or interpret replications in psychology. This concern is responded to through a combination of reframing through the resampling account of replication and by highlighting the utility of the ontological critique given this new framing. Two critical theoretical problems for the thesis are raised; the objects/properties problem and the Lakatos problem. Finally, an outline of the direction in which future research may hope to answer these key problems is presented.

Chapter 1: A philosopher′s guide to the replication crisis

Section 1.0: Introduction

This chapter begins by situating the current problem of the replication crisis in social psychology within a brief history of the response to a single, key paper. Bem (2011), and the response to the shocking finding therein, is used to frame the problem of the replication crisis as a problem of in-credibility. That is, as an epistemic problem for readers of social psychology from outside the discipline, especially researchers in other fields looking to utilise social psychological findings as premises in their own arguments. This problem is framed as an argument for the conclusion that social psychology publications, reliant on the dominant method in social psychology, do not give us novel reasons to believe their claims.

With the historical situation of the problem, as well as the form of the argument arranged, sections 1.2 and 1.3 address two avenues of responses to this argument. Section 1.2 addresses some responses which suggest that Bem′s method is, in a relevant sense, anomalous and not an indicator of the reliability of the method in social psychology more broadly. Section 1. 3 addresses responses that argue for the plausibility of reliable methods nevertheless producing anomalous results without thereby being unreliable. Having raised and responded to these claims, section 1.4 highlights that while the problem is well framed as originating in Bem historically, it may be presented in a stronger, briefer form. Section 1.5 supplements this argument with a Lakatosian defence of the treatment of social psychology as a research programme possessing properties such as reliability. Section 1.6 introduces the resampling account of replications (Machery, 2017), using this account to inform us about the challenge presented to an original conclusion by a replication failure. Section 1.7 addresses the Open Science Collaboration (2015) and responses to the findings therein from Gilbert et al. (2016). By doing so the challenge to the reliability of the method of social psychology is clarified and presented. Section 1.8 draws together the previous sections summarising the consequences of replication failures for social psychology as a research programme. Section 1.9 outlines the conclusion as well as some avenues for responses, identifying literature which explores some of these areas and highlighting those which will be further addressed in chapter 2.

Section 1.1: A (very) brief history

While there are several plausible histories that may be offered of the replication crisis, I will here outline a brief version that centres around the controversy of a single paper, and how the field responded or failed to respond. For a thorough history of how we arrived where we now are, we will likely have to wait for significantly more of the dust to settle. For the time being, the following history will suffice to introduce the central problems of the crisis, as well as the problems′ relevance for psychologists and those who read, or rely on, psychology publications.

Bem (2011) shocked psychologists, and readers in other fields. Across nine different experimental conditions, Bem found evidence for ′precognition′, specifically ″the anomalous retroactive influence of some future event on an individual's current responses″ (2011, p. 407). To be clear, this is the claim that current responses can have future causes. That, for some people, causation sometimes runs backwards. This result was found across nine different experiments, some of which reported particularly high levels of significance[1].

Schimmack (2020) summarises the response among psychologists: ″Psychologists were confronted with a choice. Either they had to believe in anomalous effects or they had to believe that psychology was an anomalous science.″ That is, given the results that Bem had generated and given the methodological similarity with many other findings across psychology, it seems that we either have reason to believe Bem′s claims on the basis of the reliability of the method used to establish them, or we take the conclusion to be a *reductio* of the method used to establish the conclusion and, thereby,

[1] Experiment 2: p=0.009; Experiment 3: p=0.007; Experiment 9: p=0.002. (Bem, 2011, p. 421)

lose our reason to believe many other findings across social psychology; i.e. that they were established via a reliable method.

i. Bem's results are absurd.

ii. Bem's results are generated by method M_B, definitionally.

iii. Method M_B is identical in all the relevant particulars with method M_P, which is widely used in social psychology.

iv. Social psychology publications which use M_p give us a novel reason to believe their claims if and only if method M_P is reliable.

v. Reliable methods do not generate absurd results.

vi. Method M_P is not reliable.

vii. Social psychology publications that use M_P do not give us a novel reason to believe their claims.

Before I move on to discuss the premises of this argument, allow me to first clarify the conclusion. This conclusion is not identical with the claim that we should not believe the conclusion of a given social psychology publication. Rather, if the argument holds, we may well reasonably believe the conclusion of a given publication, but the publication itself cannot be the reasonable basis of that belief. If a given publication finds that persons displaying race-bias in their explicit responses in an interview are more likely to display similar or related tendencies about sex on a questionnaire, we may well regard this as plausible given our personal engagement with such people in the past, or based upon our pre-existing notions about how interconnected beliefs or habits likely operate. However, if the argument holds, we have no more reason to believe the conclusion of the publication after reading it than we had before.

With the manifest importance, for social psychologists and those reading their publications, of rejecting the conclusion established, how we may we reasonably reject the conclusion?

To begin, I take it that the claim 'some people are capable of reliable precognition' is indeed absurd. Some few 'psi' researchers have defended the view that it is not, but that will not be entertained in this thesis. Similarly, since M_B is, definitionally, whatever method Bem in fact employed to gather his data, the second premise is trivially true. This leaves the premises iii-vi open to challenge.

Section 1.2: Is Bem′s method anomalous?

I shall discuss three lines of response to premise iii − the first is the attempt to replicate the results directly, the second is the claim that Bem′s statistical approach was insufficient to establish his conclusion and that 'proper′ Bayes-factor analysis demonstrates this failing, and the third is that Bem′s results were the product of ′questionable research practices′.

One immediate, positive response among psychologists was an attempt to directly replicate Bem′s results. Several replication attempts were made, by several different groups of psychologists. These replications were then brought together in the meta-analysis conducted by Galak, LeBoeuf, Nelson, and Simmons (2012). Overall, the effect size across the replications not conducted by Bem was 0.04, considerably smaller than Bem′s average effect of 0.29 and not statistically different from zero (95% CI: 0.00, 0.09). The full dataset including Bem′s original results (n = 4091) was then coded on six dimensions: ″(a) whether the experiment attempted to replicate Bem′s Experiment 8 or his Experiment 9, (b) whether it was administered online or in a lab, (c) whether it was conducted by Bem, (d) whether the software used to administer the experiment was the software originally used by Bem, (e) whether the results had already been published··· and (f) whether the experimenters conducting the replication expected to observe a psi effect.″ (Galak, LeBoeuf, Nelson, and Simmons, 2012, p. 938). Of these six dimensions, the only statistically significant factor was whether the research was conducted by Bem.

With this weight of evidence, we can confidently say that Bem′s results are anomalous and that they do not give us novel reason to believe in psi. However, it is not immediately clear what we may say regarding the third premise. There appears to be something about M_B (Bem′s exact method) that produced these results that was not available to the replicators as part of the reported method. This may be because the determining factor was innocuous and so was not reported, or it may be because it was illicit, and so was not reported.[2]

While it might be tempting to view this replication failure as a rebuttal of the third premise, closer inspection reveals that this is not the case. The failure to replicate, while it may illustrate the presence of questionable research practices in M_B, does not demonstrate that M_B and M_P are substantively different[3]. As a result, while the attempt to replicate Bem′s results is unquestionably admirable, it does not, by itself, dissolve the argument that psychology publications offer us no novel reason to believe their claims.

The second response to Bem′s results was to challenge the statistical analysis Bem employed. This response was made by Wagenmakers et al. (2011) who argue that Bem′s results only affirm his overall conclusion if the analysis is done with a one-directional null-hypothesis significance test (NHST). Wagenmakers et al. (2011) argue that NHST overstates the statistical evidence against the null hypothesis (that there is no psi-effect) because it only considers one direction of analysis: what is the probability of deriving these results given the truth of the null hypothesis? That is, given that there is no psi-effect, what are the odds of generating Bem′s results? Bem (2011) reported a one-sided p-value <0.05, that is, the odds of generating those results, given that there is no psi-effect is less than one in twenty. However, when we consider a one-sided test we do not consider the probability relative to the probability of the converse – i.e. what was the probability of generating those results if the null hypothesis was false, and how does that probability

[2] Two further possible explanations: that the results are a statistical fluke but not illicit, or that the result is the outcome of a ′file drawer problem′ will be discussed with regards to the rejection of premise v.

[3] Indeed, if one knew the replication rate obtained in the Open Science Collaboration (2015) for social psychology and knew nothing more about Bem (2011) than that it was a social psychology paper, its failure to replicate is exactly what we would expect.

relate to the probability of generating this data given that the null hypothesis is true?

As an example, consider the lottery. I have won the lottery. Let us assume that the odds of me winning the lottery at random are 1 in 1,000,000, p-value = 0.000001. The hypothesis we want to test is that I have cheated to win the lottery. Given a NHST analysis, since it is so unlikely that I have won the lottery at random, we ought to reject the null hypothesis and conclude that I cheated to win the lottery. What Bayesian statistics allows is for us to test the probability of my winning at random with the probability that I won by cheating.

Let us assume that the probability of cheating to win the lottery requires eight independent chances to fall into place, each at a rate of one in ten. This means the likelihood of successfully cheating to win the lottery is 1 in 100,000,000. A Bayesian analysis[4] would divide the probability of obtaining our result given the falsehood of the hypothesis by the probability of obtaining our result given the truth of the null hypothesis [$P(Data|H_1)/P(Data|H_0)$]. In this example, this would give us a Bayes Factor of 100. Following Jeffries (1961) this result would give us extreme evidence for H_0. Rather than rejecting the null hypothesis as the NHST recommended, we must instead reject the positive hypothesis and accept the null hypothesis – with overwhelming evidence. We are thus justified in believing that lottery winners win by chance, despite the odds against them doing so, rather than believing that they cheated.

Based on this argument, Wagenmakers et al. (2011) conclude that Bem′s results are a similar case, presenting their own analysis of his data under a Bayesian hypothesis test. Under their analysis, Bem′s data becomes significantly less persuasive. Many of the experiments go from showing weak evidence for the positive hypothesis to providing evidence for the null hypothesis (Wagenmakers et al., 2011, p. 430). The evidence from experiment 8 in particular, previously significant in favour of psi, now

[4] Wagenmakers et al.′s argument illustrates the phenomenon well but oversimplifies the alternative and shouldn′t be taken to illustrate a good Bayesian analysis of the phenomenon.

presents anecdotal evidence in favour of the null (Bayes-Factor = 2.11). They then conclude that "the evidence for precognition is either non-existent or negligible" (Wagenmakers et al., 2011, p. 430). If Wagenmakers et al. are correct, then we may conclude that the data obtained by Bem is neither absurd nor anomalous, simply misinterpreted – and therefore that the reliability of the method, excluding the statistical interpretation of the data, is not challenged.

Unfortunately, Wagenmakers et al. do not seem to be correct. As identified by Galak, LeBoeuf, Nelson and Simmons (2012) the important result is that of Bem's experiment 9. Under Wagenmakers et al.'s own Bayesian hypothesis test, this experiment provides "substantial" evidence for the existence of psi. Indeed, under their Bayesian analysis, experiment 9 offers some of the strongest evidence in either direction (Wagenmakers et al., 2011, p. 430, Table 2).

The problem of Wagenmakers et al.'s interpretation is compounded by the fact that Bayes-factor analysis comes with its own set of problems – especially the choice of a prior probability distribution (Schimmack, 2020; Simmons, Nelson, and Simmonsohn, 2011). Bem, Utts, and Johnson (2011) respond to Wagenmakers et al. (2011) by highlighting that the prior distribution adopted by the latter assumes a 50% chance of psi working in the opposite direction (the anomalous effect of current cognitions on future events), and, more importantly, that only 25% of the prior distribution was allocated to effect sizes between 0 and 1. Given that the hypothesis of psi, if true, is only committed to very small effects (or psi would not be a controversial hypothesis), this prior is entirely unsuitable for the analysis. Bem, Utts and Johnson (2011) recommend and apply a more reasonable prior to Bem's original data, that reflects the small effect sizes that psi researchers would expect of a psi hypothesis. This Bayesian analysis shows more evidence for psi, particularly in experiment 9, than Bem's original NHST analysis. By building a prior distribution based upon the data on effect sizes from similar experiments in social psychological research, and applying this new prior to the data, Bem Utts and Johnson (2011) find that the new Bayes-Factor is 0.099, rather than the 0.17 calculated by Wagenmakers et al. (2011). This means that the evidence from experiment 9 changes categorisation from substantial evidence for psi to strong evidence for psi [5] (Jeffries, 1961).

Wagenmakers et al.′s response, therefore, fails on two counts. Insofar as Bem′s conclusions are indeed results derived from the application of bad statistical practice (NHST), that statistical practice remains widespread, indeed the dominant mode of statistical analysis in the literature, something that Wagenmakers et al. highlight. That is, at best, we are offered a diagnosis of what is wrong with M_P, rather than what the substantive difference between M_B and M_P might be. On the second count, it is not clear that we may simply apply a Bayesian hypothesis test to the relevant dataset in order to generate a reliable result since the different priors offer different diagnoses of Bem′s data, and the more plausible prior gives us an even stronger case for psi than Bem originally presented. This problem for the application of Bayesian hypothesis testing is in line with suggestions made by Simmons, Nelson, and Simmonsohn (2011) that Bayesian statistics presents more ′degrees of researcher freedom′ [6] which, in simulation testing, increases the probability of generating false positive results. [7]

The third response we shall consider is that Bem′s method incorporates questionable research practices (QRP). QRP is an umbrella term for those

[5] While BF=0.099 is only just within Jeffries (1961) strong evidence for the test hypothesis category, the change from 0.17 to 0.099 remains substantial and noteworthy.

[6] While Simmons, Nelson and Simonsohn (2011) recommend against the implementation of Bayesian statistical testing in psychology, on this basis, the justification also seems *prima facie* valid against the frequentist statistical tools already employed within psychology. As such, while I concur that simply applying Bayesian tools to the existing problems in psychology is inadequate, we ought not to thereby conclude that attending to Bayesian concerns and criticisms is unnecessary or unhelpful. This point is returned to and employed in chapter two, regarding both p<0.005 and to the implementation of PPV/Prior curves.

[7] Further to the references discussed, for the case for Bayesian hypothesis testing as well as warnings about its implementation see (Efron and Tibshirani, 1986), for interpreting Bayes-factors see (Jeffries, 1961), and for criticisms of the use of Bayesian statistics in social psychology see (Simmons, Nelson, and Simmonsohn, 2011).

practices that, while not outright fraudulent, undermine the integrity of the results by increasing the potential for false-positives. The term was introduced by John, Loewenstein, and Prelec (2012) who identified and estimated the prevalence of nine kinds of QRP as well as fabrication of data. What differentiates QRP from fraud is that, generally speaking, fraud is clearly not tolerated within scientific publishing, while, generally speaking, QRP are (or perhaps, were) tolerated, at least to some extent. As part of their estimation of the prevalence of these practices, John, Loewenstein, and Prelec (2012) asked their participants (psychology researchers, n=2155) to rate each category of QRP in terms of its defensibility. These perceived defensibility ratings ranged from 0 (no, not defensible), 1 (possibly defensible), to 2 (yes, defensible). Fabrication of data (obviously fraudulent) was rated as 0.16 (SD = 0.38). By contrast, the remaining QRP were rated between 1.84 (failing to report all of a study′s dependent measures) to 1.32 (claiming results are unaffected by demographic variables when one is either unsure or knows that they are) (2012, p. 525). This shows that, at least in 2012, psychology researchers, on the whole , thought that many QRP were relatively defensible. To clarify, these are practices that systematically increase the potential for the presentation of a false positive result, without increasing the chances of a true positive result. Simmons, Nelson and Simmonsohn (2011) applying a combination of high defensibility rating QRPs to simulated data were able to generate a false positive rate of 61%, that is, 61% of the time an obviously false result would be presented with $p < 0.05$ [8]. Following a statistical analysis of Bem′s (2011) data, Schimmack (2012) concludes that the application of the ′Incredibility index′ makes it

[8] While Simmons, Nelson, and Simmonsohn (2011) conduct their analysis entirely in terms of frequentist statistics and P-values, Simmonsohn (2014) shows that the effect on Bayes-factor analysis of QRP is just as severe.

Similarly to the replication failure, and the application of Bayes-factors, this diagnosis of Bem offers us a way of identifying what it is about M $_B$ that led us to a false conclusion, however it is not a rejection of the relevant similarity of M_B to M_P, since what makes QRP so problematic is the fact that they are so widespread in psychology (and other disciplines) (John et al., 2012). As a result, none of the three responses to Bem in fact undermine premise *iii*, indeed they seem to offer support for *iii*, as well as for *vi*.

Section 1.3: Reject *v*?

One plausible response to the argument as presented above is that it is simply not the case that reliable methods do not guarantee that we will not generate absurd results. If this is the case, then we may reject *v*. In one sense, it is normal that some subset of studies will not replicate, or will generate a false, potentially absurd, result. If psychologists were to conduct studies with 80% statistical power[9], one in five studies would fail to replicate with the same procedure and sample size, even if everything goes well and all predictions were true (Schimmack, 2020). The replication failure rate is increased in cases where psychologists are testing risky hypotheses – where H_1 has a high probability of being false. Such replication failures are then evidence of a false positive result in the original experiment.

Bem (2011) is undoubtably testing a risky hypothesis, a hypothesis that is so unlikely to be true that an experiment that finds that it is true is more likely to be taken to be a *reductio* of the method than to be a true positive. However, this explanation, that such absurd results can be generated without undermining the method by which we arrive at those results, overlooks a key problem of the rate of those results.

It is one thing to say that it remains possible to generate absurd results. It is another to say that the phenomenon we are observing in psychology, or specifically in Bem (2011), is such a case. If each of Bem's experiments has the

[9] Assuming our estimated power is an accurate reflection of the 'true' power.

purported statistical power[10], then the odds of obtaining statistically significant results, nine times, in nine ostensibly independent experiments, when the null hypothesis is true are vanishingly small [11]. So small, in fact, that not enough peer-reviewed psychology papers have ever been published that we would expect such a thing to happen by chance.

This brings us to another form of the rejection of v, the file drawer problem. Rosenthal (1979) introduced the concept of a file drawer which describes the phenomenon that results from publication bias. Publication bias is the phenomenon where publishers select predominantly, or only, positive results to publish. Concerns have been raised about this practice for decades (Sterling, 1959; Rosenthal, 1979). One result of this practice is that psychologists would run multiple studies and only create a hypothesis after seeing the results so as to generate a significant, positive result (dubbed HARKing by Kerr, 1998), or would run many studies and only publish those results that arrived at a significant, positive result (estimated to have been done by 48% of research psychologists by John et al., 2012). These practices result in a set of studies that are positive and significant and thereby published, and a second set that were either significant (on Bayes-factor, not p-values) and negative, or simply insignificant and inconclusive. This second set was never published and entered into psychology′s collective, proverbial ′file drawer′, hence the name of the problem.

To illustrate the problem that the file drawer generates, over a set of studies on different hypotheses, where some studies are true positives and others are false positives, we may imagine a set of experiments where 80 experiments are run on false hypotheses and 20 studies are run on true hypotheses with 50% power[12]. In this case, we expect 14 significant, published results[13].

The advertised false positive rate of 5% ($p < 0.05$) is true for the 100 studies that were conducted, but it would be false to believe that only 5% of

[10] Such that our experiments are reliably sensitive to the smallest effect size of interest.

[11] Prod.(p-value{exp. 1, 2⋯ 9}) = 0.00000000000000007, or 1 in 140,000,000,000,000,000.

[12] Here ‘true’ power is assumed for simplicity.

[13] 80*0.05=4; 20*0.5=10; 4+10=14

the selected set of 14 studies with significant results could be false positives. In this example, we would falsely assume that at most 1 of the 14 studies is a false positive; 14*0.05 = 0.7 studies. However, in this case, we know that there are in fact 4 false positive results. We do get the correct estimate of the maximum number of false positives, if we start with the actual number of studies that were conducted, which gives a false positive risk of 5 studies, which would be a percentage of 5/14 = 36%. Thus, up to 36% of the reported 14 studies could be false positives and the actual risk is 7 times larger than the claim $p<0.05$ suggests. In short, we need to know the size of the file-drawer to estimate the percentage of reported results that are likely to be false positives.

This file-drawer problem is therefore intended to explain the increased rate of false-positives in the literature without having to posit a problem with the method M_P. It offers an explanation of the current situation in psychology in terms of a plausible bias in publication selection rather than a systemic problem with method. However, the scope of the file-drawer problem to explain the prevalence of false-positives is dependent on the analysis being of a single experiment at $p<0.05$. If psychology were populated with studies such as Bem′s that run a large number of independent experiments to test a single hypothesis, the size of the file-drawer problem shrinks rapidly. For a set of five independent studies, the probability of a false positive shrinks from $p<0.05$ for a single study to $p<0.0000003$ (i.e., 0.05^5). ″This is approximately the same stringent criterion that is being used in particle physics to claim a true discovery″ (Schimmack, 2012, p. 552). The file drawer problem for such a set of studies decreases in proportion to the stringency of these tests in weeding out false positives.

As a result, while the file drawer certainly plays a role in increasing the false-positive rates in psychology, it is insufficient to the explanation of the presence of Bem in the literature – i.e. Bem's requirement to achieve p<0.05 on nine independent studies, on paper, is so stringent that we would expect a false positive on all nine studies to occur only once in 512 billion studies given the falsehood of the null and the independence of the experiments . The file-drawer effect in publishing does not explain how M_B might be consistent with psychology's method being generally reliable. As such it does not provide an explanation for consistency between M_P and absurd results. Bem's results are absurd, and should not occur under M_P, even with an absurdly large file-drawer effect.

This is not to say that the file-drawer effect is not real and demonstrably problematic – rather, the file-drawer effect certainly occurs and is certainly problematic. But where psychologists are running multiple independent experiments in a single study what we are seeing is not the effect of the file drawer but rather a combination of questionable research practices and "poor statistical training at both undergraduate and graduate levels" (Schimmack, 2012, p. 561).

Section 1.4: Replication Failures

Eagle-eyed readers of the original argument will have identified that the above argument, while well illustrated with Bem, is actually unnecessary to establish the conclusion. We may instead parse down the original argument to the following:

> i. Psychology publications which use M_P give us a novel reason to believe their claims if, and only if, method M_P is reliable.
>
> ii. Method M_P is not reliable.
>
> iii. Psychology publications that use M_P do not give us a novel reason to believe their claims.

For the sake of maintaining clear scope, this chapter assumes the truth of the first premise.

Following Bem′s publication, and the subsequent discussion surrounding the challenge that it presented, the Nobel Laureate Daniel Kahnemann wrote an open letter to social psychologists, especially those who work in evaluative priming. Despite his relatively positive attitude towards priming results, Kahnemann writes:

> ″I see a train wreck looming⋯ I believe that you should collectively do something about this mess. To deal effectively with the doubts you should acknowledge their existence and confront them straight on, because a posture of defiant denial is self-defeating... organize an effort to examine the replicability of priming results, following a protocol that avoids the questions that have been raised and guarantees credibility among colleagues outside the field.″ (Kahneman, 2012)

Following the challenge that Bem′s paper, among others, presented to the credibility of the field, Kahneman called for replications.

Following Kahneman′s suggestion, a significant number of psychologists[14] took up the challenge and in 2015 the Open Science Collaboration (OSC, 2015) was published. The OSC focused on replicating results published in three psychology journals in 2008. These journals were the *Journal of Personality and Social Psychology* , the *Journal of Experimental Psychology: Learning, Memory and Cognition* , and *Psychological Science*.

In total, 97 replications were attempted. Discussion of these results has been widespread and detailed[15], but for our purposes three headlines need highlighting. The first such headline is that, overall, 37% of the results replicated. That is, out of the original studies that reported statistically

[14] Notable absences included the majority of the senior figures that the letter was originally addressed to.

[15] For a discussion of how the failed replications undermined confidence in the literature see Pashler and Wagenmakers (2012), for criticisms of the experimental design of the replication attempts see Bressan (2019).

significant evidence in favour of rejecting the null hypothesis, only 37% of the replications also found statistically significant evidence in the same direction. The second headline is that for social psychology studies, that figure drops to 25%. Finally, in cognitive psychology, that figure was 50%.

These three headlines indicate firstly, that a significant proportion of the results in psychology do not maintain their statistical trends (significance in a given direction) when replicated by other researchers based on the methods described in the published text, and secondly that this problem is substantially worse in social psychology than in cognitive psychology. Taxonomies of these replications and further discussion of what this means is found in section 1.7.

When each experiment is initially run, the experimenters aim for 0.8 power to establish their conclusion. This means that, given that the null hypothesis is false, the experiment should be sufficiently robust to demonstrate this falsehood 80% of the time if the experimenter′s power estimates are reasonably accurate. The remaining 20% is the scope for failure to reject a false null hypothesis.

When an experiment which accurately rejects the null hypothesis is replicated, we would expect a replication rate reflecting the power of the experiment. This is because, while the hypothesis is true, the experiment only has a given power to establish its truth. Assuming that there is some combination of true rejections of the null (accurate positive results) and false rejections of the null (false positive results) in the overall dataset, and false positive results should replicate at the false positive exclusion rate (usually $p<0.05$) we should expect to see a replication rate of the power of the experiment for those experiments that were true positives and 0.05 for those that were false positives. The overall replication rate will fall somewhere between these values depending on the makeup of the dataset. For any given replication rate, this gives us a relationship between the average power of our studies to reject a false null hypothesis, and the proportion of our hypotheses which were false positives. Assuming the original experiments possessed the target power of 0.8, we should expect an overall replication rate only slightly below 0.8.[16]

Given this, what may we say about social and cognitive psychology′s replication rates in the OSC? And what does this say about the reliability of their methods?

Section 1.5: Social psychology as research programme.

Any demarcation between social and cognitive psychology is, and should be, messy. This includes how we identify their relevant methodologies as research programmes. For the sake of this chapter, I adopt a Lakatosian framework due to its affordances in clarifying the scope of the problem. In particular, motivating the composition of the body of research as research programmes which may be treated and analysed as wholes possessing properties of their own; especially a degree of reliability.

Lakatos (introduced to philosophy of science the concept of the research programme as the fundamental unit for an account of science . Each research programme emerges in response to some problem in our understanding and represents an approach to the resolution of the problem. Each research programme has a target phenomenon or collection of phenomena which it seeks to explain and understand. In the pursuit of this understanding, each research programme consists in conceptual tools and frameworks from which stem experimental frameworks which may test and critique current understanding of the phenomena.

Lakatos (1970, 1978a) divides the conceptual tools and frameworks of a research programme into the hard-core commitments of the research programme and a protective band of auxiliary hypotheses. The distinction between hard-core commitments and auxiliary hypotheses (a distinction rooted in the works of Duhem (1954) and Quine (1951)) allows Lakatos′ reconciliation of Popperian falsificationism (Popper, 1959; and, especially,

[16] This phenomenon is sometimes called ′regression to the mean′ and has been appealed to in order to excuse the low replication rate in psychology. See Schimmack (2020) for an explanation of why such appeals are drastically understating the original problem.

1963) and the Duhem-Quine thesis. Popper′s falsificationism proposes that a prediction which is entailed by a theory, which turns out to be false, allows us to reject the theory which entailed the prediction, by force of *modus tollens*:

1. If scientific theory X, then prediction Y.
2. Not-Y.

C. Therefore, not-X.

What both Duhem (1954) and Quine (1951) highlight, with different emphases, is that a scientific theory is not a monolith from which an adequate, testable prediction can be derived. In order to predict what will happen in this experiment, under these conditions, we need our scientific theory, but also auxiliary hypotheses about the methods, outputs, equipment, etc. which are necessary to construct the full prediction. This underdetermination means that the first premise is a conjunction:

1. If X and Y and Z and⋯, then prediction Y.
2. Not-Y.

C. Therefore, not-(X and Y and Z⋯)

The problem this generates for Popper is that the negation of a conjunction only requires that one participant in the conjunction be false. Therefore, a false prediction may leave a theory entirely intact.

Lakatos uses the distinction between hard-core commitments and auxiliary hypotheses to distinguish those commitments which will ′bear the brunt′ of the falsification. If a research programme with hard-core commitments H and auxiliary hypotheses A predicts that P, which turns out to be false, then some part of A will be refuted. The research programme will preserve its hard-core commitments in this instance and proceed to test new auxiliary hypotheses.

Taken as a whole, the theoretical commitments of a research programme prescribe and proscribe the experimental and analytical tools which are the licit and illicit means of development of the research programme. (Lakatos, 1970; 1971; 1978; Musgrave, 1976) They identify the methods which researchers may and may not take to test their hypotheses, formulate and present hypotheses and findings, and to distinguish that which is at stake from that which is not. The results of a research program are the consequence of the interaction between these methods of the research programme and the target phenomena, mediated by the research programme′s participants.

Lakatos introduces two concepts which resemble my use of ′method′ which I now briefly introduce in order to distinguish them. The first is methodology, and the second is heuristic. The methodology of scientific research programmes is the meta-analytic framework which distinguishes scientific from pseudo- or non-scientific programmes. This meta-analytic framework is the judgement that some research programmes are progressive and others degenerating – where successive iterations of the programme are explanatorily and predictively superior or inferior to their predecessors. Methodology, in this sense, is the yardstick against which the research programme′s historical position is judged to determine whether it is scientific.

Heuristics in Lakatos′ sense refers to the means of progression of the research programme, these are the manners in which participants in the research programme address challenges faced by that research programme. Specifically, the licit ways in which the auxiliary hypotheses of the programme are adapted and updated is described by the programme′s heuristics. While methodology is common to all research programmes, each research programme has its own heuristics by which, and according to which, the programme responds to internal or external challenges. (Larvor, 1998, pp. 53 -56)

My use of method throughout this thesis contrasts with these two concepts in two ways. The first is that method refers solely to the practice at a given point in time and is, or imagines itself to be, ahistorical. The second is that method is defined as the set of practices undertaken by the participants in the programme as participants in the programme and the justifications for those practices which are grounded in the hard core or auxiliary hypotheses of the research programme. Method in this sense is intended to be familiar to the usage of the word within the sciences, including psychology. The conclusion which is formulated from a given study is therefore constructed through the method of the research programme subject to external constraints and considerations, out of the results obtained.

We find an effect between this set of Xs and this set of Ys under some condition C, which were each identified as such as part of the method of the research programme and sampled according to our external funding and researcher time constraints. We then extrapolated from these results, again in accordance with our method, to a more general conclusion about the operation of Xs and Ys under C, or some class of Cs.

Method allows us to describe the problem of replication as a problem for the research programme. The research programme incorporates a set of justifications of tools and approaches that make these tools or approaches licit. Method describes these features of the research programme at a time. Method, therefore, describes all the systematic relationships between the target phenomenon and the conclusions drawn within the research programme, that the programme itself considers licit means of production. Incorporating experimental design, interpretation, reporting and response; method in this sense describes everything which, if known, will not elicit demands for the authors′ apology or retraction.

The scope of a given method utilises the Lakatosian framework of a research programme as the appropriate level of description for science. In this sense, there is no ′scientific method′ though there is a ′method of particle physics′, or a ′method of social psychology′. In this way we can describe systematic features of the research programme, that are not illicit by its own reckoning, as features of the method. As such, when we describe systematic failure to replicate in a research programme, we are describing a phenomenon that arises from the method of the research programme.

Section 1.6: Replication, Replication, Replication

While we may now meaningfully speak of properties of a research programme, we have yet to clarify what is meant by replication in the sense undertaken by the Open Science Collaboration (2015) and thereby what is meant by a given replication failure, or by a systematic failure of a research programme to produce replicable findings.

This section introduces two accounts of replications to frame what happened following Kahneman′s open letter. The first taxonomy is Schmidt′s (2009) distinction between direct and conceptual replications. This is followed by Machery′s (2017) argument for a general resampling account of replications which gives us reason to reject the category of conceptual replication altogether. In its place, the resampling account offers a fourfold account of replications as well as two dimensions of extension.

Schmidt (2009) distinguishes between two kinds of replication – direct and conceptual. A direct replication is an attempt to run the experiment again, as exactly as possible. Schmidt identifies three functions which are served by a direct replication. By running a second sampling of the dataset, direct replications first control for sampling error in the original publication, and second control for artefacts where there are unexpected interactions between the experimental context and the target phenomenon. Third and finally, direct replications control for fraud by assessing whether purported phenomena can be elicited in the way described in the original publication.

A conceptual replication is an attempt to verify the underlying hypothesis of the original experiment. In order to do so, a conceptual replication requires a different experimental setup which targets the same phenomenon. Schmidt gives the example of the conceptual replication of Rosenthal and Fode′s (1963) finding that students findings of the speed that rats complete a maze can be influenced by telling the students that a random subset of the rats were maze-bright while others were maze-dull. Rosenthal and Rubin′s (1978) experiment tested whether teachers who were told that some of their pupils would show remarkable improvement over the next few months would evaluate those students′ progress over that period differently.

This replicates the target phenomenon – information conveyed to participants shapes their report of relevant outcomes – but entirely changes the experimental setup using different measures, of different kinds of participants, with substantially different contexts.

This taxonomy is the most widely employed in psychology because it affords ready description of the practice of replication from the perspective of a researcher intending to conduct such a replication.

Machery (2017) argues that we should understand replications generally (rather than taxonomizing them specifically) as a ′resampling′. On this account ″Experiment A replicates experiment B if and only if A consists of a sequence of events of the same type as B, while resampling some of its experimental components in order to assess the reliability of the original experiment.″ (2017, p. 556) This means that when we attempt to replicate some experiment, what we are trying to do is sample the r andom variables of the experiment again to see if our data evidences the same phenomenon which was apparently evidenced by the original data.

On this account, there are four dimensions of resampling which could be replicated. Schmidt′s (2009) account of a direct replication resamples the experimental units in order to perform the functions described above; we run an experimental procedure (treatment, measurement and setting) as similarly as possible to the original experiment, resampling the experimental units, in order to control for sampling error and artefacts. The variable which is intentionally changed between the original experiment and the replication is the experimental units – those entities to which we apply treatments and measure the reaction of. What the Resampling Account highlights is that this is only one of the four dimensions of experiments which could be resampled.

Machery then describes conceptual replications as resampling some combination of the original experiment′s treatment and the measurement employed. That is, a substantively novel experimental procedure is employed in order to test the conclusion which will utilise a different treatment or different measurement or both. Given Schmidt′s emphasis on generalising the conclusion drawn in the original, capturing Schmidt′s account of a conceptual replication likely also includes deliberate variation in the experimental setting. For example, an experiment which takes the conclusions from prior experiments using an online questionnaire and tests whether the findings are repeated in an in-person questionnaire with a similar population are arguably deliberately changing the experiment′s treatment, measurement, and setting.

This diagnosis of conceptual replications leads Machery to argue that conceptual replication is not a useful category because it ″does not specify what a psychologist must do (resample, change the value of a fixed factor, etc.) to an experimental component for her experiment to count as a replication.″ (2017, p. 562.) In addition to this lack of prescription when planning a replication, the statement ″Y is a conceptual replication of X″ confuses three kinds of differences between the experiments.

First, if Y changes a fixed, rather than random, variable – for example running the experiment longitudinally rather than as a one-shot – the experiment is not testing the same conclusion. It is only legitimate to statistically generalise from the observed outcomes of the experiment to the unobserved phenomena if the component is a random factor. If the original experiment held that aspect of treatment and measurement fixed, then it did not have legitimate statistical grounds to extend those findings to longitudinal studies. If the original conclusion does not statistically legitimate a given implication to be drawn for longitudinal studies, then the longitudinal studies cannot be said to be testing the same conclusion, or even a conclusion entailed by the findings of the original study.

Second, if Y is resampling from the population of treatments, measurements, or settings of the original experiment then it is testing one or more of the statistical generalisations of the original study. The findings of such an experiment necessarily have a direct bearing on the conclusion of the original study if the second study has the other features of a ′good study′ [18]. If

[18] Well-designed experimental procedures, well-defined outcomes, high statistical power, stringent p-values relative to the prior probability of the null, etc.

the same online questionnaire is being run online by the same group of experimenters on a similar demographic, but slightly different vignettes are being used to test elicit the previously identified effect, this is a resampling of the population of treatments as the original study took its vignettes to be sample vignettes from a population of such treatments which may elicit the observed effect. The original study′s treatment of its vignettes as a random variable legitimated their statistical generalisation to conclusions about a population of vignettes. The second study is testing the statistical generalisation of that first study and hence the conclusion legitimated by the statistical generalisation.

Third, if Y is sampling from a distinct population, then it is testing a distinct conclusion. If X studied the effects of a drug of a given dosage on patients with cerebral palsy, Y is testing a distinct conclusion when it studies the effect of that drug, at that dosage, on Alzheimer′s patients. Whatever the findings of Y, they constitute a novel discovery and are not a repeat of an existing finding.

These three differences between the kinds of experiment grouped as conceptual replications includes experiments which are testing conclusions which cannot be statistically legitimated by the findings of the original study, as well as those which are, and finally those which are testing an entirely novel hypothesis. Given the confusion of these, very different, experiments under the label we ought to abandon its use.

Rather than grouping subsequent experiments into direct replications, conceptual replications, the Resampling Account offers a systematic typology of replications based on the variable which is being resampled. When experimental units are being resampled from the experimental population, this is an experimental units replication. When treatments are being resampled from the population of treatments, this is a treatment replication. When measurements are being resampled from the population of measurements, this is a measurement replication. Finally, when settings are resampled, this is a setting replication. This exhausts the types of experiment which are testing the same conclusion as the original experiment by testing the statistical generalisations that legitimate that conclusion.

When a second experiment samples from a different population, one which is distinct from the population sampled in the original experiment, this is an extension (call this type 1) of the original experiment. Its findings are novel and will not necessarily have any bearing on our understanding of the reliability of the original finding. Similarly, when a second experiment changes a fixed variable in the first experiment this is also an extension (type 2) of the original experiment, testing a novel hypothesis. Each of these two types of extension can be nested in each of the four components; a novel measurement population extension (type 1, measurement), a new dosage treatment extension (type 2, treatment), etc.

This taxonomy of replication attempts brings with it a commensurate taxonomy of replication failures; experimental units replication failure, treatments replication failure, measurement replication failure, and setting replication failure. These failures share certain features which this section will close by briefly outlining. This assumes good experimental design and practice, sufficient power to reasonably exclude false-negatives, and significance attained to a high enough degree, relative to the prior probability of the hypothesis[19]. As a result, this assumes the reasonable exclusion of false negative results in the resulting replication.

Each type of replication failure gives us reason to doubt one of the statistical generalisations across a target population which motivated the original conclusion. As a result, insofar as the replication is sound, we have reason to doubt the original conclusion. We no longer have evidence that the purported effect exists across the purported population – whether this be a population of persons, dosages, implicit measures, or cultural settings. This finding underdetermines the required response from researchers, but it demands some response – both the abandonment of the original conclusion

[19] Questions surrounding what constituted adequate power and significance levels are returned to in detail in sections 2.3 and are related to prior probabilities of the falsehood of the null hypothesis in section 2.4.

and its finessing and retesting through a new mechanism, or over a different population, are entirely legitimate responses to such a replication failure. The only thing that a high-powered replication failure renders illegitimate is treating the original result as continuing to vindicate the original conclusion; continuing as though we are still justified in thinking that the purported effect extends over the purported population.

As a result of the theoretical advantages, both prescriptive to those designing replications and descriptive for those trying to interpret them, of the resampling account over Schmidt′s taxonomy, the resampling account will be utilised throughout the thesis. As such, ′replication′ is henceforth not used as a success term. It incorporates both ′successful replications′ and ′replication failures′ (Machery, 2017, p. 556, footnote 8).

Section 1.7: Replication failures as problem for social psychology.

What does this account of replications say for a research programme whose constituent findings and results tend to fail to replicate at a problematically high rate? This section begins by clarifying what this means for our ability to rely on the conclusions of that research programme. It concludes by drawing this theoretical case back to the findings of the Open Science Collaboration and subsequent replication attempts. Doing so shows we have good reason to think that the method of constructing and evidencing conclusions in social psychology, *per se*, is unreliable.

If a significant subset of good replication attempts fail, something about the conclusions drawn about the target phenomena of that research programme is not borne out when the components are resampled. That is, the conclusions express generalisations of data to phenomena, and either the data is unreliable, the generalisations are unreliable, or there is some more fundamental problem in our method.

In an individual case we might either hold a moratorium on judgment about the conclusion or constrain the population for which we take the conclusion to hold and carry on with researching the phenomena. But if it is generally true that a preponderance of conclusions in a field do not replicate this is not simply a challenge of the reliability of getting at the truth but a challenge of the reliability of the method we take to construct our conclusions. We do not simply have reason to doubt that this conclusion holds in the form it is presented, but that most of the conclusions of the research programme hold as presented – not because they might not replicate, but because however their conclusions are constructed does not ensure the strong links to well-founded evidence which should ensure higher replication rates. In this sense the method of the research programme is unreliable.

Given this worry for any given research programme whose conclusions tend not to replicate, we may ask whether social psychology is such a programme. I begin with a discussion of the Open Science Collaboration (2015) as well as a key response – that several replications were very different to the original experiments. This illustrates the importance of a resampling account in understanding replications and their consequences. It closes by highlighting subsequent work in replicating social psychology and concludes that these give us good reason to think that something about the method of social psychology makes it unreliable.

The Open Science Collaboration (2015) found that only 25% of the sampled social psychology papers replicated, these replications were mostly resampling the experimental units. They utilised the description of the relevant factors included in the methods section of the published articles to conduct replications which, as closely as possible, resembled the original study. In most cases the treatment, measurement and setting were kept the same as defined in the original study. As highlighted above, these experimental units replications give us reason to doubt that the purported effect obtains over the experimental unit population described, unlike in the ideal analysis from the previous section this reason for doubt is proportional to the quality of each replication study – its design, power, etc. In some instances, this resampling of the experimental units was done in conjunction with resampling of treatments.

Shnabel and Nadler (2008) investigated the hypothesis that victims and perpetrators are each deprived of some psychological resource. This hypothesis implies that reconciliation between victim and perpetrator requires the addressing these resource deprivations, which are different for victims and perpetrators: victims are deficient in their sense of power, perpetrators in their public moral image (Shnabel and Nadler, 2008, p. 117). Study 4 tested this hypothesis by presenting participants with vignettes where an employee, who was absent from work for 2 weeks, returned to work to find that a colleague of theirs who had temporarily filled their position had been promoted to their job, while they were demoted (p. 127). Participants were undergraduate students in Tel Aviv (75 female, 19 male; mean age 23.5). The gender of both employees was matched to be the same as the participant. The reason for the absence was maternity leave, for women, or military reserve duty, for men - the most common reasons for extended work absences in Tel Aviv. The study asked participants to imagine themselves as either the demoted or promoted employee depending on condition, before responding to a second part of the vignette where the participant imagined receiving praise from their antagonist in a subsequent staff meeting. The praise either addressed their professional skills or interpersonal skills. The study reported that messages of social acceptance was more effective in promoting reconciliation from perpetrators, whereas a message of empowerment was more effective in promoting reconciliation among victims.

As part of the Open Science Collaboration, this hypothesis was replicated (Gilbert, 2015) sampling 144 undergraduate students at the University of Virginia (62 female, 82 male; likely slightly lower average age, actual age not reported). The reasons for a work absence from the original vignettes were not relatable for the replication study′s participants since, in the US, ″reserve duty isn′t common for men, and demotions for maternity leave are illegal″ (Nosek and Gilbert, 2016). The replication study therefore changed the vignettes, for both male and female participants, to be about a victim who takes leave for a honeymoon. The replication study failed to obtain an effect.

Nosek and Gilbert (2016) take these changes to be both necessary (to accommodate a change in setting between the original and replications studies) and irrelevant for the replication because they shared the same structural features. On a resampling account, the claim is that they are part of the same population of vignettes which were party to the generalisations made to by Shnabel and Nadler in drawing their conclusion. While Gilbert et al. (2016) identify this as reason to contest that the second experiment constitutes a replication because of the change in both setting and treatment, the resampling account clarifies what is at stake if one takes this response.

If a resampling fails to find an effect, with good power to do so, and the response offered is to claim that the change in setting or treatment is sufficient to claim that this attempted resampling is not a genuine resampling, then we are obliged to constrain the scope the conclusion drawn in the original article. If the replication attempt has sampled from a population that the original experimenters consider illegitimate, then the original conclusion can be clarified to make this explicit.

Shnabel and Nadler summarise their hypothesis: ″victims must restore their sense of power, whereas perpetrators must restore their public moral image″ (2008, p. 116), and summarise their findings that ″being a perpetrator threatens one′s public moral image, resulting in a greater need for social acceptance, and that being a victim threatens one′s sense of power, resulting in a greater need for power″ (2008, p. 129) and ″reconciliation between adversaries depends on satisfying relevant emotional needs and restoring their damaged psychological resources.″ (2008, p. 130)

The response to this replication from Gilbert et al. (2016) is that the replication differed from Schnabel and Nadler′s (2008) experiment in substantial ways: an ″original study that asked Israelis to imagine the consequences of military service was replicated by asking Americans to imagine the consequences of a honeymoon″ (2016, 1037b). That is, the replication attempt is engaging a different population with a different treatment in a different setting.

What the resampling account of replications highlights is that, if this response is to be taken seriously, that it is illegitimate to consider the second experiment to be a replication of the first, then this must be because they are sampling different populations (if both experiments are ′good′ by other metrics). If we take it to be true that the replication established no evidence for the claim that American victims must restore their sense of power, whereas American perpetrators must restore their public moral image, and that American victims are a subset of victims, then we must refine the original conclusion such that it does not rely on a population extension which the response considers over-broad.

On Gilbert et al.′s response, Schnabel and Nadler ought to have concluded only for victims/perpetrators from Tel Aviv and not for victims/perpetrators *per se*. If this response it taken, the pressure to contract to the population is in tension with the central hypothesis, the needs-based model of reconciliation: that victims and perpetrators are deprived of unique psychological resources and effective reconciliation involves meeting these needs. If we must contract our claim to Tel Aviv students, this model of is no longer a model of reconciliation *per se* for victims *qua* victim, or perpetrators *qua* perpetrator. The internal logic of the model is that becoming a victim and becoming a perpetrator involved particular deprivations which are necessary to redress for reconciliation. If this is only the case for those in Tel Aviv, there is no model-internal explanation for why this should be the case. This is not to say that the response cannot be made, only that the theoretical cost to the original hypothesis of making the response is substantial and overlooked by Gilbert et al. (2016).

The clarification offered by the resampling account makes responses of this sort far less palatable as it clarifies the theoretical cost to be borne by the original study: the richness and scope of your claims must be curtailed to exclude the populations to which you consider generalisation illegitimate.

This is illustrated by highlighting how a response can work against the replication: that in the replication, the vignettes were less relatable to the participants and failed to evoke the relevant victim/perpetrator responses as a result. That is, the change in treatments, while still successfully structured as a victim/perpetrator treatment, fails to gain traction because the treatment must also be relatable to the participants to elicit the relevant effects. To test this claim, Baranski et al. (2020) , as part of the Many Labs 5 project, attempted to test the effect of changes in the vignettes across multiple labs. By beginning their study with a pilot study, to obtain feedback from participants about vignettes that would, or would not, be relatable as victim/perpetrator treatments, the study presented a more relatable vignette. The victim condition told participants to imagine that being a recently unemployed university student. On returning from visiting family for 2 weeks their roommate tells them that they had found a new roommate who could commit to paying the next year′s rent and that they had to move out. The perpetrator was told to imagine finding a more reliable roommate to replace their recently unemployed current roommate, who had left for 2 weeks to visit family and telling them they would have to move out.

By running a pilot study to check that the vignette was sampling a relatable victim/perpetrator treatment, Baranski et al. replicated the original effect from Shnabel and Nadler (2008) successfully and by testing moderation by protocol (the effect of changing treatment) found a significant moderation effect. By running the experiment across multiple labs with a much larger sample of participants (N=2738) the subsequent replication was able to isolate the kind of resampling which failed to obtain in the original replication.

By highlighting the sampled populations of an experiment and identifying replications by the populations they resample, we can make responses to replications much more specific and precise. Gilbert et al. (2016)

are wrong to highlight the difference between military service and a honeymoon, the relevant population of victim/perpetrator vignettes is still being sampled[20]. They would be correct to highlight that the treatment is not being resampled, since the unrelatability of the new vignettes means that the population of *relatable* victim/perpetrator vignettes has not been successfully resampled.

Gilbert et al. further offered the reply that the structure of the Open Science Collaboration′s replications doesn′t provide enough power to give a reliable estimation of the replication rate. Even if we are successfully resampling the original experiment, we must be doing so with sufficient power to reliably evidence true effects. The findings of the Open Science Collaboration with those of the Many Labs Project (Klein et al., 2014) which ran replications of each of the thirteen target studies 35 or 36 times across different labs and pooled the data. Taking the pooled data on each experiment, the Many Labs Project found that it replicated – where they found an effect in the same direction as the original study – for 85% of the 13 original studies[21]. Gilbert et al. go on to note that only 34% of these replications found an effect size within the confidence interval of the original study – the more powerful design found some effects much larger and some much smaller than the original, only a third were close in magnitude. They conclude that the low power of the OSC has led to an underestimation of the ″actual rate of replication″ (1037b). This amounts to a counterfactual claim – if we were to run each of the replications in the OSC 35 or 36 times we would find that we observe an effect in the same direction as the original experiment far more often than running the experiment once.

This claim raises a slightly different problem for the original studies. Increasing power increases our ability to detect an effect if it is present, even for very small effects. Meehl (1990) argues that in psychology we ought to be

[20] This is highlighted by the strength of the manipulation checks obtained by both Schnabel and Nadler (2008) and Gilbert (2015).

[21] Four variations of the anchoring experiment (Jacowitz and Kahneman, 1995) were replicated, bringing the total experiments replicated to 16.

concerned about the ʹcrud factorʹ – in psychology everything is, to some extent, correlated with everything else. As a result, correlational studies should always expect to find some correlation. While randomisation in experimental studies is intended to filter out these background correlations, the power of randomisation to do so depends in part on our ability to randomise along the dimensions of all of the relevant factors, which we want to control for. Our ability to randomise our experimental conditions in this way in turn depends on the quality of our theoretical framework surrounding the purported effect, something we have reason to be concerned about in psychology (Irvine, 2021). As a result, even experimental studies should be wary of ʹcrudʹ in their results especially in areas where the mechanisms of the purported effect are less well theorised and demonstrated.

Given the potential to obtain a ʹcrudʹ result, if all we are interested in is obtaining a non-zero effect, we have only to increase our sample size sufficiently to guarantee finding a statistically significant effect. If a single study is designed to find a small effect (*d*=0.3) with good power (0.8, n=175), running that experiment 36 times and collating the data (n=6300) gives us enough power to detect a minute effect (*d*=0.05) at the same rate (power = 0.8012)[22]. The problem is that doing so can lead us to use these large studies to postulate a special effect between the targets of the study, when what we are observing is in fact ʹcrudʹ – the basic interconnectivity of psychology. Our effects will be real, and our diagnoses of them will be false. Avoiding this requires sensitivity to the actual effect sizes we obtain and the stability of those effect sizes (especially the confidence intervals for very large sample sizes), as well as to the theoretical structures which inform our generalisations (Irvine, 2021).

In the case of the MLP, as highlighted by Gilbert et al., only 34% of the replications found an effect within the confidence intervals of the original study, and the original studyʹs effect size fell within the confidence intervals

[22] All calculated for two-tailed independent t-test.

of the replications for only 12.5% of the experiments [23]. That successful large-scale replications vindicate the existence of some effect between factors, is not the same as the claim that a successful large-scale replication vindicates the conclusion of the original study. This is especially the case if the effect size is volatile, and the original conclusion is founded on an effect size which is unlikely to fall within the confidence intervals of a large-scale replication′s effect size. The MLP vindicates the claim that some psychological studies are observing a genuine effect, in the right direction, but also indicates that those studies are often poor indicators of actual effect size.

When we offer an explanation of a phenomenon, the magnitude of the effect is integral to that explanation. A small effect and a very large effect cannot be substituted for one another in an explanation. Jacowitz and Kahneman′s (1995) introduce a method for measuring anchoring effects. Their observed effect size for the anchoring effect on participants estimates of babies born per day in the US is around d=1, a very large effect size for psychology. The replication found an effect size of around d=2.4, an astronomical effect size for psychology. An explanation of a large effect of anchoring on subsequent estimated by participants of the number of babies born in the US might rely on the involvement of some social outsourcing of knowledge to the ′anchor′ answer, where the subsequent response assumes some significant reliability of the original answer. The latter effect size is so large that this explanation would undervalue the effect of the anchor on the estimate: an effect size of 2.4 is something closer to epistemic deference to the anchor than merely being informed by it. An explanation of a smaller effect size will grossly underrate the relevance and influence of a larger effect, and an explanation based on the latter would give greatly undue credit to the former.

[23] Low vs. High category scales (Schwarz et al., 1985) and Sunk Costs (Oppenheimer et al., 2009); also notable, Correlation between implicit and explicit math attitudes (Nosek et al., 2002) is a near miss.

While this response from Gilbert et al. does indicate that replication rates are likely somewhat better than indicated in the OSC, it also emphasises that the conclusions drawn in the original articles are unlikely to be founded on a true estimate of effect size. They are even unlikely to include the true value of the effect size in their confidence intervals. If our confidence intervals, more often than not, fail to include the true effect size, we are failing to account for some significant factors in our analysis, and hence in our conclusions.

Section 1.8: Replication and research programmes

We may now combine the discussion of the consequences of a given failed replication with the framework of research programmes; what does the relatively low replication rate in social psychology say about the research programme of social psychology′s method?

If a significant subset of good replication attempts fail, this represents a challenge to the general method of conclusion construction in the research programme. The conclusions are not bearing out when the components are resampled. Either no effect is found, or the confidence intervals for effect size are substantially different from the true effect. It is not simply that we have reason to doubt that such and such conclusion. Rather, we have reason to doubt most of the conclusions of the research programme. In this sense the method of the research programme is unreliable. Something about the way social psychologists hypothesise, formulate methods, run experiments, collate data, construct conclusions, present data, and peer review, is unreliable.

If we must accept the published corpus of social psychology as it stands − where our only heuristic for judging the reliability of a given finding is its presence in the literature − this commits us the refined version of the argument against reliability:

i. Method M_P is not reliable.

ii. Social psychology publications which use M_P give us a novel reason to believe their claims if, and only if, method M_P is reliable.

iii. Social psychology publications that use M_P do not give us a novel reason to believe their claims.

Section 1.9: Concluding remarks

Something about how social psychologists hypothesise, formulate methods, run experiments, collate data, construct conclusions, present data, and peer review, is unreliable and needs re-addressing. Each area of this overarching method admits of several responses[24] currently available in the literature[25]. In particular, the overarching concern about the robustness of our theoretical framework is the target of the response offered in section 2.6.

One option which will not be further pursued in this thesis is addressing the tendency of psychological studies to be underpowered. Underpowered studies not only have a lower chance of presenting a statistically significant effect when the null is false, they also have a lower positive predictive value for statistically significant results[26] (Button et al., 2013, p. 366). This means that low-powered studies that find significant results have a significantly lower[27] chance of representing true rejections of the null than well-powered studies which finds a significant result. Addressing this challenge will require some significant shift in the norms of reporting our results and in experimental design; either a more stringent significance level (Benjamin et al., 2018; also discussed as a selection heuristic for readers in section 2.3), a specific justification of the appropriate alpha (Lakens et al., 2017), adoption of Bayesian statistical methods (Wagenmakers, 2007), or a combination of approaches from more extensive statistical training to multi-lab projects

[24] Some of which are the topic of sections 2.3 and 2.4.

[25] In line with the Lakatosian approach adopted throughout the thesis, I take any given rational response to be underdetermined by the problem.

[26] This point will be returned to in greater depth in section 2.3.

[27] For a prior of 0.1, an experiment with significant results ($p<0.05$) and 0.5 power represents a true rejection of the null only 50% of the time. At the same prior, an experiment with significant results and 0.95 power will represent a true rejection of the null 66% of the time.

(Munafo et al., 2017).

Turning social psychology into a progressive research programme depends on one or more of these approaches being successful in improving the statistical power of studies in social psychology. The proposals of chapter two present a problem for social psychology even if one or more of these proposals is successful in increasing power, but this does not undermine the importance of improving social psychology′s practices in this area.

The consideration of replication failures for method in social psychology is crystalised by the resampling account of replication into a clear injunction for social psychologists. If we have reason to reject the generalisations that ground our conclusions, as a resampling account of replications indicates, we ought to moderate those generalisations and reconsider the populations to which effects, even well evidenced effects, can be legitimately generalised. This requires that we explicitly state those populations in initial findings and be willing to revise these generalisations given further evidence about their legitimacy (Machery, 2017, p. 565). In practice, this requires both a transparency of authors with readers and a transparency of authors with themselves. The former is a communication problem − how to effectively communicate the generalisations we are making and our grounds for doing so. The latter is a self-reflective epistemic practice − how to identify where and to what we are generalising, and how to identify the generalisations for which we have warrant.

Finally, if we need to re-address our generalisations this could indicate a more general worry: the theoretical frameworks, the hard-core commitments and auxiliary hypotheses, of our research programme need rethinking. Concerns about the scope of our generalisations would be one way this is expressed, but it applies equally to concerns about the individuation of entities, assumptions about how and when interactions occur, or more fundamental questions about which posits of the research programme are, or should be, basic. By addressing this more general version of the problem, the proposal in chapter 2 also offers some traction in addressing the specific problems of generalisations.

Chapter Two: Vivisecting Social Psychology

Section 2.0: Introduction

Following the conclusion established in the previous chapter, social psychological publications do not give us substantive reasons to believe their claims because they do not have a reliable method for constructing reliable conclusions. This chapter takes up this problem and proposes a series of responses. Following much of the recent literature (Benjamin et al., 2018; Nosek et al., 2021; Schimmack and Brunner, 2018; Schimmack and Bartos, 2020, *inter alia*), the chapter begins with a focus on the reliability of the statistical trends upon which the conclusions in the literature are predicated. With some plausible, potentially fruitful responses to the statistical problem clarified, the problem is restated as a problem of the theoretical framework through which the data is interpreted and within which the conclusions and claims are situated. By clarifying the scope of this problem in relation to the problem of the reliability of data, the proposed solution is presented as a pragmatic dialectical tool for the clarification and critique of the ontological status of the conclusions we are presented with in the literature.

The argument addresses two audiences. The first audience are the readers of social psychology, largely those who work within related fields, who go on to use social psychology′s findings as the basis of some further work. The second audience are those social psychologists working within the field who appreciate the breadth of the challenge the replication crisis in social psychology represents.

The problem of the challenge presented in chapter one is set out and framed for each audience according to their demands of social psychology and responsibilities to it as good epistemic interlocutors in section one. Section two sets out the first response available to readers of social psychology: to disregard social psychology′s conclusions as the reasonable basis of beliefs or as grounds for further research. This section presents the radical solution in order to emphasise the magnitude of the problem and as the first horn of the dilemma that will be put to readers. The alternative solutions are epistemically demanding, perhaps prohibitively demanding. The radical solution is presented to highlight that this is the only epistemically acceptable alternative to a full engagement in the epistemically demanding options. Section three introduces the first of the epistemically demanding options: engage only with those social psychology conclusions founded on p<0.005, rather than p<0.05 (Benjamin et al., 2018). The most epistemically demanding option is introduced in section four: implementation of PPV/prior curves. Section five highlights that these statistical solutions only address a part of the problem that the replication crisis has revealed; the theoretical framework within which the claims and conclusions of social psychology are situated is contradictory, ambiguous or non-existent. In Lakatosian terms, the research programme is either unclear about or lacking a hard-core theoretical commitment which will resist falsification, and as a result is also unclear about or lacking a well-defined protective belt of auxiliary hypotheses which are at risk of falsification. Section six presents a systematic, optimistic approach to resolving this deeper problem: the ontological critique and how it is implemented by readers and authors. The conclusion draws together what these several conclusions mean, in practice, for readers and authors of social psychology. In particular, the link between the criticism of the theoretical framework and the identification of the generalisations being made is made explicit. A sample ontological critique form which could be appended to papers, preregistrations or funding applications is presented in the appendix.

Section 2.1: The problem

The argument from chapter one established that, addressing the field as a whole, social psychology has an unacceptably high false-positive rate, and as such any given publication *qua* its membership in that set carries an unacceptably high false-positive risk. If we select a random publication and know nothing more about it than that it is a part of the field of social psychology, we would be epistemically reckless in treating it as good grounds for forming or justifying beliefs, or in predicating further research upon it.

Simultaneously, we may posit that it is not truly random which studies are reliable and which are not. The reliability of any given publication will depend on a multitude of factors besides it being a social psychology publication. Everything from the robustness of the experimental methodology, to the statistical approaches applied, to the interpretations offered will be a relevant factor in the reliability of the conclusions the publication draws.

Finally, we may posit that many, if not most, of these factors are obscured in, or omitted from, the final publication. For reasons of their perceived irrelevance to the reliability of the conclusion, conformity with existing research or publication culture, intentional fraud or mere oversight, these factors are systematically excluded from publications in social psychology, as in many other fields.

The problem that arises from this combination of factors has two faces, one each for its two audiences. The first audience are readers of social psychology who may wish to predicate further research upon the findings they read therein or to form, or justify, beliefs thereon. The second audience are those authors who work within social psychology who are both engaged with the research that has come before them and attempting to produce research upon which others may rely.

The first face of the problem is that readers do not have access to most of the factors that make a given publication reliable or unreliable, and so must rely on some schematic proxy for the true determinants of reliability. What should they rely on in order that they avoid being epistemically reckless?

The second face of the problem is that authors need a way to ensure and communicate the reliability of their particular work in such a way that it is differentiable from unreliable work, while simultaneously avoiding predicating that work on previous, unreliable work in the field. How should they ensure their reliability? How should this guarantee be communicated?

And how should they avoid propagating the unreliability of previous work into new work?

The faces of these problems will be treated as distinct, though the boundaries between the audiences is at best fuzzy and at worst non-existent. This analysis of our position in relation to social psychology deliberately privileges what we demand of the field over other relevant concerns. The cost of doing so is balanced against the clarity this offers our practical analysis of what we must do in response to the replication crisis. While this is a cost we should pay in order to proceed, the messiness of the distinction between audiences should be borne in mind when we come to apply the resulting recommendations, both for ourselves and others.

What follows is a series of proposed responses to the replication crisis, which will be evaluated in terms of their efficacy in answering the problems as presented to each audience.

Section 2.2: Disregard social psychology.

The first response to the problem which must be raised is primarily a response for readers, though a variation of it is available to researchers. Readers may disregard social psychological research as providing any relevant evidence. For authors, the corollary may mean either treating any and all research which precedes their own as failing to offer any relevant evidence which may inform their current research questions, or they may treat social psychology as a degenerative research programme (Lakatos, 1978) and may decide to abandon its pursuit on that basis.

This first response is the most radical and is presented here for two key reasons. The first is that it is an option which many researchers have adopted and will continue to adopt and deserves to be seriously addressed as a result. The second is dialectical. The options that will be presented in the rest of this chapter are, to varying degrees, epistemically demanding. The dialectical purpose this first proposal serves is to highlight that the alternative to these demanding options is not 'business as usual', but the abandonment of social psychology as inadequate to any task to which we might wish to put it, both by readers and by authors.

The lack of *prima facie* obvious means of differentiating ′good′ papers from ′bad′ papers, even by those with experience in social psychology [28], is the root of the intractability of the problem and the reason why the other responses are epistemically demanding.

Disregarding social psychology as evidence does not necessarily require one to cease interest in what is being done within the field, nor even necessarily to stop reading the material published within it. However, it is an injunction not to regard the data, models, theories or conclusions of social psychology as providing us with justification for our existing beliefs, or for the formation of novel beliefs, in order that we avoid being epistemically reckless .

[28] Citation rates in psychology do not correlate with replication success. Papers founded on experiments that do not replicate are cited at near identical rates to papers founded on experiments that do (Yang, Youyou, and Uzzi, 2020, figure 2. 1), for both direct and second-degree citations. It is not clear whether these citations represent negative citations of papers whose experiments do not replicate, however, for the citation rates to diverge significantly between results that do and do not replicate, a preponderance of these negative citations would need to fall on the papers whose experiments do not replicate, and negative citations would need to make up a substantial minority of citations. Findings in some fields suggest a negative citation rate far lower than needed (for example 2.4% in immunology: Catalini, Lacetera, and Oettl, 2015), even if every negative citation ′correctly′ identified a result that would not replicate. As such, the conclusion that the citations of results that do not replicate represent negative citations seems unlikely. This is compounded by the prevalence in psychology of reasons to cite negatively which are not correlated with replicability, such as trivially false null hypotheses (e.g. ′crud factor′: Meehl, 1970), poor experimental design (Cohen, 1962), poor interpretation of otherwise reliable results (Schimmack, 2020), etc.

The benefit to adopting this strategy, going forward, is the avoidance of some unknown but significant proportion of false beliefs. Moreover, the subject matter of social psychology includes many questions of substantial social and moral weight, wrong answers to which bear costs we would not want to countenance. However, the strategy also incorporates significant costs. I begin with the clear compliment of the benefit of this strategy: the increased false negative rate. Stating this risk epistemically highlights and clarifies the more substantial true risk: that this approach closes lines of enquiry more broadly than we would likely countenance, or for reasons we are unlikely to endorse.

The most obvious cost of this approach is the increase in false-negative rate, which will simply become the true positive rate. Expressing this risk epistemically alters its operation slightly from the more familiar statistical representation, since we are not simply worried about rejecting a true test hypothesis but of not forming beliefs that would be true. The cost to scientific progress of false negative errors is sometimes considered lesser than the cost of false positives because of the disparity in the rate at which they are accepted into the literature and their subsequent persistence in the literature. That is, for the project of advancing a given area of science, false positives have a warping effect on the field in a way that false negatives do not.

The questions being tackled by social psychology are ones on which many people already hold beliefs; about race, class, gender and about our tendencies to respond to such categories under typical and atypical circumstances. These are not issues on which we have no settled ideas, for which we can await an improved science of human behaviour without consequence. If we are currently wrong about these issues, that failing has consequences. There are epistemic, moral and social consequences to accepting a result from social psychology that is false, and epistemic, moral and social consequences to failing to accept a result from social psychology that is true and surprising to us. We may decide that social psychology is too unreliable to serve as the basis for new beliefs, but the epistemic framing highlights that this is not a question of the static reliability of the findings but of their reliability relative to our existing notions. This is further problematised by the fact that the continued study of social psychology is among the means by which we may come to know how reliable or unreliable our existing notions are.

This brings us to the second cost of disregarding social psychology as evidence, that this closes lines of enquiry. Disregarding social psychology as evidence closes many lines of enquiry into its subject matter: we are no longer willing to countenance the findings about that subject matter as being a good foundation for knowledge. While social psychology is unusual in its replication rate, it is not entirely unique. Neuroscience in particular has been shown to be unreliable due to its small sample sizes, leaving studies underpowered to establish the purported effects (Button et al., 2013). If we are committed to the purely abstract, statistical approach of avoiding false positives *per se*, then we would disregard both neuroscience and social psychology equally. However, recognising the epistemic dimension of this problem highlights that we do not avoid false positives *per se*, but according to the weight of the questions being investigated and the implications of arriving at various conclusions, true or false.

The second cost can be represented as a dilemma of disregarding social psychology. If we are to disregard social psychology as evidence, we must either do so because we care only about false positives *per se* and be consistent in rejecting other areas with similar problems. Or we must do so on the basis of the epistemic consequences of accepting these particular false positive risks because we believe the topics and questions of social psychology are less important than those of, for example, neuroscience, and we must make this motivation explicit as a key premise of our argument. The former is unpalatable because of the scope of investigation which becomes closed to us, while the latter requires an affirmation that topics which have shaped many aspects of society are relatively unimportant or undeserving of study.

For those readers and authors unwilling or unable to engage in the more epistemically demanding options set out below, this option may be the most epistemically cautious. However, its costs mean that widespread implementation bears substantial and unknown epistemic risks; individually, for the research community, and for society at large. This emphasises the importance of implementing a strategy of distinguishing reliable studies from unreliable studies in more detail than their subject matter; some candidates strategies are the topics of the following sections.

Section 2.3: P-values<0.005.

One option is to treat as significant only those findings which reach p<0.005[29]. Benjamin et al. (2018) argue that the current standard for statistical significance in many fields, including social psychology, of p<0.05 is too high and that a more stringent test should be adopted for the report of novel discoveries. The argument for introducing a more stringent test begins with analysing what is happening in a null hypothesis significance test (NHST), where significance is set at P<0.05, in Bayesian terms. Their analysis highlights that the insensitivity of the NHST to the prior probabilities of obtaining particular results causes the point value test result to obscure a significantly higher false positive rate than is being advertised, especially when testing risky hypotheses.

The high false positive risk of P<0.05 leads the authors to propose the adoption of P<0.005 as the level of significance for claiming a new finding, with 0.005<P<0.05 being redefined as ″suggestive evidence″ (2017, p. 8). While the paper has generated a great deal of attention and comment, its proposals have yet to be adopted.

The differences between the two audiences have implications for what this strategy looks like in practice as well as the argument for its

[29] Or equivalent. P<0.005 is close to a 3-sigma rule, so for tests that generate non-p-value results should be treated as significant if they reach 3-sigma or greater (Benjamin et al., 2018, p. 8).

implementation. Benjamin et al. (2018) are arguing for a change in norms within scientific practice, in our terms, this is an argument that the strategy be adopted by authors (and editors) for their publications going forward.

What follows is a parallel argument to that presented by Benjamin et al (2018), targeted at readers and authors addressing pre-existing work rather than authors looking forward to future work. While the positive argument for adopting the strategy is essentially the same, the asymmetry in what p-values signify between audiences presents novel challenges.

Readers of social psychology may adopt a rule of treating p-values as significant if and only if they reach $P<0.005$ and results that reach $0.005<P<0.05$ as suggestive evidence. When a social psychology experiment reports $P<0.005$ we are justified in believing that this represents an effect that is likely to replicate and remain robust; this experiment has given us good evidence against the null hypothesis. When a social psychology experiment reports $0.005<P<0.05$ we may regard this as indicative of a plausible avenue for future research but not as significant evidence in favour of rejecting the null hypothesis.

Null-hypothesis significance tests are insensitive to prior odds of a test hypothesis being found to be true. This insensitivity is the motivation Benjamin et al. (2018) offer for their proposal. To illustrate the extent of this insensitivity they plot the false positive rate against statistical power[30] for $P=0.05$ and $P=0.005$ at three different prior probabilities (1:5, an unlikely hypothesis; 1:10, a risky hypothesis; and 1:40, a very risky hypothesis) (Figure 2.1). As power increases, the false positive rate for $P<0.05$ drops to a limit of 20% at a prior of 1:5 and 100% power. For a prior of 1:40, the lower limit of the false positive rate for significance at $P<0.05$ is 66%. By comparison, the minimum false positive rates at $p=0.005$ at these priors is 2% and 17% respectively. Some fields, particularly genomics and high-energy physics, have already adopted 3-sigma and 5-sigma rules for similar reasons

[30] Power is the probability that a test finds a statistically significant rejection of the null, given that the null is not true.

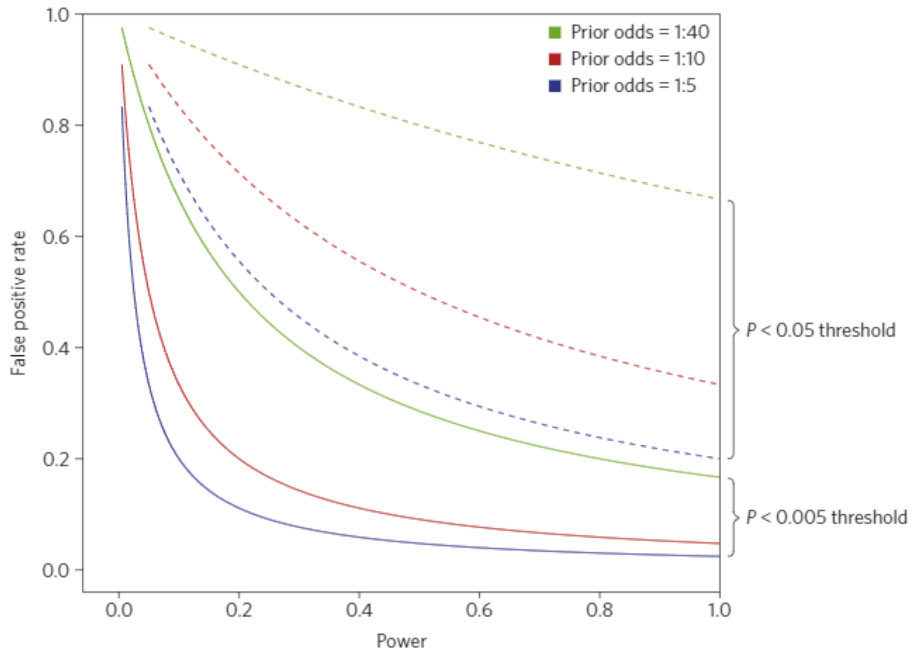

Figure 2.1: False positive rate/power curves for different priors and p-values (Benjamin et al., 2018, Fig. 2, p. 8, R-code: Supplementary Materials, p. 6.)

What this means is that studies conducted at P<0.05 on unlikely hypotheses have a false positive rate four times higher than readers might reasonably expect, even with 100% power to detect a true effect. For more unlikely hypotheses still, P<0.05 becomes wholly inadequate to give us confidence that a rejection of the null represents a true rejection of the null. Whereas, at the P<0.005 threshold the false positive rate remains low even with quite long prior odds. Importantly, this low false positive rate remains even as power falls from 100% to levels more commonly seen in social psychology (Cohen, 1962 estimates mean power in abnormal psychology is around 0.50; Bakker et al., 2012, estimates only 0.35 power for psychology generally).

By adopting this heuristic, readers can reduce the rate at which they are appropriating research that reports false positive findings substantially, without engaging in more demanding interrogation of the data. This enables readers to continue to engage with social psychology without being epistemically reckless in accepting unacceptably high false-positive rates. For many readers, this strategy will reduce false-positive rates to an acceptable level, even for very risky hypotheses.

The primary cost associated with this strategy is that it dramatically reduces the proportion of psychological research which offers us significant grounds to reject the null. This may limit research in several ways. One key constraint is the reduced likelihood of having multiple reliable research papers on a single research question to compare.

If one is interested in investigating sex-based implicit biases in questionnaire responses at P<0.05, there are hundreds, if not thousands, of relevant papers which examine the related phenomena. Within this large set of relevant papers, we may expect there to be several with the combination of robust experimental designs, clear expression of conclusions, and sufficiently large and diverse sample size that we may feel well-resourced for further research based on the findings. While it is difficult to say what proportion of research reaches P<0.005 when it has not been designed to, the overwhelming majority is excluded. Following the above argument this is a good thing, since the majority of this research is unreliable [31]. However, it makes it unlikely that a substantial body of evidence can be collated on a sufficiently narrowly stated question to make research based upon this evidence, by readers of social psychology, possible.

This highlights one of the key differences between the proposals Benjamin et al. (2018) make for authors and this proposal for readers. Their proposal is for a change in norms of publication and reporting; pushing the corpus as a whole to adapt to a more stringent standard. By presenting a proposal for readers, without a commensurate change from authors, the scope for investigation contracts considerably. For many research topics, this contraction will be prohibitive of this strategy and represents the key cost associated with selecting for P<0.005 as readers.

[31] Either it is examining hypothesis which are trivially true, in which case the research does not offer new information (or at least, the statistical inferences from the data do not) or the hypotheses are somewhat to very risky, as a good (in the Popperian sense) research proposal should be, in which case the false positive rates are unacceptably high (>50%).

One concern that may be raised against this strategy is the result that selecting for P<0.005 has on the statistical power of studies. This is relevant for readers as the statistical power of a study is one parameter that determines the positive predictive value; the probability that a given rejection of the null represents a true rejection of the null. The positive predictive value is derived from the following table:

	H_0 rejected	H_0 not rejected.
H_0 True	α	1-α
H_0 not true.	1-β	β

Where α is the false-positive rate, and β is the false negative rate. The pre-test probability of the null is also incorporated as R (the pre-test probability of any given probed effect being a true effect over the total effects probed). The positive predictive value asks what the probability that a rejection of the null (column one) is a true rejection of the null (row two). This gives us the formula:

$$\text{PPV} = \frac{\text{(Reject null given the null is false [= Power * Pre-study odds])}}{\text{(Total rejections of the null [= Power * PSO, plus false positive rate])}}$$

Or:

$$\text{PPV} = ([1 - \beta] * R) / ([1 - \beta] * R + \alpha)$$ [32]

With the positive predictive value derived we can ask what the move from selecting P<0.05 to selecting P<0.005 does to the statistical power of the study.

[32] As presented in Button et al., 2013, p. 366.

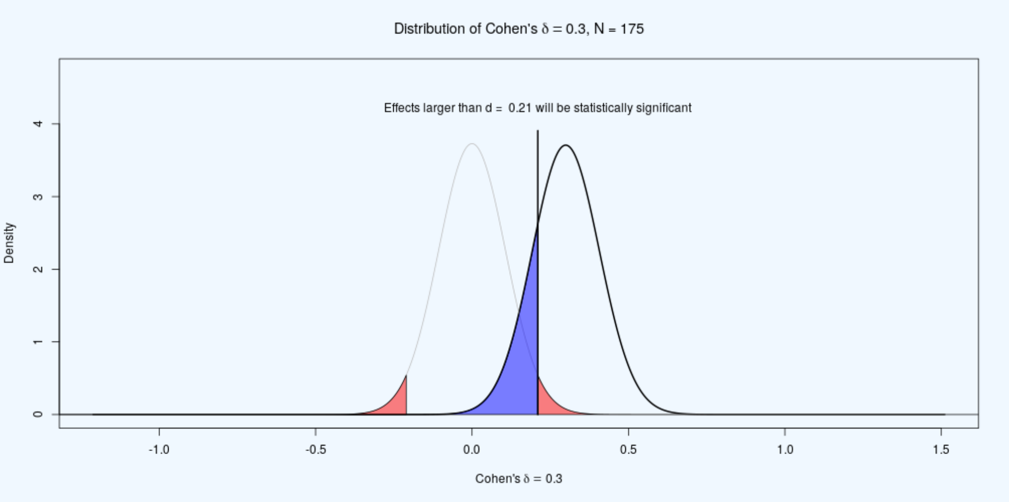

Figure 2.2: Probability distribution for two-tailed t-test, δ=0.3, N=175, P<0.05, 1-β=0.8.

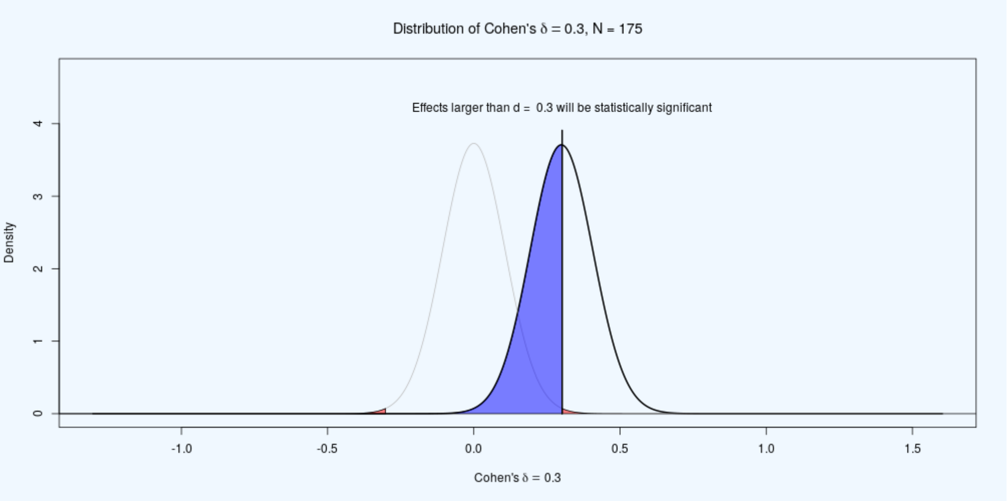

Figure 2.3: Probability distribution for two-tailed t-test, δ=0.3, N=175, P<0.005, 1-β=0.5.

Figures 2.2 and 2.3[33] show the probability distributions of results for a two-tailed t-test under the null and test hypotheses for an effect δ=0.3, N=175, p<0.05 and δ=0.3, N=175, p<0.005 respectively. The vertical lines show the significance level, which lies at δ=0.21 and δ=0.3 respectively. The area under the right-hand curve shaded blue represents the probability of a

[33] Both figures 2.2 and 2.3 are generated using; https://tinyurl.com/3mebkfkd (Lakens, 2020a). Code available at https://github.com/Lakens/p-curves (Lakens, 2020b). Code for similar tools for other statistical tests at; https://github.com/arcaldwell49/Superpower (Caldwell and Lakens, 2020).

false negative. The two red tails under the left-hand curve represents the false positive rate. The power of the study to reject the null hypothesis, given that the null hypothesis is not true and given the effect size of the hypothesis, is the area under the right-hand curve minus the blue shaded false negative area.

Since the positive predictive value is a fraction with the power of the study over the power plus a constant, as power falls we would expect the positive predictive value to fall as power becomes a smaller proportion of the denominator[34]. If we hold the false positive rate fixed, this is indeed what we observe. However, the fall in power is occurring as a result of a more stringent false-positive rate. In the shift from figure 2.2 to figure 2.3 we see a fall in power from 0.8 to 0.5, while the false positive rate falls from 0.05 to 0.005. The fall in false positive rate is substantially larger than the fall in power, meaning that power represents a greater proportion of the denominator [35]. As a result, when the fall in power occurs as a result of a fall in false positive rate the positive predictive value increases. That is, even though power has fallen to obtain a significant result, we can have greater confidence that a positive result indicates a true rejection of the null.

For authors, this problem has a somewhat different face. Since power falls significantly if we hold N fixed under a more stringent α, we need to increase sample size to maintain previous power levels. The exact quantity of this increase depends on the original power level we are seeking to maintain . The reward of doing so is the confidence that authors can have in their own results, and that they can demonstrate readers should have in their results. A well powered study at $P<0.005$ will have a very high PPV unless the prior odds of null being untrue were exceptionally low.

[34] $\mathbb{R}$: $\{A, B\}$; $A \geq 0$; $B > 0$: $x = A/(A+B)$; $x \rightarrow 0$ as $A \rightarrow 0$ and $x \rightarrow 1$ as $A \rightarrow \infty$.

[35] Power falls to 62.5% of its prior value and false positives fall to 10% of the prior rate.

While the P<0.005 criterion is substantially more reliable than P<0.05, the problems it brings in reducing our available dataset makes it difficult to employ in practice for readers interested in asking their own research questions using social psychological studies. For authors, it is wholeheartedly recommended as an indication of reliability of one′s own results in conjunction with recognition of the new power requirement. This may make some studies of small effects prohibitively costly to run, but to run them without these stringent safeguards wastes time and resources on generating results that could not be relied upon.

Overall, the second proposal is useful for some readers of social psychology who are interested in areas or questions where a significant subset of the literature reaches the more stringent p-value significance requirement. In such cases the limitations of the more stringent significance requirement are somewhat ameliorated by the good fortune of having enough material to tackle the question at hand. Aside from this circumstance however, the second proposal is limited by the fact that most authors are not using this stricter significance test and most findings do not reach significance at this level. As a result, it is sufficient for ensuring reliability for both readers and authors but cannot be implemented by readers outside the few areas with large numbers of such p-values.

Section 2.4: PPV/Prior curve

The chief problem facing selection for p<0.005 is the attendant loss of breadth of existing literature on which we may rely, both as readers and as authors. This makes the approach inadequate for those interested in many of the questions posed by social psychology. What makes this approach readily epistemically implementable is the combination of a simple heuristic and the relative availability of the information on which the heuristic be applied. What makes it reliable is the effect that such a selection has on the likelihood of selecting a true positive. This section presents an alternative that retains the reliability of the previous suggestion, which avoids some of the problems of scope, but at the cost of ease of implementation. Rather than use p<0.005 as a proxy for a high PPV, we can instead calculate the PPV, for those results we are interested in relying on, manually.

This approach is intended as a pragmatic tool for those with frequentist training, or working in frequentist fields, who are sensitive to the Bayesian concerns raised in the previous section. As a result, it represents a potentially illuminating tool, but a fundamentally unprincipled one – it uses frequentist analysis while using Bayesian priors for a given experiment (rather than background probabilities for whole fields). Insofar as it is useful for informing us about the reliability of a given result over a range of priors, it serves its purpose. In the long run, even if it is useful, it would need to be superseded by more principled changes in statistical methods and training.

Suppose we are interested in an experiment conducted which showed us that there is a connection between feeling happy (evaluated by some implicit measure) and reporting thinking that we are happy (evaluated by an explicit measure). This experiment was conducted with an estimated power of 0.8 and found a significant effect in favour of rejecting the null hypothesis (p=0.045). For this experiment, we can plot the PPV against a range of priors:

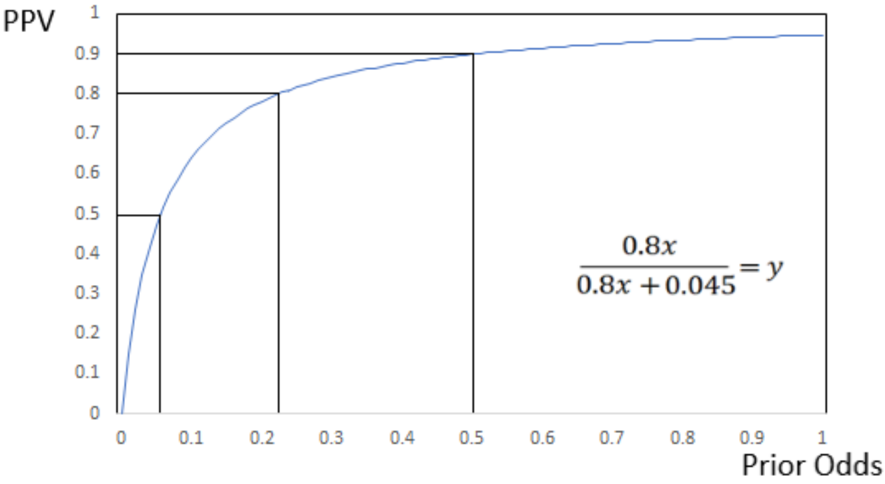

Figure 2.4: Graph and equation relating PPV (y) with prior odds (x) for power 0.8 and p=0.045.

This graph plots the PPV of this experiment against priors on the x axis. For test hypotheses with high prior odds (0.5<x) the PPV exceeds 0.9 and we can have a high degree of certainty that the finding represents a true rejection of the null hypothesis. Such test hypotheses are, however, not those which are asking interesting questions since they seek evidence that we ought to reject a null hypothesis which, on balance, we would reject prior to the experiment. Interesting hypotheses are risky. Focusing on the segment of the curve 0.25<x<0.5, we have a segment where prior risks give us a reason to be relatively confident that the result represents a true rejection of the null (PPV>0.8), but which gives us some warrant to investigate auxiliary support [37] for, or further testing of, the test hypothesis. Third, we have a segment 0.07<x<0.25, where the PPV begins to drop off significantly and in which we should consider ourselves obligated to find auxiliary support before predicating other research or beliefs upon it. Below x<0.07 the PPV drops precipitously. Within this segment of the curve, it is unlikely that the finding represents a true rejection of the null.

With the structure of the finding for a variety of prior odds outlined, we can ask what this exercise does for us as readers and authors in social psychology.

We identified the range of priors across which we would readily believe the significant result to represent a true rejection of the null, PPV>0.9, and those ranges across which we may want some supplementary support for the rejection of the null and for which it becomes necessary, PPV>0.8 and PPV>0.5 respectively. Finally, we have a range across which it is unlikely that the significant result represents a true rejection of the null, PPV<0.5. With

[37] Support originating from some source other than the sheer strength of the evidence offered, which remains independent of our priors, which give us a further reason to reject the null. Such auxiliary support might be the way in which the finding corresponds with other similar findings, unavailable at the time of experimental design, and therefore independent, or they may be dependent on this finding, but independent of the strength of the evidence, such as coherence with a theory in a related but distinct field. Finally, and perhaps most commonly in practice, explanatory elegance may give us a reason to reject a null hypothesis.

these ranges identified we can interrogate our prior and where within these ranges we fall.

In the experiment described above we may decide that our prior for the null that there is no relationship between feeling happy and reporting happiness is that it is trivially false ($P(-H_0)\approx 1$). In which case the finding very likely represents a true rejection of the null, though since the PPV should be used to update future priors we are obliged to become less sure of our commitment to the falsehood of the null than we were before this finding ($P_1(-H_0)\approx 1$; $P_2(-H_0)\approx 0.95$). Alternatively, we may decide that, in general, a person′s understanding of their own emotional state is exceptionally poor and that we should expect these measures to be relatively independent. If we decided that we had a prior of $x\approx 0.25$ of the null being false, we instead should take the result to offer us good reason to think the null is more likely false than not, though we might want further ancillary support for the claim before predicating further research on it. As before, we are obliged to update our priors for future experiments ($P_1(-H_0)\approx 0.25$; $P_2(-H_0)\approx 0.8$).

Suppose the experiment instead purported to find evidence that the rating a child gives to their happiness on their first day in school correlates with their happiness at age 25. We might assign a very low prior, given the sheer breadth and variety of confounding factors and noise, as well as preconceptions we may have about the relative independence of the first measure from the child′s general wellbeing, etc. Suppose this means that we allocate a prior of $x\approx 0.01$. Now the finding represents something which, while statistically significant, is still overwhelmingly more likely to be a false positive than a true rejection of the null ($PPV\approx 0.15$).

This approach deliberately inverts the more usual Bayesian inferential procedure of identifying our priors and only then relating our priors to the probability of the novel data given that prior to arrive at an updated confidence value (Robert, 2001; Carlin and Louis, 2008). Given the challenges currently faced in social psychology, this inversion is intended to force us to ask specific and searching questions about our priors for our hypothesis. It assumes our priors are not readily available for deployment and will require some interrogation as part of their rigorous implementation. Insofar as this is not necessary because of the clarity and availability of our priors, or the procedure is either in practice unimplementable or has unforeseen deleterious consequences, the order can be reversed with little loss to the core utility of the tool.

The application of this tool for readers allows us to identify the key statistics of a paper and derive a simple graphical representation of the strength of evidence it offers. It also asks us to clarify our prior assumptions about the likelihood of the null being true or false and to offer some justification for that judgement. It also highlights those areas in which the judgement is volatile or stable; if we think it is unlikely that the null is false it matters a great deal precisely how unlikely because this has a dramatic effect on the PPV whereas if we think the null is almost certainly false the consequences of being imprecise about how likely this is are far less noticeable for the PPV.

PPV/Prior graphs allow us to interpret a frequentist statistic in a way which is sensitive to Bayesian criticisms and concerns. These criticisms centre around the opaqueness of the underlying assumptions of frequentist approaches (Wagenmakers, 2007; Efron and DiCiccio, 1996) which distort the judgements our evidence legitimates. Bayesian methods relax some of these assumptions and explicitly incorporate them in the analysis. Rather than adopting Bayesian statistical approaches wholesale, PPV/prior curves allow us to remain sensitive to the prior probability of the falsehood of the null hypothesis when interpreting frequentist data. This retains the key advantages of a frequentist analysis (Efron and DiCiccio, 1996) while enabling us to be sensitive to some of these Bayesian criticisms.

As readers we can produce a graph, like the one above, substituting in the relevant values for the PPV equation for the experiment we are interested in. Depending on our requirements of the evidence we then set the PPV levels we consider to be excellent, good, acceptable and unacceptable and determine the ranges of priors that correspond to these levels. With this established we clarify how likely we believe it is that the null hypothesis is false and offer some considerations and motivations for this judgement. We are then in a relatively informed position to judge whether we want to predicate our own beliefs or further research on this claim.

For authors, this approach can be implemented in two ways. The first is in experimental design. If we are doing interesting, novel research on a risky test hypothesis we should presume our prior to be low. Given such a low prior we should design an experiment with sufficient power that we can offer readers a high PPV in the case of a significant finding. This first implementation also complements the adoption of the previous suggestion of $P<0.005$. By using and reaching such a stringent p-value, readers can be assured of a high PPV (0.89) even with priors as low as 1 in 20 ($x=0.05$).

This introduces the second implementation; transparency about PPV. As part of reporting the results the PPV/Prior curve should be published and readers should be allowed to judge for themselves whether the evidence represents a likely true rejection of the null. This is particularly useful for studies which propose risky, unlikely test hypotheses. Those studies which claim to be ground-breaking, novel, radical or their synonyms must present a clear, readily interpreted tool which backs up these claims with the strength of their data.

There are two costs to this approach for readers. The first cost is the relative epistemic demandingness to readers of running their own post-hoc analysis of the data presented in an article. First, the approach necessitates a close reading of the article to establish the p-value we need to investigate. Second, and more strenuously, we need to estimate the power of the study to establish its finding. This is often not directly reported in the study and, if

The second cost of this approach is that, given the rate of surprising findings in social psychology, which obtain marginal significance, with low power, it incorporates a significant rate of exclusion. Unlike the selection for P<0.005, these exclusions are not solely based on the p-value but on a combination of the power of the study, the prior of rejecting the null and the p-value. This allows studies which have higher p-values but which are well designed to test a low-risk hypothesis to remain relevant to future research, while also excluding studies which obtain a very high level of significance for rejecting a trivially false test hypotheses (e.g. Bem, 2011; Bem, Utts and Johnson, 2011). In this way its exclusion cost is mitigated by the greater nuance it offers in its exclusion criteria.

Overall, the PPV/prior curve is more demanding than selecting for p<0.005 but allows for nuanced and transparent interpretation of the strength of evidence we are being offered. For authors, who already have to engage in statistical analysis of their data, much of it more complex than deriving a PPV/prior curve, there is no clear cost of implementing the strategy except to discourage running studies which test incredibly risky hypotheses with insufficient power and relaxed significance criteria. Such studies offer little reason to believe their conclusions even if they reach significance and their loss is arguably no cost at all. It also rewards diligent researchers, testing well-defined hypotheses with well-powered studies with a clear tool for readers to show the weight of evidence even for risky hypotheses.

Section 2.5: The persistence of the problem

By applying the third, or in some cases the second, approach outlined above readers of social psychology can ensure that the statistical trends on which they predicate their conclusions have a good chance of remaining robust and are likely to be replicable. Authors implementing either approach also offer tools for their readers to clearly distinguish their research from less robust research in their field.

In this manner some of the most pernicious problems of the replication crisis can be avoided or mitigated, including the dedication of time and energy by readers into explaining purported phenomena that are unlikely to exist, the waste of time and resources by authors in conducting experiments that are inadequate to test their risky hypotheses, and the loss of confidence by both readers and authors in the field. However, the statement of this problem in epistemic terms highlights a disconnect between what this ameliorative project resolves, and the central problem of social psychology as stated in chapter 1.

Our problem is that social psychological papers do not give us a novel reason to believe their claims. The ameliorative statistical project focuses on the statistical trends in the data which are the grounds of these claims. However, this is only part of the story of the construction of the claims made by social psychology. The raw data is gathered according to an experimental procedure following the categorisation of experimental outcomes and this data is then subjected to a variety of statistical tests and finally the outputs of those statistical tests is interpreted according to the coda used to classify the experimental outcomes. This interpretation is the conclusion of a published article. The conducting of low powered studies which replicate poorly is not unique to social psychology, since it occurs in other areas including neuroscience (Button et al., 2013) and cognitive psychology, but a replication rate of 25% is peculiar to social psychology. I propose that social psychology′s rate of replication is not simply a function of the statistical power of studies but also an indication of the falsehood or at least imprecision in the underlying ontological commitments of the research programme (c.f. Muthukrishna and Henrich, 2019; Devezer et al, 2020).

In social psychology, I propose that the relative volatility of actual replications is due to the codification of the raw data in terms that are ontologically poorly expressed −contradictory, incomplete or imprecise. Attempting to describe statistical trends under such uncertainty is analogous to offering answers to incoherent questions and wondering why our answers do not make sense, giving us null hypotheses that are trivially false or test hypotheses that are so open to interpretation that weak results can be used to reject the null.

Following the Lakatosian framework adopted in the previous chapter and throughout, the ontological commitments of social psychology are the hard core of the research programme. These are the theoretical commitments that will resist contradiction at the expense of the protective belt of auxiliary hypotheses. Research programmes rely on the combination of these types of theoretical commitments to offer explanations of phenomena and make predictions of novel phenomena. In Lakatosian terms, the hard core of the research programme of social psychology has not been examined with sufficient rigour. Due to the heterogeneity of commitments in the research programme, if it were to be stated plainly, it seems unlikely to receive widespread assent by practicing social psychologists. This is in spite of the fact that it forms the basis of the codification of much of their data, as well as experiment design, appropriate statistical practice, and terms for expressing conclusions. From this we may either conclude that social psychology does not actually possess a proper hard core, or we may conclude that the hard core it does possess is internally incoherent or incomplete.

There are several plausible responses to these options. One is to reject that social psychology is or can become a progressive research programme, effectively adopting the first approach outlined in the statistical responses to the replication crisis discussed above. Another is to substitute in a central model as a new hard-core commitment from a related area which is already well-defined and to go forward using this as the foundation of social psychology. It either tests auxiliary hypotheses around this new hard core or offers a competing hard core which it must motivate as a superior alternative. This is the approach adopted by Muthukrishna and Henrich (2019) who propose adopting the dual-inheritance theory in social psychology to address its theoretical shortcomings.

The option section six presents is more explicitly Lakatosian than Muthukrishna and Henrich′s by avoiding offering explicit instruction of the direction of future research or introducing new norms to social psychology. The primary pragmatic advantage of this approach is the openness to social psychologists being the drivers of theoretical advancement in the field. It is hoped that such an approach makes the proposal more palatable and therefore more likely to receive uptake. The aim of this option is to offer a clear framework for presenting and clarifying the theoretical framework that is being relied upon by a given publication. This transparency introduces no new norms but unpacks the justifications of explanations and predictions as targets for normative judgements.

The nature of the problem as one of inconsistent commitments and lack of clarity across the research community makes an exhaustive demonstration of these inconsistencies in social psychology unfeasible. However, in subsequent chapters I use the tools developed in the following section to illustrate the prevalence of the problem in attitude psychology with the aim of both offering specific critiques and illustrating the presence of precisely this problem within social psychology more generally.

Section 2.6: Ontological critique

The ontological critique that I propose consists in a sequence of challenges to clarify the ontological status of claims. It is intended to be run in conjunction with the ameliorative project outlined above to improve the reliability of statistical trends. The questions are proposed generally but are followed by comments on the application of the critique for readers, authors and editors.

The first challenge is not actually ontological, but axiological:

1) *The axiology challenge: What do we care about?*

1a: In the research programme in general?

1b: In this project in particular?

This challenge specifically calls for an interrogation and statement of the motivations which are legitimate research concerns which may be appealed to in answer to criticism. A full and frank answer to this challenge will exhaustively identify the dimensions in which a finding may be exemplary, robust, interesting, trivial, or frail. They may include concerns such as accuracy, precision, coherence with other areas of research, simplicity, elegance of explanation, coherence with a broader political or social concern, etc. There is no licit or illicit list of answers to this challenge, simply those answers to which the author or reader believe legitimate recourse may be made.

Answers to these questions are often assumed to be straightforward by philosophers, especially philosophers of science, but the clear explicit statement of these motivations is often surprising. For example, Bem famously responded in interview "I'm all for rigor, but I prefer other people do it. I see its importance—it's fun for some people—but I don't have the patience for it. If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, 'Will this replicate or will this not?'" (Engber, 2017). This is not a 'bad' or 'wrong' answer from Bem, but had this been clearly stated as part of the original article, readers would have been clear about the kind of engagement they were involved in. When reading Bem's articles we are not observers of an inquiry into a research question but the target of a persuasion attempt making use of a convenient statistic. Knowing this, we respond to the information we are presented very differently.

Social psychology sits at the crux of many important questions and the interrogation of what is being asked, and what the licit modes of answer are, is necessary for open enquiry into the importance and reliability of the answers generated.

2) *The objects challenge: What are the proposed objects?*

When an explanation is being offered, either of the interpretation of data or of the categorisation of experimental outcomes, we may ask what the objects being posited are. What needs to exist such that the offered explanation is literally true, at that level of description ? This may be a neuron, an empathic connection between participants, the individuals themselves, tendencies to ϕ, situations, beliefs, desires, the anomalous retroactive influence of future events on current cognition, etc. Whatever is purported to exist is a relevant and important answer and a full answer to this challenge will exclude no objects that play any role in subsequent explanations or predictions.

Once we have a full enumeration of the relevant objects we may begin asking questions about the justification for inclusions or exclusions to that list as well as challenging distinctions or dichotomies that are proposed therein. This allows us to challenge the categorisation of datapoints or experimental design in terms of the effect it will have on the construction of the dataset as a whole.

3) *The Properties challenge: What are the proposed properties of these objects?*

With the enumeration of the objects themselves, we may begin interrogating their purported properties. How do these objects behave? Under what conditions? What are they capable of and incapable of? A full answer to this challenge will enumerate all those properties of the objects outlined in the previous challenge that are necessary for the explanations offered by the model.

4) *The Explanation challenge: How does the conjunction of the objects and their properties give rise to the observed evidence?*

This fourth question brings together the previous two and allows us to ask what role the ontological commitments the author is explicitly committing to play in the explanations being offered. At this point, gaps and inconsistencies, where there are any, begin to become apparent. Those explanations offered by the research are stated in terms of the objects and their properties with no inclusions that were not part of the answer to the previous challenges.

The fourth challenge highlights the features of the explanation when the effect is elicited but it asks about all of the observed evidence, not simply the successes. What is happening when the effect does not occur? Does the list of proposed objects offer us tools to explain failures? What are the confounding factors and what power do they have to confound? Do we have a complete list prior to the results or are we offering explanations, or positing the means of explanations, *post hoc*?

5) *The Falsification challenge: What would it take to falsify these explanations?*

5a: What would have to occur for the explanation to be falsified?

5b: What would be falsified if this occurred?

The fifth challenge takes the clarification from question 4 and asks it for counterfactual cases in which the explanation would be falsified. This utilises the clarification of the confounding factors identified by the explanation challenge and asks what it would take for a result to be more than merely confounded by some factor, but in fact falsifying of the explanation offered. It then further asks what, in such cases, is being falsified?

This question has three heuristic uses. The first is to identify when a theory has built in so many, or such powerful, confounding factors that it becomes immune to falsification.

Suppose we have a model of understanding which describes the time taken for an individual to comprehend a written word in terms of other observable factors. Built into this model is a stochastic element which may vary by up to 50% of the value of the whole function. When our model, excluding the stochastic element, predicts that a subject will take 6.2 seconds to register understanding of a written word and she in fact takes 3.9 we may either see this as potentially problematic for the model, or put the divergence down to the stochastic element in the model. This model would be almost impossible to falsify because its stochastic element is such a powerful confounding factor. Identification of cases where this has occurred helps prevent stagnation in a research programme as researchers only find confirmatory evidence of an unfalsifiable explanation. Once identified, this allows for heuristic progression within the programme as these elements are reduced in power and new, clearly defined elements are incorporated in an attempt to explain what was previously excluded.

The second heuristic use of this question is to find the conditions under which the explanation might be falsified in order that researchers might try to bring this about. By clarifying the conditions of explanation failure our ability to pursue crises or find that we are unable to bring them about, can be actively encouraged.

The third heuristic use of the challenge is to identify the non-novel effects which the explanation entails that we should expect to find. When we explain an interaction between a drug and a cancer cell in a petri dish, we are committed to claims about future experiments which test that quantity of that drug on that kind of cancer cell in a petri dish, according to the measurement procedure used. Given the mechanics of our explanation we might posit that the class of treatment procedures T_1 will elicit the effect, but that T_2 will not, while on T_3 we have no reason to expect either way on the basis of this explanation. Even more simply, our explanation will banally entail the prediction that the same effect will tend to be elicited by the same treatment under the same setting for the same measurement procedures.

These banal predictions are as important in a research programme in crisis as the radical, novel predictions. Novel predictions are important because they are risky. In a research programme in crisis, even banal predictions are risky. Predicting that a lab in Argentina will replicate a priming result obtained in Massachusetts – even if it is strictly entailed by the explanations we have offered for the effect in Massachusetts – remains risky in social psychology.

In the terms set out in the previous chapter, this challenge demands the statement of the generalisations entailed by the form of the explanations we are offering for a phenomenon, as well as those generalisations we are making which are adopted on the basis of something less strict than entailment. When we explain a phenomenon as following from an essential property of the population of experimental units – e.g., humans are, and conceive of each other as, goal directed animals (Myowa-Yamakoshi, Scola and Hirata, 2012) – it is entailed by our explanation that we expect the effect to generalise across that population, in this case all humans. When we explain a phenomenon as associated with properties related to, though not essential to, a population of experimental units – e.g., career-focussed students tend to find the transition to higher education easier than those without clear career trajectories, because that focus often helps when higher education presents setbacks and obstacles (Hassel and Ridout, 2018). Here we have theoretical reason to generalise across the population, but the generalisation is not entailed by our theoretical framework because of the recognition of a mediated relation between the operator and the population of experimental units. This is important for interpreting subsequent replications and for designing those replications.

If these effects do not replicate upon resampling, especially if in general this does not occur, then our explanation is false. This ensures that the population extension of any given explanation is explicitly invoked and can be tested, not only for novel predictions but for banal predictions which may turn out to be false.

This heuristic use has two edges which cut in different directions. This is the first – the clarification of how our explanations extend and the populations we take ourselves to be sampling from and generalising to. The second is that the awareness of this extension, and its explicit statement, encourages us to be more modest and precise about our claims. Rather than offering an explanation of ′priming effects for humans′ we might instead offer an explanation of ′priming effects of culturally specific stress factors on response timings for undergraduate students′. We are still generalising, and can still be resampled and held to account, but we are only generalising to undergraduate students and a population of culturally specific stress factors.

The second part of the fifth challenge asks which elements of the explanation do we maintain our commitment to in the event of a falsification, and which bear the brunt of the falsification? This pushes us to identify the hard core/auxiliary hypothesis distinction for a particular, imagined, falsification of the explanation. This includes though extensions which are warranted by the explanation, which fail, and replications which resample one of the relevant populations and fail to replicate the effect. By highlighting what will and will not be jettisoned in the event of a falsification, the overall commitments of the research programme and their status is gradually established, not in a single report but over continuous utilisation of the approach within a research programme.

With the sequence of the ontological critique outlined and an ideal answer to each defined, we may ask what this means for readers and authors.

Readers should clarify for themselves how the author would respond to each of the challenges. To be clear, readers are enjoined to engage in a rational reconstruction of the foundations of the paper, not biography of the author. For the most part, this is an imaginative and reconstructive exercise that fills in the blanks left in most social psychology papers. Doing so is elucidating because of the extent to which it clarifies what is at stake and what is purporting to be demonstrated therein. This should be combined with some aspect of the statistical ameliorative project outlined above in order to be sure that the results upon which that we predicate our own investigations and beliefs are both clear and reliable.

For readers, the ontological critique acts as both a reading tool and a negative heuristic. In cases where it is unclear what is being claimed or defended readers may either disregard the paper as providing valuable evidence or may offer a reconstruction that fixes the apparent problems. Many uses of social psychology literature by philosophers engages in this ontological reconstruction and clarification, sometimes explicitly, sometimes not (e.g., Webber, 2016; Machery, 2016).

Authors also ought to implement the ontological critique as an accompaniment to a standard publication as well as consideration during experimental design. The form, appendix, offers an example of how the critique might be implemented by authors and editors to clarify the status of claims being made within the publication. This makes it available for reference by readers, cross-reference during meta-analysis, and as an additional tool for pre-registration. By carefully stating the explanations the experiment will offer of a variety of outcomes it becomes possible to identify where confounding factors are built into an experimental design and allows for precise feedback to authors.

Unlike for readers the stage at which authors engage with the ontological critique′s challenges matters. If the critique is engaged with during experimental design, the challenges allow us to clarify what we will be investigating and what the data can and cannot tell us. It clarifies the explanations available to us and what the data will tell us if it arrives in one expected form or another. It also allows us to identify potential challenges or flaws during preregistration when this is possible. If the critique is engaged with following data-gathering, then the challenges amount to a rational reconstruction. While this is similar to what readers are enjoined to do, the incentive structure for researchers is substantially different and presents challenges in the implementation of the challenges in a clear and open manner. As such, if the challenges are not engaged with during experiment design, this should be declared explicitly so that readers may evaluate the claims through an appropriately critical lens.

Section 2.7: Concluding remarks

In summary, readers who wish to continue engaging with social psychology should take steps to clarify and double-check the statistical evidence for the purported phenomena. Where it is possible to implement the second option of treating as indicative, rather than significant, claims based on 0.05>p>0.005 this is a low-cost strategy which offers good reliability. Where this is not possible, or where more nuanced interpretation of results is needed, the third option of implementing PPV/Prior curve to the data should be used. Finally, once the robustness of the statistical trend is clarified, the ontological status of the claims it makes and the warrant for those claims must be interrogated by implementing the challenges of the ontological critique.

Authors should implement the more appropriate significance level of P<0.005, treating results of 0.05>p>0.005 as indicative but not significant. Furthermore, as part of their publication they should include a PPV/Prior curve which illustrates the likelihood that the presented result represents a true rejection of the null across a range of priors. This increases the reliability of the data itself and gives readers a clear tool to interpret the evidence given the novelty or unexpectedness of the claims being made.

This programme improves the epistemic standing of the claims themselves by ensuring that a positive result is more likely to in fact be a true result and ensuring that this is clearly communicated to readers. Furthermore, by improving the quality of the available evidence we are seeking to explain, as well as offering ever more precise analyses of the ontological status of the explanations we offer, we may hope and expect social psychology to progress both heuristically and evidentially as a research programme. We may reasonably hope that such practices will clarify and construct a clearer and more robust hard core of theoretical commitments.

Finally, by clarifying the broad category of the theoretical commitments being made and the justifications for the explanations offered we are able to outline how our theoretical commitments extend our explanations to particular populations. In particular, the explanation and prediction challenges identify the construction of the explanations and the populations over which we expect these results to remain robust. Even when predictions cannot be novel, they can explicitly identify populations over which effects are predicted to reliably extend – this includes populations of experimental units, treatments, measurements and settings. This explicitly calls for the clarification of what the explanation takes to be a replication by resampling the relevant populations, and what would be a theory-motivated extension. By engaging with the critique, we might hope to offer ourselves and our readers clarity about the generalisations made from data to explanation and the theoretical justifications for doing so.

Chapter Three: Deciding Against Desiderata

Section 3.0: Introduction

Contemporary research in social psychology has emphasised the development and analysis of the phenomena of implicit cognition. This analysis is of concern to many areas of philosophy, most extensively in the literature on virtue theory, through identifying non-rational elements in decision-making processes (Rees, 2016). How such a challenge should be presented, what it represents, and how it may be met are currently contested in the literature. Answering these questions requires a precise conception of implicit cognition.

Several models of the phenomena have been proposed including the description of implicit biases as patchy endorsements (Levy, 2015), the contrast of implicit biases with beliefs resulting in their description as ′aliefs′ (Gendler, 2008), and the description of attitudes in general as traits with implicit measures identifying parts of whole cognitions (Machery, 2016). Holroyd (2016) and Holroyd, Scaife and Stafford (2017) contribute to this debate by proposing desiderata for a successful account of implicit cognition.

I argue that while the application of desiderata may resolve some instances of model choice in implicit cognition, problem cases are both extant in the literature and are relevant to current model choice. As a result, the desiderata approach fails to meet its purported aims and as a result the ontological critique is to be preferred, if it proves fruitful in gaining traction on these problem cases.

Section 3.1 challenges the distinctness of the above desiderata and, in doing so, reframes the analysis of models of implicit bias offered by Holroyd et al. in terms of the model′s relevant explanatory and predictive power. Section 3.2 presents a theoretical problem for analysis of models through desiderata, where any such analysis will be insufficient to settle key cases. Section 3.3 introduces the first, more minor version of such a problem from the contemporary literature. Section 3.4 introduces a more radical and insurmountable version of this problem highlighted within, though not by, Holroyd (2016; Holroyd, Scaife and Stafford, 2017). Section 3.5 offers reasons to reject three prima facie plausible responses to the problem of desiderata. Section 3.6 clarifies why undecidability remains a problem for the desiderata approach even if we are radically pluralist about our models. Finally, the section 3.7 presents the proposal of the ontological critique as a specific response to the problem faced by the desiderata approach. The chapter concludes in section 3.8 that if the ontological critique can be shown to be fruitful in addressing the problem cases discussed it is to be preferred to the desiderata approach as a tool for evaluating models in social psychology.

Section 3.1: Reducing the desiderata

Holroyd, Scaife and Stafford (2017) identify the following desiderata for a successful account of implicit cognition:

"D1: To distinguish implicit from explicit mental states or processes;

D2: to capture interesting cases of dissonance between agents' professed values and the cognitions driving responses to these measures;

D3: to formulate interventions for changing bias, or blocking discriminatory outcomes;

D4: to accommodate or explain the full range of the phenomena captured by indirect measures; and

D5: to gain traction in addressing problems of marginalisation and under-representation, and draw attention to complicity in these problems." (Holroyd, Scaife and Stafford, 2017, P. 3)

In assessing desiderata, we may ask what would constitute excelling in each regard.

To meet desideratum 1 a model must offer the means of making a distinction in the entities or properties proposed, or an explanation of how the appearance of such a distinction may emerge in spite of no such distinction in the entities or properties proposed. Furthermore, it must offer this explanation in a manner that fits the actual distinction in the evidence.

The second desideratum states that a good model will capture a particular subset of the evidence; namely those cases where the agent's professed values and the implicit measures diverge. In order that a model capture the 'interesting cases' it is necessary that it has either a means of making the distinction between the explicit and implicit or an explanation of the appearance of such a distinction, as well as a reason for the distinction becoming apparent in those key cases. By elaborating what would constitute meeting the second desideratum it becomes clear that it is in fact a restatement of the first desideratum of explanation for a subset of the same evidence. In the former, a good model will offer a means of making the distinction and must fit the actual appearance of the distinction in the evidence 'overall'. In the latter, a good model will offer a means of making a distinction and must fit the actual appearance of the distinction in the evidence, qua interesting cases.

The third desideratum states that a good model of implicit cognition will formulate interventions for changing bias, or blocking discriminatory outcomes. In order to meet this desideratum a model must meet the first desideratum by offering the means of a distinction and accurately identifying the distinction in the evidence. Furthermore, it must do so in a manner that predicts future occurrences of the distinction. These predictions necessitate a description of how the divergences occur such that interventions can be formulated to prevent or sufficiently alter the outcomes.

A model meeting the third desideratum goes beyond mere explanation: predictions are also necessary. Moreover, these predictions should specifically elucidate how interventions in the process that instantiates the divergence can alter the outcomes in the interesting and ethically relevant cases. This desideratum establishes the need for accurate predictions and states what would constitute making such a prediction: identifying when a phenomenon will occur and identifying the mechanism with sufficient precision that an experimenter can control the outcome.

The fourth desideratum states that a good model of implicit cognition will also accommodate or explain the full range of phenomena captured by indirect measures. For a model to meet this desideratum it must offer an explanation of the so-called ʹinteresting casesʹ where a divergence occurs between the implicit and explicit measures, and the more common case where it does not. As such, any model that meets the first desideratum necessarily meets the fourth and to the extent that the fourth desideratum is not met, neither is the first.

The fifth desideratum states that a good model of implicit cognition will allow us to gain traction in addressing problems of marginalisation and under -representation, while attending to ethical and epistemic complicity in this regard. The latter part of this desideratum emphasises a dimension of the evidence that requires explanation – its relevant ethical and epistemic properties. The latter calls for accurate predictions and interventions following the third desideratum. The fifth desideratum combines the explanatory project of the first desideratum with the predictive project of the third.

By laying out exactly what would constitute meeting the desiderata, we may now group them into two sets. The first includes desiderata one, two, four and the latter part of five. The second, desiderata three and the former part of five. The first group represents those desiderata which are achieved by models which explain the evidence which is already available, with varying emphases. The second group represents those desiderata which are achieved by models which provide predictions of salient cases, with sufficient precision to that we may formulate interventions.

Within the first group, any model which successfully meets the first desideratum, has necessarily met both the second and fourth desiderata. T he meeting of the first desiderata requires the explanation of a dataset, while desiderata two and four require the explanation of subsets of that dataset. Similarly, for the second group, any model that has formulated interventions for blocking discriminatory outcomes has also, necessarily, gained traction in addressing those same problems of discrimination and marginalisation.

I propose that each of these groups be reduced to a single desideratum. The first group reduces to the desideratum A:

A. *The explanation of the relevant available evidence through distinguishing kinds of mental states or processes, in accordance with the implicit and explicit psychological evidence.*

Similarly, the second group reduces to the desideratum B:

B. *The successful prediction of phenomena through addressing epistemic and ethical problems surrounding marginalisation and under-representation, in part by drawing attention to relevant complicity.*

This reduction concludes that what we want from a model of implicit cognition is that it meets these desiderata – a good model will meet both *A* and *B* to a greater extent than a less good model.

One response that could be made to such a reduction is that the five desiderata are not simply normative guides to a good model, but are worth keeping for the practical benefits their emphasis affords. For example, the presence of a desideratum that emphasises the need to explain the full range of phenomena associated with implicit measures motivates psychologists and philosophers interested in modelling such phenomena to offer a model that explains why the 'interesting cases' of divergence are also significantly in the minority (Holroyd, 2016). As a result, we ought to maintain the full complement of desiderata in order to preserve their ability to guide model-making of implicit cognition.

I offer two counterarguments to this response, one direct and one indirect.

The direct response is that we need to independently justify the emphases we offer in additional desiderata. There is, in principle, a non-finite set of desiderata which would place varying emphases on different aspects of the available data, or on predictions we might make, which would be reducible to A and B. Suppose it to be the case that if everything is emphasised, then nothing is emphasised. If our aim is to emphasise some particular subset of that non-finite set of reducible desiderata, then it is necessary that we include some part of that set and exclude other parts. This places a requirement on such a response that they must also motivate the particular emphases given by the further desiderata over and above the alternate possible emphases that other reducible desiderata would offer. In the absence of such reasons for desiderata-choice any emphasis is insufficiently motivated.

The second response is indirect, in that it does not respond that we ought not to emphasise but that there is an advantage to maintaining only the two desiderata A and B. This advantage is one of coherence with other areas of the philosophy of science. Lakatos [38] utilises and refines the concepts of explanatory and predictive power. These two properties are the yardsticks for the reckoning of some research programmes as science and other as pseudoscience. While the question of "is this research programme science?" may not be relevant here, what remains pertinent is our goal when engaged in the improvement of the research programme. To this end, we seek to improve the explanatory and predictive power of the models and theories of the research programme.

When deciding between two models, their ability to distinguish mental states or processes in accordance with the available evidence describes the relevant explanatory power of the models - how well the model synchronically fits the data. When explanatory power does not decide the matter, either practically or in principle, we utilise predictive power – what novel predictions may be offered by each model which turn out to be true. This is the diachronic instantiation of the model's fit with the data. The relevant predictive power for the project of a philosophical understanding of implicit cognition is articulated by their ability to address the epistemic and ethical problems surrounding marginalisation and underrepresentation. As

[38] While responses to Lakatos' conception of science and the methodology of scientific research programmes exist, these largely focus on the role of rationality or realism within this conception. Further discussion of the larger 'Lakatos problem' is undertaken in section 7.7.

such, this reduction identifies the relevant explanatory power with meeting desideratum (A) and the relevant predictive power with meeting desideratum (B).

The question of what we want from a model of implicit cognition is a specific form of our more general aim of the assessment of competing models and theories in the sciences. As such, the advantage offered by reducing the desiderata is that our answers to specific problems can be directly informed by the substantial existing literature on the more general problem.

Section 3.2: The problem of desiderata

In those cases where theory choice partially determines the relevant dataset, the desideratum A will generate a vicious cycle of dependency which will not yield a single stable conclusion as to the quality of the model. In comparisons between models which partially determine different relevant datasets there will not necessarily be a stable conclusion as to which has greater explanatory power. Furthermore, this problem will not necessarily be practically resolvable by the application of desideratum B. The models may not necessarily be clearly distinguished in terms of predictive power. As a result, cases may occur of comparisons between theories which cannot be decided either by the application of desideratum A, in principle, or the application of desideratum B, in practice. As such, the problem of desiderata may be stated that: the desiderata are insufficient for model choice in implicit cognition.

The substantiation of this claim will be the focus of this and the following two sections. In this section, the outline of the problem is introduced and the theoretical argument for such a problem is given. In sections three and four, two cases of this problem are introduced. The first case illustrates a significant but not insurmountable version of this problem. The second illustrates one which is insurmountable with the tools so far addressed available. In this way the problem is shown to occur in the literature, even in its most vociferous form, while distinguishing this stronger form from the more common, but still problematic, surmountable form.

The theoretical case for such a problem begins with the application of desideratum A.

A. *The explanation of the relevant available evidence through distinguishing kinds of mental states or processes, in accordance with the implicit and explicit psychological evidence.*

This desideratum describes a relation of explanation obtaining between the target data and the model. Explanation is a one-directional relation of fit which obtains when, and to the extent that, both the parts and whole of the target data are elucidated and accounted for by the model. The existence of this relation is indicated by the presence in the model of the means to distinguish kinds of mental states or processes in a manner which aligns with the distinctions present in the target data. To apply the desideratum to a model, we must first identify the relata: the model being evaluated and the data of the field and phenomena the model is intended to explain. Then we need to identify the features of the model that distinguish kinds of mental states or processes, identify the distinctions in the data, and evaluate the correspondence between the two.

This evaluation of correspondence, while informative in itself, is primarily useful as a means to compare and contrast two different models. In doing so we identify the model which presents the greatest correspondence with the distinctions observed in the phenomena. All else being equal, this gives us reason to use the model with the greatest degree of correspondence.

Suppose we want to compare two models of a given set of target phenomena. The first model offers an explanation such that every member of that set is well explained. According to desideratum A, this first model maximally explains the relevant evidence, which is constituted by the set of target phenomena. The second model offers an explanation of each member of a subset of this set of target phenomena, which is large enough that it includes a preponderance of the total set. In addition to offering explanations of each member of this subset, the second model also offers us reason to believe that those members of the total set which are not members of the subset are also not relevant evidence in need of explanation – they are measurement errors, statistical flukes, the results of conceptual confusions, etc. On grounds of desideratum A, which model ought we prefer?

One way to conceive of this distinction is that of the difference between explaining a data point and explaining it away. In the former case the data point is regarded as representing the target phenomena in some important sense. As such, in order that a model may be considered successful it is necessary that it be able to explain the data. In the latter case the data point is not regarded as representing the target phenomena, whether through experimental noise, error, outlier, influences external to the target phenomena, etc. Because of this, not only does the model have no onus to offer an explanation of this data point, but the explanation of this data point as part of the target phenomena would counts against the model.

For example, if a planetary model explains why a given planet is in a particular location in terms of its relation to other planets and plots its projected orbit as part of the explanation of other planets′ orbits, this is a strong explanation. If the observed planet turns out to be an imperfection in the lens of the telescope being used to observe said planets, then the fact that the model explained the experimental error as a data point (a non-existent planet′s orbit) as part of its explanation of other parts of the data set (the orbits of the other planets) shows, at minimum, an imprecision in the model and perhaps a more fundamental problem for the model.

It might seem that the first model is the superior explanation since it explains more of the data. That would indeed be the right conclusion, were it not for the fact that the second model rules as inadmissible precisely that part of the data set explained by the first but not by the second. That is, the second model does not fail to explain that part of the data set: it entails that a model should not try to explain it as relevant evidence.

If the first model is, as it were, in fact the better model, then it is a failing of the second model that it does not offer an explanation of the excluded parts of the dataset. If the second model is, in fact, the better model then it is a failing of the first model that it does offer explanations of those parts of the dataset. In the absence of the ability to know which of these is the case, the desideratum treats the two cases as equal and incommensurable.

In a case like this, if the first model explains vastly more data than the second, or if the second explains hardly any data at all, then it would seem plausible to prefer the first even though the second rules out the additional data that the first explains[39]. But in a case where the two models are relevantly close in scope of explanatory value, where what the second explains represents most of what the first explains, it is not at all clear which we should prefer.

Four potential resolutions present themselves. The first is that problem cases are resolvable in terms of their predictive power, I.e., when desideratum A does not offer a neat resolution, desideratum B will. The second is that we change the desideratum such that it becomes sensitive to the relationship between the available data and the relevant data as well as to the concordance between the relevant data and the model. The third is that we introduce a new desideratum that is sensitive to the scope of the dataset that a model explains. The fourth is that we expand what we consider 'explain' to mean, such that, by giving us reasons to exclude it from the set of relevant data, our second model has offered an 'explanation' of the excluded data.

[39] This sensitivity to scope is necessary to preclude possible model which entail that all, or nearly all, data-points ought to be explained away rather than explained. Conspiracy theories around climate change appear to be examples of this kind; the data-points are fraudulent fabrications of a nefarious cabal of self-serving scientists and therefore ought not to be explained by a model of human-driven climate change but rather explained away. Scope concerns are, however, only adequate in extreme fringe cases. Once one model explains a substantial subset of the data explained by the other, scope is inadequate to decide between the two models.

Section 3.3: The good case

The first case of the problem emerges from the divergence between how Tanesini (forthcoming) and Webber (2016) understand the concept of attitude strength within the context of attitude psychology. Tanesini offers a philosophically rigorous version of Haddock and Maio′s model of attitudes, while Webber offers an interpretation of Mischel and Shoda′s Cognitive-Affective Personality System. Both engage with the available psychological evidence on behavioural patterns and present their respective models as explanations of the phenomena that the evidence represents. As part of this engagement, both engage with the phenomenon broadly referred to as ′attitude strength′.

Following many interpretations by psychologists working in the field (Fazio, 1990; Brannon et al., 2007; Petrocelli et al., 2007; Clarkson et al., 2009; Zunick et al., 2017; *inter alia*), Tanesini analyses the phenomenon of attitude strength as constituted by four independent but complementary variables: accessibility, extremity, centrality, and certainty.

Strength as *accessibility* is defined as the potential for an attitude to influence behaviour. This type of strength is a property of the associative link between the representation of the object of the attitude and the valence of the attitude. The stronger the association is, the more likely it is that anything that triggers the representation of the object will also trigger the valence (Tanesini, forthcoming, P. 10., Fazio et al., 1986). The greater the likelihood of the valence being accessed and influencing the response to the stimulus, the greater the attitude′s accessibility.

Strength as *attitude extremity* refers to the severity of the valence of the attitude. An attitude that is one of mild dislike has a lower extremity than one of visceral loathing. This strength is evidenced in the kind of reaction to the stimulation of the representation of the object of the attitude. More extreme attitudes are evidenced in more pronounced responses to the stimulus. The greater the severity of the valence of an attitude, the greater the attitude′s extremity.

Strength as *centrality to self-conception* refers to how some attitudes are more closely tied to a person′s self-conception than others. This changes how such attitudes influence behaviour and how they may be changed over time. An attitude that is entirely central to someone′s self-conception will be difficult to change and subject to reinforcement mechanisms. Furthermore, the kinds of responses a high-centrality attitude evinces are typified by their association with self-evaluation: defensiveness, pride, confidence, etc.

Strength as *certainty* is the degree of commitment the individual has to their attitude. In this this degree of commitment determines some of the attitude′s interaction properties, such as the role it forms in knowledge acquisition and how it is changed. Tanesini distinguishes between two notions of attitude certainty; ″The first is clarity which measures the subject′s certainty that a statement expresses her attitude. The second is correctness that refers to the subject′s certainty that her attitude is accurate or correct (Petrocelli et al., 2007). Attitude certainty as correctness is opposed to feelings of doubt about the rightness or truth of one′s attitude.″ (Tanesini, forthcoming, P.10)

These four dimensions of strength each explain different aspects of why and how particular attitudes give rise to the kinds of behaviours and how reliably. The aggregation of these dimensions of strength are explicitly described by Tanesini as ″an aggregate measure of several distinct factors″ (Tanesini, forthcoming, P.10).

Webber, by contrast, defines attitude strength as ″the degree to which the attitude is embedded in your cognitive system. An attitude that is strong in this sense is not easily changed by persuasion or reconsideration″ (Webber, 2016). This concept of strength is well illustrated by the Greenpeace experiment (Holland et al. 2002). Participants were asked, as part of a lengthy questionnaire, their attitudes towards Greenpeace. They were also asked how certain they were in their attitude, how important this attitude was to them, how it related to their self-image, and whether it was tied to values they considered important. The first question elicited their attitude towards Greenpeace. The latter four determined its strength. Participants returned a week later for what they were told was an unrelated experiment for which they were paid entirely in coins. They were then offered the opportunity to donate some of their payment to Greenpeace and then asked to fill out a short questionnaire on Greenpeace.

For those participants who held strong attitudes towards Greenpeace, as measured by the four latter questions on the original survey, their attitude towards Greenpeace, from the first question on the original survey, was found to be a good predictor of their eventual donation and responses to the final survey. Whereas the original responses of those participants who did not hold strong attitudes were not good predictors either of donation or of non-donation or of answers to the final survey. From these findings it was concluded that strong attitudes are consistently manifested in evaluative judgements whereas weak attitudes are not.

Where Tanesini′s view holds that the four measures track different dimensions of strength, each a property of the attitude in its own right. Webber holds that they track the same property, which is best triangulated by the aggregation of these measures.

Suppose we wanted to compare these two explanations of this key phenomenon in attitude psychology and evaluate their explanatory power. How could we do so?

Both Tanesini and Webber offer the means to make distinctions between kinds of mental states, namely stronger and weaker mental states, in accordance with the usage of strength in the psychological literature. Furthermore, the dimensions Tanesini identifies are directly led by the experimental literature. Webber′s definition of strength, however, is also compatible with this experimental literature, by distinguishing the dimensions of strength that Tanesini identifies as emergent phenomena of a single property.

The second part of the desideratum is the comparison between the distinctions being drawn in the model and those found in the evidence. Given that the distinctions match one another very closely, the models offer very similar sets of explanations of the effects of strength on attitudes. However, the models differ in how they address cases of the explanation of experiments that focus on a single measure, such as centrality. For Tanesini, centrality tracks a kind of response, while for Webber it is only one possible manifestation of a strong attitude that it results in the right kind of reaction. Such experiments are explained by Tanesini as part of the relevant evidence about the role of strength in behaviour. However, Webber's explanation is that such an experiment is failing to control for the key component of the phenomenon, there will be cases the experiment misses where the attitude was strong yet not extreme while there will also be cases where the attitude was extreme but not very strong. As such, they are not part of the relevant evidence of attitude strength for Webber.

This makes the divergence in the two explanations a case of the desiderata problem introduced above. Purely on the basis of the available evidence it is not clear which of the two models we should prefer, not because they both offer explanations of all the relevant evidence, but because the sets of evidence they explain as relevant evidence are non-identical.

However, it is a relatively minor version of the problem. This is because experiments can, in principle, be formulated that directly test the two theories. Each theory makes a prediction about the connection between the four measures of strength and how we might expect changes in these outcomes to track one another as well as how these changes should become apparent in the behavioural outcomes of the individual. In particular, if

Section 3.4: The bad case

While many cases of the problem of desiderata are cases of the above kind, where the dispute may, in principle, be settled by the application of desideratum B. There remain cases which cannot be resolved by the simple application of desideratum B. This section outlines one such case: Machery′s (2016) trait theory of attitudes. In doing so I demonstrate the inadequacy of the application of desiderata A and B for its assessment.

Machery (2016) begins with the question ʺwhat kind of things are attitudes?ʺ (2016, p. 104) offering an account which he calls the trait picture. This is the claim that attitudes are traits:

ʺA trait is a disposition to perceive, attend, cognise, and behave in a particular way in a range of social and non-social situations. Within a species, there are individual differences with respect to a particular trait; some organisms have more of it, others less. This variation can be measured, and it is predictive of their behaviour and cognition.ʺ (2016, p. 111)

Machery then differentiates between broad-track and narrow-track dispositions. Because of their influence on a wide range of situations and their manifestation in behaviour, cognition and affective responses, traits are broad-track dispositions. Machery argues that attitudes have a psychological basis, which comprises ʺa motley assortment of mental states and processesʺ (2016, p. 112). But, according to Machery, an attitude is not identical with any part of its psychological basis, or with that basis as a whole. Rather, the language of attitudes, like the language of traits, describes only abstract formal structures that can be realised by such a psychological basis.

Given this picture of attitudes as traits, we may ask what implicit measures are detecting? Machery argues that under this description of attitudes, we ought to understand different explicit and implicit measures to be detecting different artefacts of the psychological basis of attitudes: ʺFor instance, the affect misattribution procedure and the implicit association test may respectively measure automatic emotional reactions and concept associations.″ (2016, p. 113) Thus, Machery can explain how several measurements designed to observe the same attitude can give different, even opposing, results. By tracking different underlying components of the target attitude, they can appear to contradict, when in fact neither is tracking the attitude itself. Each are measuring the direction of rotation of a single cog in the engine and, from that, inferring the direction of the locomotive's travel.

Setting aside a more comprehensive engagement with Machery′s argument and model for chapter 4, how does the trait picture fit with the desiderata introduced above?

A model meets desideratum A by offering the means of making a distinction between kinds of mental states or processes insofar as this distinction accords with the available psychological evidence. The means of distinction between mental states is the proposal that states possess the properties of ″automaticity″ and ″introspectability″ to a greater or lesser degree. This allows Machery to distinguish types of mental state or process according to their possession of these two properties. When a mental state or process appears ′implicit′, this is because it possesses high automaticity, low introspectability, or both. That is, when an experiment identifies something that is repeatable and identifiable in the agent′s cognitive structure, but which is not reported by the agent upon reflection, this is explained in terms of the properties of some combination of the levels of automaticity and introspectability that the relevant processes possess.

The second part of Machery′s explanation provides an argument for the trait picture by arguing that the trait picture offers the best available explanation of the following phenomena: [1] the low correlations between indirect measures; [2] the variation in the correlations between indirect measures; [3] the variation in the indirect measures; [4] and the low predictive value of indirect measures. Machery shows how his distinction explains and fits each of these important phenomena. The relevant point for us to note, is that the explanation Machery offers for these phenomena excludes certain aspects of the apparent evidence as relevant evidence.

Within social psychology and research on implicit cognition, it is common to find experiments which aggregate or take a weighted average of scores across a variety of tests. These experiments express their conclusions and the phenomena they describe in terms of these composite metrics. A composite metric of an entity or phenomenon is often taken by the experimenters to be more reliable than a simple metric due to the reduction of the scope for experimental noise, reduction of the influence of unusual or outlying test results, and the accommodation of the necessary experimental distance from the phenomena common in social psychology. However, these advantages only hold in cases where there is in fact a single phenomenon or entity being measured. When a series of measures are, in fact, measuring distinct and divergent phenomena they become less reliable than simple metrics since their composition systematically occludes information about the relevant phenomena.

If Machery′s trait picture is correct, then these composite metrics are misleading since they are not tracking a single phenomenon but rather aggregating across a disparate range of phenomena. This subset of the available evidence is excluded from the set of relevant evidence by Machery′s trait picture on these grounds. Here we have a case where a model gives us good grounds to believe that evidence for which explanation is offered by other models[40] ought not to be considered relevant evidence. Following the argument laid out in section 3.3, the explanation of composite metrics as relevant evidence now counts in favour of those models which explain it only on the condition that Machery′s trait picture is false, just as the explanation which includes a novel planet is stronger for its doing so on the condition that it is not a scratch on the telescope′s lens.

This presents another case of the problem of desiderata. Where it diverges from the case outlined above is that there seems no prospect of

[40] Composite metrics of one kind or another are either explained or explicitly addressed by Tanesini (forthcoming), Webber (2015; 2016), Holroyd and Kelly (2016), Levy (2015), Schwitzgebel (2010), Gendler (2008a; 2008b), *inter alia*.

settling the matter by reference to predictive power. Where Tanesini and Webber each offer some substantive empirical predictions, Machery's model operates at a greater level of abstraction. Precisely how this abstraction should be understood, especially its relationship with predictions, is addressed more substantively in the following chapter. For the purposes of this section, the model lacks the relevant detail to make substantive, novel predictions against which we may apply desideratum B. The argument for this claim will be made in chapter 4, especially sections 4.5 and 4.6.

This provides us with a 'bad case' of the problem of desiderata. A case where desideratum A is unable to evaluate the model because of the contestation over what counts as relevant evidence compounded by the lack of relevant predictions to allow us to appeal to desideratum B. Applying the desiderata to such a case fails to decide on the basis of explanation, and there is no clear path to settling the decision based on predictions.

Section 3.5: Two Apparent Solutions

As raised at the end of section two, there are two *prima facie* plausible solutions to the problem of desiderata that need to be considered. In this section I consider each of these solutions and argue that they are each undesirable. The solutions are [1] to change desideratum A to accommodate the various modes of explanation, or [2] to introduce a new desideratum to decide problem cases.

The problem of desiderata arises in part because of desideratum A's focus on those aspects of the relevant evidence as the target of explanation. When we have cases of the problem of desiderata, this evaluation is insufficient to decide between different models. It may be argued, therefore, that the problem of desiderata in fact represents the failure of desideratum A to accurately track the true explanatory power of the models in question.

As introduced above, explanation is a one-directional relation of fit which obtains when, and to the extent that, both the parts and whole of the target data are elucidated and accounted for by the model. When we apply the desideratum, we begin by identifying the relata: the model being evaluated and the data of the field and phenomena the model is intended to explain. With the relata identified, we may then identify the features of the model that distinguish kinds of mental states or processes, identify the distinctions in the data, and evaluate the fit of the distinctions in the model with the distinctions in the data.

In the cases outlined above, the first step identified the data to be explained, and thereby the scope of desideratum A, with the relevant evidence excluding the remaining apparent evidence. One ameliorative project for the problem of desiderata may therefore be to explicitly include the explanation of the apparent evidence as relevant to the assessment of the model's explanatory power. Two directions of this ameliorative project are tentatively outlined below:

A_i. The explanation of the relevant available evidence through distinguishing kinds of mental states or processes, in accordance with the implicit and explicit psychological evidence, insofar as the set of relevant evidence is a sufficiently large subset of the available evidence.

A_{ii}. The explanation of the entire available evidence, incorporating both explanation of the relevant evidence and the justification of the exclusion of the remaining available evidence, through distinguishing kinds of mental states or processes, in accordance with the implicit and explicit psychological evidence.

Each direction for desideratum A attempts to address the problem of desiderata by incorporating the relationship between the model and the reasonably excluded apparent evidence into the assessment of explanatory power. The former conditions the quality of the explanation of the relevant evidence on the scale of the relevant evidence as a proportion of the total apparent evidence. The latter expands the explanations we ask of the target model to include the justification of the exclusion of the remaining evidence as part of the explanation.

However, the problems of doing so are threefold. First, and least problematically, what counts as sufficiently large is likely to be problematically interpretable. Second, even with an uncontroversially large subset, there will be divergence between the sets of data that two models explain. Both may explain a substantial subset of the available evidence, but if those subsets are not identical then the problem is not resolved. This is exacerbated in cases where we may wish to compare more than two models. Third and finally, the desiderata are useful tools precisely because of the clarity they provide by isolating a key metric of the quality of the model under analysis. By having a single desideratum monitor two independent metrics the primary utility of applying desiderata is undermined.

As such, the changing of desideratum A to make it sensitive to the relationship between the available data and the relevant data as well as to the concordance between the relevant data and the model is not a desirable option if it can be avoided.

If reframing the expression of desideratum A is undesirable, why not introduce a new desideratum C to settle problematic cases? For example, we may introduce the desiderata C_1 or C_2:

C_1. All other things being equal, the better model will offer the most complete explanation of the available evidence.

or

C_2. All other things being equal, the set of relevant evidence will be the majority of the available evidence for a good model.

or

C_3. Where desiderata A and B are inconclusive, the simpler model is the better.

These desiderata offer examples of an additional metric by which we might judge a model ʹgoodʹ in order to settle undecidable cases. However, unlike the desiderata A and B, cases can be imagined, and likely exist in the literature, which make the statements untrue. In the case of C_1, a model which explains all the available data that is not the result of experimental error, is worsened by the further inclusion of experimental errors as relevant evidence. Similarly, in exceptionally noisy research programmes or programmes in crisis, the majority of available evidence may well not be relevant evidence. Finally, the simpler model in undecidable cases may lack the detail which makes it applicable to the experiment we are trying to design, or to the phenomenon we are investigating. The relatively uncontroversial nature of A and B likely derives from their relationship with explanation and prediction – core aims of the research programme as a science, as an attempt to understand some set of phenomena. A new desideratum will either risk 'stepping on the toes' of these desiderata or deriving its normative force from something more controversial, or less universal to research programmes.

The advantage of desiderata approaches is to offer a relatively straightforward analysis of the qualities of our models against some relatively well-founded normative judgement. What this shows is that the approach for desiderata A and B fails to decide in important cases between relevant candidate models and that the most obvious approaches to improving this problem, insofar as they succeed, do so at the cost of the advantages of the approach. As a result, the desiderata approach fails to enable us to decide between problematic cases and the most obvious responses do not offer clear traction in addressing the problem and bring with them substantial costs to the approach.

Section 3.6: Why undecidability is a problem

One response available from the desiderata approach is to deny that the inability to decide between two models M_1 and M_2, on the basis of desiderata A and B, makes those desiderata unfit for purpose. The purpose of these desiderata is to identify what would make a given model 'good', in the context of research on implicit cognition. At least *prima facie*, two 'good' models may be undecidable. In this case, the desiderata have done their job since either model is good, and therefore either may be used. It is this last point, that undermines the *prima facie* plausibility of the response – that the goodness of a model, and therein the purpose of desiderata, is for and depends upon the use to which it is put.

The relationship between the utility of a model and its explanatory and predictive power addressed in the desiderata requires more unpacking. One example of the complexity in this relationship is the way that a model or framework offers an account of some data, conditional on, among other things, some account of the mechanism of measurement. Both parts are essential to the explanation offered by the model and the evaluation of each is necessary to evaluating the utility of the model for our current project. If we want to know if the trait picture of attitudes is good for our current project, we interrogate its explanations not just in their ability to account for prior data, or to predict future trends, but to do so in a manner which offers an implementable account of the mechanism of treatment and measurement which we intend to use in our further research.

This requires more than an additional desideratum, but rather a breaking down of the desiderata. The desiderata A and B represent the important features of explanation and prediction, but utility of a model depends on more than just the totality of each of these properties; it depends on the structure and properties of the parts that make up those explanations and predictions. What the argument from undecidability illustrates is the way that a desiderata approach fails to draw distinctions between models which are substantially relevant to their utility. In this sense, the reply from a desiderata approach is not wrong in saying that the desiderata identify 'good' models, rather the sense of 'good' being employed is significantly underspecified in a way which undermines the usefulness of desiderata in identifying what we want from a model of implicit cognition.

While we may be, and perhaps should be, pluralists about models, especially about good models, if what makes a model good is not adequately triangulated, or articulated, through desiderata, then we need some richer tool to offer us the relevant traction to resource our judgements about the adequacy, or inadequacy, of a given model.

Section 3.7: The proposal

In place of the desiderata problematised above, this section proposes the application of the ontological critique outlined in chapter 2. This proposal runs in three stages. The first briefly adapts the general form from chapter 2 to the specific question at hand. The second applies this specific version of the ontological critique to the problem of desiderata. The third discussed the relationship between the ontological critique and the desiderata. These stages clarify the manner of application of the ontological critique, specify the theoretical advantages of the critique over the desiderata approach, and identifies the theoretical move as one of situating the desiderata within a more comprehensive theoretical toolkit, respectively.

To reiterate from chapter 2, the ontological critique consists in the following challenges:

1: What do we care about?

1a: In the research programme in general?

1b: In this project in particular?

2: What are the proposed objects?

3: What are the proposed properties of these objects?

4: How does the conjunction of the objects and their properties give rise to the observed evidence?

5: What would it take to falsify this explanation?

5a: What would have to occur for the explanation to be falsified?

5b: What would be falsified if this occurred?

The general form is then substituted into the specific by clarifying the research programme and the particular project:

1a’: What do we care about in research programme of social psychology?

1b’: What do we care about in the project of understanding implicit cognition?

Here we answer the first question with the principles to which researchers may licitly appeal when arguing the advantage or disadvantage of some position. In our present case, and perhaps in many others, this answer will likely include some variation of; accuracy, precision, scope, coherence, fruitfulness, simplicity, clarity, transparency, elegance, etc. If a given researcher argues for X over Y and their reasons for preferring X are challenged, the complete response to 1a will exhaust those principles to which they may licitly appeal.

The answer to 1b will add a layer of detail to the answer to 1a. Where the answer to 1a may list both precision and simplicity, 1b may specify that our understanding of implicit cognition should be precise enough to formulate testable interventions or that simplicity is a licit appeal but cannot outweigh precision in any quantity. This additional layer of detail is added at the project specific level because they may vary between aspects of a research programme, as well as because they may attach to particular targets within the project, i.e. we care more about accuracy in regard to the specific findings of certain implicit bias results regarding race, over similar findings in other areas, because of their moral and epistemic standing.

With the two teleological questions answered, we may address the remaining questions to each model under examination. For example:

2’: What are the proposed objects of Machery’s Trait Picture (2016)?

5’: What would it take to falsify Levy’s account of Patchy Endorsements (2015)?

The answers to the remaining questions clarify the ontological status of the picture presented and thereby highlights unsupported suppositions or overstated conclusions. By demanding specific answers to each question in turn we develop a more precise and transparent account of each model. In conjunction with the answers to 1a′ and 1b′, we have a framework for the heuristic advancement of each account as well as a clear statement of the licit terms of engagement between accounts of a shared target phenomenon.

The second stage of this section is to clarify how this ontological critique applies to the challenge of desiderata introduced above. That challenge was the claim that there are cases where the application of desideratum A will render some models incommensurable because they may give us reasons to believe that some datapoints ought not to be explained as part of the relevant evidence, while other models give us no such reason and provide an explanation of those datapoints. This problem is compounded by the problem that some such cases cannot be settled by the application of desideratum B since many such models, especially in social psychology where the metaphysical grounds of many concepts are up for grabs, offer explanations in terms of mechanisms that are not directly available to measurement. In Machery′s case, no prediction of novel phenomena can be made which would substantively diverge from predictions made by any other competing model since Machery′s model is a claim about the kinds of relationships between the ontological entities that produce patterns of behaviour rather than about how those patterns of behaviour emerge from those entities.

If our only tools for deciding between models are the desiderata A and B, then we are left unable to decide, not only in principle between a great proportion of the possible explanations, but between many of the currently available explanations. This is not because these models are similarly good but because the metric by which we adjudge them to be good does not apply to each aptly enough to bear the weight of a comparison between them.

The problem depends upon our only tools for differentiation being those of the quantity of explanation offered and the availability of risky predictions. Challenges 1a′ and 1b′ demands the clarification of the licit appeals we may make which will include but go beyond the explanation and prediction the desiderata focus on. 1a′ will likely allow appeals to simplicity or precision as deciding factors between models. 1b′ will specify the datapoints, trends or phenomena that we care about explaining to a greater degree than others and will call for the clear statement of this as the grounds for judgements on model choice.

It is perhaps clear at this point that the appeal to the ontological critique does not necessarily resolve difficult cases by its mere application. By introducing greater variety of licit avenues of appeal for researchers advancing one model or another the picture is perhaps complicated rather than resolved. Furthermore, it remains possible that two models may have wildly different strengths and researchers may disagree about the relative weight that those qualities bear on model choice.

If the application of the ontological critique is (as highlighted in chapter 2) more epistemically demanding, presents a less simple comparison structure between models, and does not resolve the problem of desiderata, why bother with it? Why not simply apply our desiderata and accept that model choice is always complex and not always resolvable?

The advantage of the ontological critique lies in its openness demand. It is not simply the injunction to weigh each question that will resolve our difficulties in model choice, rather it is the practice of stating plainly our answers to those questions which will enable heuristic advancement. In the case where two researchers disagree about the relative weights of qualities of the models under examination, by calling for explicit answers to 1a′ and 1b′ we clarify that the disagreement is one within the research programme about what we care about. This is a dispute that, even if not readily resolvable, may be made clear in order that it receives proper attention in the literature and proper care in the laboratory. Furthermore, by calling on authors to state how the objects of their model, in conjunction with their purported properties, give rise to their explanations, not only the fact of explanation or prediction may be evaluated but also the quality.

If a model with minor auxiliary hypotheses entails a given outcome this is a strong prediction as well as grounds for a real risk to the model. If a model offers an explanation only in conjunction with dozens of extrinsic contributing factors, some of which are particularly powerful confounding factors, in addition to the core objects of the model then it is a less strong explanation.

By calling for the explicit statement of these judgement calls within the text of publications, the precision with which claims must be made is improved and the scope for critical response is expanded.

The advantage is not necessarily in the ability to fix problems of model choice at the point of application, though when we are lucky this may prove to be the case, but rather in the long term opens up debates in the field to greater, more precise scrutiny.

Evaluated as a long-run approach to heuristic advancement, we may now situate it more appropriately in relation to the desiderata approach. Where the desiderata approach clearly asks for two identifiable properties of the models under assessment, their explanatory and predictive power, the ontological critique both calls for the identification of the relevant properties of each model and their explicit statement in-text. In cases where the former settles arguments between the two models, it does so at least as well and clearly as the desiderata approach, since the desiderata approach represents a particularly narrow response to challenges 1a′ and 1b′ while the ontological critique is capable of offering these modes of distinction as well as other grounds of model choice. Where a simple decision cannot be made, the clear explicit statement of the answers to the challenges gives the ontological critique utility to future literature that the desiderata approach lacks.

Section 3.8: Conclusion

The desiderata approach to model choice in implicit cognition, while initially promising, has a problematic outcome theoretically where some models cannot be decided between. This problem is not only theoretical, however, but has occurred in the literature (Tanesini, forthcoming; Webber, 2016; Machery, 2016) and may be reasonably expected in a significant proportion of actual cases. This approach is contrasted with the ontological critique which presents not only greater scope for synchronic model choice by appreciating dimensions of analysis that go beyond the desiderata of Holroyd, but also by providing an injunction towards a particular kind of openness in our theorising, explanations, and predictions. The application of the ontological critique provides scope for the long-term heuristic advancement of the research programme of social psychology in general and the project of understanding implicit cognition specifically: which is what we really want from our models of implicit cognition.

Chapter 4: Unpacking the Trait Picture

Section 4.0: Introduction

This chapter introduces the ʹtrait pictureʹ of attitudes proposed by Machery (2016) and his argument against the ʹFreudian pictureʹ of attitudes. The chapter argues that we ought to reject most of Macheryʹs arguments against the Freudian picture and to accept exactly one. This clarifies precisely which commitments of the Freudian picture should be rejected, making the argument directly relevant to contemporary philosophical interest in attitude psychology. With Macheryʹs negative programme clarified, his positive programme, the trait picture of attitudes, is analysed using the ontological critique developed in chapter 2.

The several conclusions are, first, that we ought to accept Macheryʹs argument against a one of the claims that part-constitutes the ʹFreudian pictureʹ, a claim which receives widespread assent. Second, the trait picture requires the introduction and investigation of auxiliary hypotheses to offer a substantive account of its target phenomena. Third and finally, the method introduced in chapter two is demonstrated to be fruitful when applied to a relevant case in the literature.

Section one introduces Macheryʹs negative programme and introduces his foil: the Freudian picture. This also highlights the relevance of the Freudian picture, even if it is unclear that any psychologists or philosophers affirm the whole picture. Section two unpacks Macheryʹs arguments against the Freudian picture. Concluding that we should accept exactly one of these arguments against the Freudian picture, it closes by identifying the key problematic elements of the Freudian picture. Section three addresses Macheryʹs trait picture in the light of the more precise scope of his argument before applying the ontological critique.

This presents Machery as an exemplary case in establishing the importance of the ontological critique. The chapter concludes with a discussion of the implications of the preceding sections: (1) that we ought to offer a model that is sensitive to both the correlation between implicit and explicit measures and to their marked divergence; (2) that Machery′s proposal that implicit measures track parts of the target of explicit measures is a plausible metaphysical interpretation of these phenomena; (3) that unpacking the trait picture constructively demonstrates where it is in need of development to meet its aims; and (4) that the assessment of such a model is fruitfully conducted through the lens of the ontological critique established in chapter two.

Section 4.1: Machery′s Freudian problem

Machery frames his enquiry in terms of competing views about the ″nature of attitudes″ (2016, p. 104) in attitude psychology; ″what kind of things are attitudes?″ (2016, p. 104). Machery′s negative programme begins by introducing an argumentative foil: the Freudian picture.

The Freudian picture is characterised by a series of interconnected theoretical claims and commitments regarding the interpretation of phenomena in attitude psychology:

1. Attitudes are mental states.

2. Attitudes are distinguished from other psychological relations by ″the nature of their formal objects, by their valence, and by their functional properties″ (Machery, 2016, p. 105).

3. Attitudes may be distinguished between explicit and implicit along two dimensions: ″automaticity and introspectability″ (Machery, 2016, p. 106). Explicit attitudes are mental states that their possessor may be consciously aware of and may intervene in the activation of, while implicit attitudes are those which their possessor is typically not aware of and typically may not intervene in the activation of.

4. Attitudes are individuated by their valence and their object.

With the claims of the Freudian Picture laid out, Machery focuses on three of its further commitments. Firstly, claims [1] and [2] collectively entail the claim that [5] attitudes have the same ontological status as other kinds of mental states such as beliefs, emotions, desires, intentions and wishes

(Machery, 2016, p. 107).

Therefore, the conjunction of their formal object, valence and functional properties distinguishes attitudes from beliefs or desires. For example, attitudes and beliefs may share an object but not a valence while attitudes and desires may share both an object and valence but not a functional structure. The conjunction of these properties distinguishes an attitude from a mental state of another kind. From this it follows that [5] attitudes have the same ontological status as beliefs and desires and other kinds of mental states in the same way that beliefs and desires share an ontological status. Attitudes are distinguished in their particular object , valence and functional structure in the same way that we distinguish beliefs from desires or other kinds of mental states; they possess all those ontological properties shared across all members of the set of kinds of mental states except those properties that explicitly differentiate attitudes from other members.

The second commitment is Machery′s proposal that the Freudian picture tacitly includes the further claim that ″mental states can be occurrent″ (Machery, 2016, p. 107). If mental states can be occurrent then, as mental states, attitudes can be occurrent. This occurrence is framed such that mental states ″are psychological events that cause further psychological events″ (Machery, 2016, p. 107). This commits the Freudian theorist to the claim that [6] attitudes can be occurrent mental events where occurrent mental events are psychological events that can cause further psychological events.

Thirdly, Machery proposes the introduction of an ancillary commitment to materialism[41]. Materialism in this context is the claim that brain states are the ontological basis of mental states. If the devotee of the Freudian picture accepts this claim, then they are committed to the further claim that, as mental states, [7] brain states are the ontological basis of attitudes.

[41] Specifically, Machery advocates for the introduction of a commitment to non-reductive materialism. While this is a well-regarded position in philosophy, it is unnecessary for the seventh claim.

It is not clear that any one psychologist or philosopher advocates all seven claims. Machery highlights the example of Wilson, Lindsey and Schooler (2000, p. 102) as an apparent endorsement of at least claims [3] and [6]; Gendler (2008a, p. 642) as an example of endorsement of at least claims [1] and [6]; and finally, Kriegel (2012, p. 475) of at least claims [1] and [6]. While many theorists are committed to parts of the picture, if the argument is against the conjunction of these claims, it is not clear whom the argument is opposing.

I propose that it remains fruitful to engage with the Freudian Picture even if no theorist in philosophy or psychology affirms all of its claims. This is because it sets a framework for the criticism of particular parts of the Freudian Picture, especially those parts which have the widest appeal. This framework is an argumentative strategy that illustrates a single conclusion: that different implicit measures do not all measure the same kind of object. By illustrating this claim the Freudian Picture provides an argumentative framework that need not be accepted by any given theorist while offering a conclusion that does not depend on the wholesale acceptance of the Freudian Picture itself.

Section 4.2: Predictive Validity of Implicit Measures

Machery's first argument against the Freudian picture runs:

i. If the Freudian Picture is true, then implicit measures should have a high predictive validity for macro-level behaviour.

ii. Implicit measures do not have a high predictive validity for macro-level behaviour, they are weak predictors.

C. The Freudian Picture is not true.

Machery states that, to hear philosophers, psychologists and science popularisers talk about implicit attitudes, presumably from the standpoint of something like the Freudian Picture, "one may get the *erroneous impression* that indirect measures of attitudes are *excellent predictors of biased behaviour.*" (2016, p. 120, emphasis mine)

The justification for the first premise may be reconstructed by building upon claims [5] and [6]; that attitudes have the same ontological status as other kinds of mental states and that they occur in such a way that they can participate in causal chains. Beliefs about particular objects cause us to behave in particular ways towards those objects. Our belief that snakes are dangerous causes us to keep our distance from them. Our belief does this in virtue of its participation in psychological causal chains. Since attitudes are analogous to beliefs in these relevant senses of ontological status and causal efficacy, the prediction of object relevant behaviour on the basis of belief gives us a reason to believe that our attitudes also enable the prediction of object relevant behaviour. Thus, if the Freudian picture is true, implicit measures should have a high predictive validity for macro-level behaviour.

As Machery highlights, the second premise is now considered to be relatively well established. In particular, the two meta-analyses that he cites have been widely influential in changing how psychologists and philosophers view the findings of implicit measures.[42] If we accept the two premises, then the conclusion follows.

While Machery says that the Freudian Picture theorist may respond with *ad hoc* adjustments to meet the challenges, there seems to be a more

[42] The two meta-analyses cited by Machery are Greenwald et al. (2009) and Oswald et al. (2013). The conclusions of these meta-analyses diverge. Greenwald et al. (2009) conclude that the low predictive validity of IATs is significantly higher than those of self-report measures and therefore useful in predicting biased behaviour. Oswald et al. (2013) conclude that IAT scores offer only a ʺpoor prediction of racial and ethnic discriminationʺ (pp. 171, 183) and provide ʺlittle insight into who will discriminate against whomʺ (p. 188). Greenwald et al.ʹs sampling criteria identified effects where there was a prior theoretical justification for the existence of a given effect, while Oswald et al. included effect sizes where no theoretical justification for an effect was available. Greenwald, Banaji and Nosek argue that we should conclude from these findings that ʺsmall effect sizes affecting many people ⋯ repeatedly can have great societal significance.ʺ (2015, p. 560) However, the variation in predictive validity when priors are explicitly incorporated in sampling choices likely indicates the presence of a further covariate of the implicit measure and biased behaviour. Thus, while some of the text of Greenwald at al. (2009) and Greenwald, Banaji and Nosek (2015) seems to speak against Macheryʹs conclusion, their findings do not.

principled response to this challenge. The Freudian picture does not entail a commitment to the first premise.

The Freudian Picture adherent needn′t commit to the general predictive validity of implicit measures for biased behaviour − only for particularly high-speed non-deliberative behaviour and even then a single implicit measure is susceptible to other confounding factors. That is, the Freudian theorist expects an implicit measure to predict a class of behaviours only under very strict, relatively unusual circumstances. While the meta-analyses cited by Machery rule out the general predictive validity of implicit measures, it is not clear to what extent this conclusion bears out if the behavioural measures which are not high-speed or are deliberative are excluded and whether other confounding factors systematically confound these predictions.

The Freudian theorist would not expect an IAT to identify a tendency to express race-based biases in one′s deliberative behaviour. Rather the IAT may be expected to predict, for example, minor body-language changes toward particular people under circumstances where deliberation about that behaviour is constrained, in the absence of confounding factors. These confounding factors may range from the subject′s beliefs about the norms of behaviour in that social setting to their own emotional state at the time of that interaction.

That is, the Freudian theorist can and likely would reject the claim that the conjunction of [5] and [6] in fact entails premise i. Rather the properties of attitudes which make them distinct from other mental states also make them distinct in their influence on macro-level behaviour. By rejecting that [5] and [6] entail premise i, the argument against i no longer bears on the soundness of the Freudian picture.

Predictions of the sort we would expect from the Freudian theorist are not adequately addressed or refuted by the available meta-analyses. As such, while further analysis may bear the conclusion out, Machery′s first argument is not sufficient to reject the Freudian picture.

Section 4.3: Context Variation

Machery's second argument runs as an inference to the best explanation:

i. Implicit measures have high context variation.

ii. Other things being equal, the best explanation of (i) will offer novel and surprising predictions about the context variations of implicit measures, such that interventions eliciting or confounding these variations may be designed and implemented.

iii. The Trait Picture offers novel predictions of the nature of this context variation.

iv. The Freudian Picture does not offer such predictions.

c. We should accept the Trait Picture over the Freudian Picture (other things being equal).

Machery motivates the first premise with reference to several experiments that seem to demonstrate context variation in implicit measures. What Machery means by context variation is that ″indirect measures... [vary] from context to context, depending on subtle features of subjects′ environment″ (2016, p. 118). There is significant consensus in the literature that such a phenomenon is widespread and consistent enough that it should be understood as a distinctive property of attitudes (Rydell and Gawronski, 2009; Gawronski and Sritharan, 2010; Peck et al., 2013).

The second premise states the importance of specific, implementable predictions to good psychological science.[43]

In order for a model to meet the third premise we expect a set of (1) concrete predictions, (2) that follow from the adoption of the Trait Picture, (3) about when context dependence will and will not occur, that (4) are substantively novel, and (5) which turn out to be true. Each of these

[43] For further information about the importance of prediction and its role in science, omitted the sake of brevity, see Popper (2002, pp. 9-10), Kuhn (1977, Chapter 13), and Lakatos (1978, pp. 34-5; 1970, p. 116-8). See also the discussion of the importance and role of prediction in section 1.5 and section 2.6, as well as more cursory discussion in section 3.1.

conditions may be met to varying degrees, and our interest in this premise is to establish the extent to which the Trait Picture meets them. Similarly, the fourth premise claims that, to whatever extent the trait picture does meet these criteria, the Freudian picture does not.

The Freudian Picture theorist may respond to Machery that, if we clarify the standards we wish to apply to each picture, it becomes clear that under those conditions which lead us to regard the third premise as true we are also committed to the falsehood of the fourth premise and under those conditions which lead us to regard the fourth premise as true we are also committed to the falsehood of the third premise. In either case, the argument fails.

To motivate the claim that we have concrete predictions, that meet the above requirements, Machery offers the example of Schaller, Park and Meuller (2003), who found that subjects who believe in a dangerous world present a higher implicit association test result associating black persons with danger when the test is taken in a dark room. Machery proposes that the Trait picture offers us the prediction that this would be the case. To clarify, our familiarity with the kinds of objects postulated by the Trait theory (″good old fashioned mental states and processes, such as emotions, self-control, and so on″ (Machery, 2016, p. 118)), and their means and modes of operation, is what allows us to make a common-sense prediction about when these implicit tests will yield higher or lower correspondences [44].

Following the analysis above Machery is committed to the claims that: (1) Darkness would cause those who believe in a dangerous world to demonstrate greater correspondence between black people and threat when tested in a dark room than when tested in well-lit conditions; (2) This follows from the adoption of the Trait picture because of the foundation of the Trait

[44] Precisely how this understanding yields these predictions is not explicitly outlined, but we may offer a plausible reconstruction. For example, our understanding of the mechanics of these sorts of mental states and processes allows us to extrapolate the likely behavioural consequences of them to predict their interactions and outcomes in these sorts of novel situations.

picture in well-understood mental states; (3) In this case the understanding that an individual who has an overall schema or heuristic for the heightened recognition of danger will more readily recognise some particular entity as a threat when they are in a state that amplifies the influence of that overall schema or heuristic, such as darkness; (4) This prediction is substantively novel and surprising; (5) Following the research by Schaller, Park and Meuller (2003) it seems likely to be true.

By breaking down the claim into its component commitments like this two things become clear. First, the example is not a prediction. The experiment that seems to illustrate the claim was conducted a decade and a half before the ʻpredictionʼ and its results were available to Machery. Prediction is preferred to explanation because of the flexibility of the explanation of data by a model and the relative inflexibility in predicting an unknown outcome.

It is also false that this experimental result is substantively novel or surprising. From the framework of the argument itself, it is because the Trait picture utilises familiar concepts that our ʻpredictionʼ comes about. The Trait picture is presented as good precisely because its predictions are unsurprising. This is not in itself a theoretical vice; if the predictions turn out to be true then all is well. However, when the prediction no longer needs to be surprising, prediction is not substantively harder to offer than explanation. As highlighted in section 2.6, in a research programme in crisis, even banal predictions remain risky and are therefore valuable. However, the relevant risk is only incurred when predictions are made prior to the outcome being known.

Machery emphasises the predictive quality of the Trait picture over that of the Freudian picture. He admits that the Freudian picture is as well equipped to explain the phenomena (2016, p. 118), but claims that the Trait picture offers more than mere explanation – it offers prediction also. But since any such prediction would be neither novel nor surprising, on his own account, it does not possess the qualities that make a good prediction superior to a good explanation. Predictive power offer us no reason to accept the Trait picture over the Freudian picture.

This point may be further illustrated if we were to lower the bar of prediction from ʹsurprising and novel, true predictionʹ to ʹany true prediction will doʹ. Here we may turn to the fourth premise that the Freudian picture does not offer the predictions we want from a model of implicit cognition since it does not offer predictions about when we may expect to find context variation. Suppose the Freudian theorist is told of an upcoming, unprecedented experiment to test the relationship between darkness, stress and race responses. On what basis would a Freudian theorist make a prediction of the outcome?

The Freudian theorist seeks to explain the effects of implicit cognitions in terms of their being a kind of mental state, having a valence, object and functional structure, and which is highly automatic and possesses a low degree of introspectability. In this case the Freudian theorist will identify the attempt to isolate those participants who have a negative explicit attitude towards the world at large and a negative implicit attitude towards black people. This set of participants, along with the control, will be exposed to darkness (a stress inducing environment for diurnal animals). What kind of interaction, on this analysis, would we expect the Freudian theorist to give?

I propose that the Freudian theorist would expect to see an interaction in precisely the same manner as the Trait theorist. The group of participants who have both the belief in a dangerous world and a negative implicit attitude towards black people would be expected to exhibit their implicit attitude to a greater extent under stress conditions that constrain introspection and empower automaticity. This prediction is no better than what the Trait theorist offers, it being entirely unsurprising.

By clarifying the standard adopted for the third and fourth premises we find that they cannot be simultaneously true. If we adopt the high standard implied by Macheryʹs argument we find the third premise to be false, but if we lower the bar in order to consider the third premise true then we find the fourth premise becomes false. In either case, the argument fails to establish its conclusion. We may therefore reject the second argument.

Section 4.4: Correlations Between Implicit Measures

Machery′s third argument runs:

i. If the Freudian Picture is true, then we should find a high correlation between implicit measures or a systematic divergence between these measures.

ii. We do not find a high correlation between implicit measures or a systematic divergence between these measures.

iii. c. The Freudian Picture is not true.

The first premise is intended to be derivable from the conjunction of claims [3], [4], [5] and [6]. That is, if we may distinguish between explicit and implicit attitudes, and if implicit attitudes have the same ontological status as other kinds of mental states and if occurrent implicit attitudes are psychological events that cause further psychological events, then these implicit attitudes (identified by their high automaticity and low introspectability) will behave like other mental states (such as a desire) and be detectable stably by multiple different measures. Thus, independent evidence of the existence of such a mental state would allow for a conclusion that such a mental state exists to explain the convergent experimental outcomes. If a given mental state exists and multiple measures track its presence, we should see these correlating or systematically diverging.

Machery provides as evidence for the second premise, Nosek et al. (2007, p. 274), Sherman et al. (2003), Olsen and Fazio (2003), and Bar-Anan and Nosek (2012). I further recommend Schimmack (2019) for an analysis of the IAT. The second premise is no longer considered controversial.

One may respond to Machery that the consequent of the first premise does not follow from the antecedent. This is because the premise overlooks attitude functions. If a particular attitude′s structure codifies the attitude′s function (Tanesini, forthcoming), the need the attitude serves, then we should expect to see divergence between implicit measures which attempt to elicit responses through differently codified stimuli even if those stimuli are apparently the same object. For example, an attitude formed through an ego defensive function and an attitude formed through a knowledge acquisition function may be connected to the same object and possess the same valence, but their functional structure will differ. This divergence may mean that different implicit measures will record different results depending on the relation between the attitude′s function and the measure′s framing of the attitude object.

While this response has some success in characterising the low correlations between different indirect measures in some cases, it does not account for those cases where there is no such shift in the framing of the object and where low correlations are still observed.

There is a more pointed challenge to this response still, that if variation across function were the explanation for low correlations between measures, we should expect to see these low correlations remain relatively stable, which they do not. That is, while we may not expect two measures to correlate, the response available to the Freudian Picture implies that the low correlations between measures should be relatively stable and systematic, however, the studies cited above do not find stable or systematic covariation between measures. The response also tacks close to positing separate attitudes for separate measures and, as Machery points out, ″It is bad scientific practice to postulate a theoretical entity for every measure.″ (2016, p. 117)

As such, the first premise is true, and inescapable for the Freudian theorist, in a way the first argument′s was not. The evidence for the second premise is also robust, with a variety of studies converging on the same conclusion. We ought, therefore, to reject the Freudian Picture as the conjunction of all seven claims on the basis of this argument.

Section 4.5: The Core Argument

Given the above analysis of Machery, and the rejection of two of his arguments, what conclusion are we now committed to? If we accept the third argument the conjunction of the seven claims from section one can be rejected. However, it was far from clear that any theorist avows all seven claims. In order to demonstrate that the argument in fact targets only a subset of the claims introduced above, a subset which has far more widespread acceptance, we must identify the claims that are party to the successful argument. Which of the claims that characterise the Freudian Picture are individually necessary and jointly sufficient to establish the first premise?

i. If the Freudian Picture is true, then we should find a high correlation between implicit measures or a systematic divergence between these measures.

Attitudes have particular functional properties that distinguish them from other kinds of psychological relations. These functional properties can be isolated and identified by relevant experiments to track the presence and valence of a given attitude. Specifically, the distinction between implicit and explicit attitudes is grounded in the presence or absence of automaticity and introspectability. Those measures which depend upon high automaticity and low introspectability track only those attitudes which have those features of being highly automatic and unavailable to introspection; implicit attitudes. This is the consequence of the Freudian picture′s commitment to claim [3].

These implicit measures come in various kinds from Implicit Association Tests (IAT) to Evaluative Priming (EP) to the Affect Misattribution Procedure (AMP)[45]. Each of these tests removes the opportunity for deliberative control or introspection and each aims to control for the object of the attitude in order to test its valence and stability under different circumstances. The claim that ′implicit measures track implicit attitudes′ raises the question of whether all of these measures can or do track the same implicit attitudes.

[45] Methods developed by Greenwald, McGhee, and Schwartz (1998); Fazio, Sanbonmatsu, Powell, and Kardes (1986); and Payne, Cheng, Govorun, and Stewart (2005), respectively.

From [4], the Freudian Picture is further committed to the individuation of attitudes by object and valence. Each measure controls for the object of an attitude and tests its valence under different conditions. Thus, the Freudian picture is committed to the claim that when an implicit measure offers a relatively stable report of an attitude towards some object of a given valence, that measure is tracking an attitude. Since this is true for each measure, when those measures are tracking the same object we should expect those measures to display relatively high correlation, as measures of a single object, or that the measures should diverge in a systematic manner as different measures identify different properties of their shared object of observation.

From this we can see that premise i from the original argument may be reframed in terms of commitments [3] and [4]:

i'. If [3] implicit attitudes are distinguished on the basis of high automaticity and low introspectability and [4] attitudes are individuated on the basis of valence and object, then those experiments which ensure high automaticity and low introspectability and which control for object are tracking the same target object (an implicit attitude).

ii'. If experiments are tracking the same target object, then the experiments should present data on valence that correlate or that systematically diverge.

iii'. If experiments present data that neither correlate, nor systematically diverge, then those experiments are not tracking the same target object.

iv'. The Implicit Association Test, Evaluative Priming, and Affect Misattribution Procedure (and such like) present data that neither correlates, nor systematically diverges.

v'. The IAT, EP, and AMP (and such like) are not tracking the same target object.

vi'. The IAT, EP, and AMP (and such like) ensure high automaticity and low introspectability and control for object.

vii'. The conjunction of [3] and [4] is false. It is not the case that implicit attitudes are distinguished on the basis of high automaticity and low introspectability and attitudes are distinguished from one another on the basis of object and valence; one or both of these claims is false.

Attitude psychology is committed to the truth of [4] as part of its hard-core commitments. This claim was illustrated by the partitioning of attitude psychology into the two programmes of explicit and implicit attitude psychology in the 1980s following the apparent discovery of attitudes toward an object with multiple, diverging valences – interpreted as multiple attitudes toward that object. That is, the commitment to [4] is robust enough that the research programme will split into two new programmes with distinct experimental approaches rather than abandon it. This commitment is, furthermore, unsurprising since preserving the individuation of attitudes by object and valence preserves many of our folk practices of attitude attribution – "S likes/dislikes O". Attitude psychologists are committed to [4], and by the above argument are therefore committed to the falsehood of [3].

This precision affords a more significant conclusion than Machery's original conclusion that the conjunction of [1]-[7] is untenable. [3] is widely held among philosophers working on social psychology and is ubiquitous among social psychologists. An exhaustive list of those committed to [3] might take up the remainder of this chapter but some notable figures in psychology and in philosophy who explicitly endorse it include Greenwald et al. (1998), Fazio et al. (1986), Haddock and Maio (2015), Gendler (2011), Kelly and Roedder (2008), Holroyd and Kelly (2016), among many others.

Section 4.6: Ontological Critique of the Trait Picture

With Machery's negative programme concluded, we are now in a position to apply the proposed ontological critique to Machery's positive programme: the trait picture of attitudes.

The central claim of the Trait Picture is that ″attitudes are traits⋯ broad track dispositions to behave and cognize (have thoughts, attend, emote, and so on) toward an object⋯ in a way that reflects some preference. ″ (Machery, 2016, p. 112)

. What distinguishes the trait picture from other models of attitudes is the properties it attributes to them. On the trait picture, to say that S has attitude X toward some object O is to say that S has a broad-track disposition to behave and cognize in a particular way towards O in a way which reflects the affective tenor X has for them. By broad-track disposition, Machery means a tendency to respond to a wide range of inputs with a wide range of outputs where the codification of each is related but inputs and outputs do not directly correspond one-to-one. My attitude that cows are dangerous is a broad track disposition towards cows that reflects the affective tenor of the attitude. The disposition being broad track means that seeing a cow in the same field as me does not entail running for the nearest stile. I could instead behave in any of a wide variety of other ways which are all expressions of the attitude′s valence.

On the trait picture, attitudes have a psychological basis that consists of a ″motley assortment of mental states and processes″ (Machery, 2016, p. 112). This basis includes mental states, affects and emotions, as well as regulatory processes. The basis of each attitude will consist in some heterogenous mix of these psychological phenomena. Different aspects of the base will be more or less influential on the attitude overall, while some will be more or less influential on the attitude in combination with some other external or internal stimulus.

This description of the trait picture can now be formalised by applying the ontological critique from chapter 2. The ontological critique consists in five challenges: Axiology, Objects, Properties, Explanation, and Falsification.

The first challenge is the axiology challenge: What do we care about? What do we care about in the research programme of social psychology in general, and what do we care about in the project of attitude psychology in particular? *An ideal answer to this challenge identifies the dimensions of excellence or deficiency which are legitimate recourse for advocating or criticising the model.* In general, we care about the accuracy and precision of the model′s explanations, and our ability to offer fruitful predictions on their basis. Furthermore, we care about the simplicity of the model as a theoretical virtue as well as its coherence with other related areas of research. Within attitude psychology specifically we want an explanation of the variations in and between implicit and explicit measures toward some target object, an explanation of the low predictive validity of implicit measures for macro-level behaviour, and ideally, coherence with models in cognitive psychology, or other related fields. Finally, we want a model that offers predictions of when responses will be elicited and how interventions can induce or prevent responses.

The second challenge is the objects challenge: What are the proposed objects of the model? *An ideal answer will identify all those entities to explanations of phenomena will refer. Central objects are those which are novel to the model or given a novel exposition within the model. Peripheral objects are those which are not novel to the model, but which are necessary to offer the model's explanations*. The peripheral objects include dispositions (broad and narrow track), traits, mental states and processes, target objects (which an attitude is about), persons, and behaviour. The central objects are attitudes, parts of attitudes, valences.

The third challenge is the Properties challenge: What are the proposed properties of these objects? *An ideal answer will enumerate all those properties of the objects outlined in the previous challenge that are necessary for the explanations offered by the model including their relations with other objects of the model.* Dispositions are tendencies for something to occur, given an impetus. Traits are a kind of disposition. Mental states and processes are those things that include, relate and group; beliefs, desires, emotions, and so on and incorporate various coherence and interaction mechanisms. Mental states can cause behaviour and can have target objects. Target objects are what a mental state or process, or attitude is about[46]. Persons, for the sake of this model, are those things which can have mental states and attitudes. Behaviour is the broad category of responses persons can give to stimuli.

Attitudes are traits. They are about a target object and a possess a valence. They code behaviour. As such, they are broad track dispositions for persons to behave in such and such way in response to a target object in accordance with the attitude′s valence. Attitudes are constituted by parts of attitudes which may not share the valence of the whole toward the target. Parts of attitudes are mental states and processes.

The fourth challenge is that of explanation: How does the conjunction of the objects and their properties give rise to the observed evidence? *An ideal answer clarifies what role the ontological commitments play in our explanations.* The trait picture explains the occasional divergence of implicit and explicit measures by identifying the measures as tracking distinct objects. Explicit measures track beliefs or judgements. Implicit measures track parts of attitudes. Since they do not track the same object, we should expect them to be able to diverge. The general correspondence of implicit and explicit measures is explained by their target objects both playing a role in the attitude toward the target object. The attitude itself captures and explains where we find general consistency in object relevant behaviour. This gives us reason to expect measures to align but only imperfectly. Finally, the low predictive validity of implicit measures for macro-level behaviour is explained by identifying that these measures track only a part of an attitude. This gives, at best, a weak guide to the valence of the overall attitude which will govern behaviour.

The fifth and final challenge is the falsification challenge: What would it take to falsify these explanations? Specifically, what would have to occur for the explanation to be falsified and what would be falsified if this occurred? *An ideal answer identifies some potential results that would falsify the model.*

[46] Aboutness is here construed deliberately broadly, to capture related notions such as reference, directedness, and the like.

Given such a result, it further identifies the features of the model we take to be at stake in such a falsification.

Examples of findings that would falsify the trait picture might include evidence that behaviour lacks the sort of object-oriented consistency that would arise from a broad track trait with a valence. Alternatively, we might find evidence that attitudes do not have an object-stable valence. In the former case behavioural traits of all kinds would be at stake, while in the latter attitudes would cease to be a useful category of explanation – their parts and mental states would then be sufficient to explain the phenomena. What this further clarifies is that the lack of precision means that the population extensions over which we should expect these results is incredibly broad.

The critique highlights both the strength of the trait picture in offering the explanations it does, and the level of abstraction it is operating at. This abstraction appears to be what Holroyd is referring to with the claim ″ So far as the trait view offers us a theory of attitudes, it tells us that these attitudes are not themselves implicit cognitions. But it does not give the rest of the story about implicit cognition.″ (Holroyd, 2016, p. 174) While we have a broad categorisation of the relations between attitudes, explicit and implicit measures, we do not have enough detail to construct interventions to test these relations. This criticism from Holroyd can be refined by highlighting its consequences for the fifth challenge. Because of its level of abstraction, the trait picture presents claims that could only be falsified by behavioural patterns not being codifiable by object or by valence. While we need not be Popperian and insist on the utter falsifiability of our theories, this demonstrates the first part of the picture which needs refinement. At present the model possesses commitments which could only be falsified if our understanding of psychology is overturned wholesale. To improve the model, it needs sufficient expansion of its core commitments or the introduction of auxiliary hypotheses to allow it to make substantive predictions about interventions. One way to do this would be to offer specific claims about the generalisation populations; ″the class of settings X have the relevant properties″, ″persons who display trait Y in A and B control measures″. Simply adding greater detail to our accounts of the relevant populations offers both greater precision and claims which may be tested by more specific falsifications.

Given Machery′s argument for the trait picture over the Freudian, on the basis of its use of ″good old-fashioned mental states and processes″ (Machery, 2016, p. 119) addressed in section 4.3, the data and phenomena the picture is trying to explain and predict are well known and familiar to us. As a result the direction of effect of these predictions is unlikely to be novel; that there is a positive link between Xs and Ys is not a novel prediction. However, in conjunction with sufficient auxiliary hypotheses the trait picture may well offer a precise enough account to make predictions which will accord with our intuitions in general, but which will be novel in their specificity. We may be able to specify pivot-point conditions where otherwise reliable effects can be surprisingly confounded, or surprisingly elicited, or we may be able to make predictions about the size of effect we would expect to find. These point predictions or pivot predictions do not tell us that there is an effect but instead where to look for surprising features of familiar phenomena. This level of specificity is needed to offer sufficiently risky predictions for familiar phenomena.

The second criticism the critique raises stems from the identification of objects or properties that bear substantial weight in the explanations being offered by the picture without similarly substantial exposition of these objects or properties. In this case, the criticism revolves around the role of coherence mechanisms. One of the core explanations the trait picture seeks to offer is the general alignment of implicit and explicit measures while allowing for their divergence. The trait picture′s explanation of the divergence is that the target objects of the measures are distinct, while it explains their general alignment by means of coherence mechanisms. These coherence mechanisms are mentioned only in passing by Machery but play a key role in the trait picture. This is the second part of the trait picture which needs refinement.

Section 4.7: Concluding remarks

With the trait picture outlined and codified we can relate it back to the conclusion of the negative programme. The conclusion to the argument in the previous section was that the conjunction of [3] implicit attitudes are distinguished on the basis of high automaticity and low introspectability and [4] attitudes are individuated on the basis of object is false. Since attitude psychology is predicated on the truth of [4], we have good reason to reject [3]. The trait picture offers us one way to reject the claim: there are no implicit attitudes.

Machery frames the Trait Picture as a way of understanding attitudes such that we do not commit the cardinal sin of mistaking them for mental states. What we are now able to clarify, is that the motivation for adopting the trait picture is that it offers a plausible account of the evidence without relinquishing the commitment to object oriented attitudes. The reason to prefer this over other accounts which maintain implicit attitudes is that maintaining [3] given [4] runs contrary to the best available evidence.

Contrary to Machery′s claim that the Trait picture offers the best explanation to some otherwise puzzling findings (2016, p. 125), the Trait Picture′s advantage is rather that it does not depend on claims contrary to the best available evidence.

The first conclusion we ought to draw from this analysis of Machery is that attitude psychology has significant work to do to explain the evidence it has available. It appears that in order to do so it must reject the claim that implicit attitudes are distinguished on the basis of high automaticity and low introspectability. Machery′s trait picture offers a broad framework for doing so which has the advantage of accuracy in this regard and coherence with other fields. At present, the model lacks precision and requires auxiliary hypotheses to make substantive predictions. Furthermore, it requires an elaboration of coherence mechanisms that vindicates their use in the explanations offered. These heuristic progressions would leave the trait picture as among the most robust accounts of attitude currently available. Their implementation would enable substantive and novel predictions, testable by interventions. Without them, the account does not fully resource its explanations and could only be falsified by results that would overturn our understanding of psychology wholesale.

The breakdown of the dimensions of excellence or deficiency for a model in attitude psychology as well as the structured approach to diagnosing a model′s place along these dimensions enables a clarification of the problem in the field and how the trait picture resolves that problem, as well as clarifying those dimensions on which the model needs improvement. This is not simply an attempt to identify which model is right or wrong, but to identify the strengths and weaknesses of a model relative to our needs as researchers. This allows us to clarify the next steps in the research programme: the introduction of auxiliary hypotheses that allow for precise predictions which may be tested. In the event that these predictions are falsified we also have a clear tool for identifying the target of the falsification and what will survive nature ′saying no′. Given social psychology′s position, this illustrates the importance of the ontological critique. It is not enough to identify problems as they stand; we must also identify fruitful paths forward from here.

Chapter 5: Ontological Critique and the Mainstream Model of Attitudes

Section 5.0: Introduction

This chapter presents and discusses the mainstream model of attitudes developed by Haddock and Maio (2010; 2012; 2015), as articulated by Tanesini (forthcoming). Section 5.1 introduces the mainstream model as it is presented by Tanesini; enumerating the features and properties of attitudes for the mainstream model. Section 5.2 presents two *prima facie* challenges to the mainstream model. The first is that certainty as strength is ambiguously interpretable, to a greater extent than recognised by Tanesini. The second is that the grounds of the implicit/explicit distinction are not internal to the mainstream model but are relevant to the core explanations the model seeks to offer. Section 5.3 presents Tanesini's preferred response to the latter challenge; situating the mainstream model of attitudes within the MODE model developed by Fazio (et al. 1986; 1990; Fazio and Olsen, 2014). MODE and the resulting combination of Mainstream-MODE are then outlined. Section 5.4 presents a final problem for the MODE-mainstream model, the *secunda facie* challenge; the apparent duplication of entities within the model. Section 5.5 applies the first four challenges of the ontological critique to the Mainstream-MODE model. In doing so relevant systematic differences between the two parts of the model are highlighted and, by identifying the relevant objects and properties, the apparent duplication of entities is dissolved. Section 5.6 applies the falsification challenge and explores what it would take to falsify the model as well as ways in which the model might constructively respond to such findings. The chapter concludes by identifying the power of the ontological critique in resolving particular kinds of challenge, while identifying those challenges which the critique may clarify but not resolve, and by highlighting the strengths of the MODE-mainstream model and the areas in which theoretical progress may be called for.

Section 5.1: Mainstream model *a la* Tanesini

Tanesini (forthcoming) argues that at least some intellectual virtues and vices should be understood as collections of attitudes. This situates her argument within the context of virtue and vice epistemology. I shall focus on the framing of attitudes presented by Tanesini, without addressing the corollary framing of virtues and vices, as such the conclusions drawn are conclusions about the mainstream model utilising Tanesini′s explication, and do not necessarily extend to conclusions about the argument Tanesini presents.

The model of attitudes outlined by Tanesini is the mainstream model in attitude psychology. Its key proponents, and the primary sources that Tanesini makes recourse to, are Haddock and Maio (2012, 2015). I present the model as represented by Tanesini because of the concise, precise, and clear representation of the features and properties of attitudes as well as their relation to the experimental data. While it does not differ in substance from the presentation by Haddock and Maio, the clarity of its presentation and discussion makes it a superior candidate for the following analysis. [47]

Tanesini introduces attitudes as ″summary evaluations of objects″ which ″may be thought of as likes or dislikes.″ (forthcoming, p. 6) The framing that Tanesini gives of attitudes and attitude psychology breaks down into their features, properties and measurement.

[47] A philosopher′s representation has further advantages, especially avoidance of basic conceptual confusions. Haddock and Maio make several such errors, e.g. ″To introduce different measures of attitude, we have elected to distinguish them on the basis of whether they are explicit (i.e., direct) or implicit (i.e., indirect). The distinction between explicit and implicit processes has a long history within psychology. Psychologists usually think of explicit processes as those that require conscious attention, while implicit processes are those that do not require conscious attention.″ (Haddock and Maio, 2015, p. 11) In this passage the authors move from implicit and explicit as properties of the experiment to properties of psychological processes without any justification, assuming there are substantive parallels between the two uses of the terms as well as introducing the concepts of consciousness and attention, which are not further elaborated upon. Such confusions are not necessary to the model and are absent from Tanesini′s representation.

Four features of attitudes are identified: object, content, structure, and function.

The object of an attitude is what the attitude is about. Attitudes are intentional, and as such have some object which they are about [48]. This object may be a concrete particular, a group, a value, or an abstraction. Sam′s attitude that ″salty food is revolting″ has, as its object, ″salty food″. Sam could instead have the attitude that ″this anchovy is revolting″, or that ″saltiness is revolting″. For the mainstream view, attitudes are partially individuated by their object. We distinguish between the attitudes ″salty food is revolting″ and ″mathematics is elegant″ in part by their distinct objects.

An attitude′s content is the informational basis on which the attitude was formed. It is generally taken to consist in three parts: cognitive, affective, and behavioural. The cognitive element is the set of beliefs one has about the object of the attitude. Sam′s attitude about salty food may have been formed from an informational base which includes the belief that salty food is bad for them. This belief is a cognitive component of the content of their attitude. The affective element is the emotions and feelings that one has about or towards the object. Feelings of revulsion or nausea which accompany the thought of salty food may constitute affective components of the content of their attitude. The behavioural element is the set of memories of past behaviours and experiences regarding the object. These components can include memories of particular events, such as a cake Sam mistakenly baked with salt rather than sugar, or behavioural habits that have been built out of these events, such as picking salty ingredients off shared pizzas. This diverse informational base is ′summarised′ in the attitude′s evaluation of the object.

An attitude′s structure describes the valenced structural features of the content of an attitude. The contemporary view, advanced by Haddock and Maio (2012), is that the content is structured along two dimensions; that an attitude′s valence is not an aggregate evaluation varying from very

[48] Some psychologists prefer the terms ′refer to′ or ′are directed towards′ or ′consist in the evaluation of′, I use ′about′ to avoid confusions and disagreements surrounding reference, direction and constitution since they do not have a bearing on this argument.

favourable to very negative. Instead, positive and negative elements are aggregated separately. Therefore, valence may be represented as a point in Cartesian space where one axis measures levels of positivity and the other of negativity. (Haddock and Maio, 2012, pp. xxv-xxvi) This co-ordinate summarises the positive and negative valenced components of the content and represents another part of how attitudes are distinguished. We distinguish between Sam′s attitudes towards mathematics and salty food by their target object but also by their different valences and the affective tenor which realises these – likely more positive than negative for elegance, overwhelmingly negative for revulsion.

The cartesian representation is adopted by Haddock and Maio (2012) to explain some unusual findings in the psychological evidence. A scale may run below the axis and become negative, so why is a negative evaluation not simply measured as a lower point on the same scale as the positive evaluation? Firstly, we find that the same attitude may have both positive and negative characteristics which are only elicited under specific conditions. This means that if we treat the valence of an attitude as a summary, or overall, evaluation we would identify attitudes with both highly positive and highly negative valences with those with both low positive and negative valences. Thus, the cartesian approach offers an explanation of the divergent behavioural patterns that mark these distinct attitudes, which the summary approach does not.

Finally, attitudes are formed in order to meet a variety of needs. Some psychologists categorise attitudes by their function, that is, by the psychological need which is met by the formation of the attitude. Tanesini emphasises that not all psychologists share this approach but adopts it because of how function relates to the epistemic standing of such attitudes, thus making the information pertinent to her overall project. The six functions[49] Tanesini identifies as having relatively wide acceptance are:

[49] Haddock and Maio (2012; 2015), as well as Tanesini (forthcoming), highlight that attitudes may have and perform more than one of these functions and this list is not taken to be exhaustive of the functions an attitude may have or perform.

1. Object appraisal; the need to evaluate.
2. Knowledge; "the need to make sense of the world" (Tanesini, forthcoming, p. 14) which may be expressed motivationally. E.g. "If having an accurate account of the target guides the formation and revision of an attitude, that attitude is said to have a knowledge function." (Tanesini, forthcoming, p. 9)
3. Instrumental; the need to maximise hedonistic reward.
4. Ego-defensive; formed in response to a need to defend the ego against real or presumed threats.
5. Social adjustive; The need to belong to one's elective social group.
6. Value expressive; the need to give expression to one's values.

Attitudes may be distinguished from one another on the basis of their performing, or emerging from, different functions. For example, Sam's attitude that "salty food is revolting" was primarily formed in response to their object appraisal function: in response to their need to evaluate salty food, they formed this attitude. Jordan's attitude that "salty food is revolting" was formed in response to their need to fit in with a particular social clique. Their attitude is primarily formed in response to their social-adjustive function. The behaviour which is elicited by the access of these attitudes will be structured differently, and be sensitive to different features of the situation, than one another. We might expect that Sam's behaviour toward salty food will be insensitive to their company, while Jordans may be somewhat more sensitive to company. Alternatively we mat posit that their attitudes will be changed, strengthened or undermined by different interventions.

The relationships between these features of attitudes imply something like the following model:

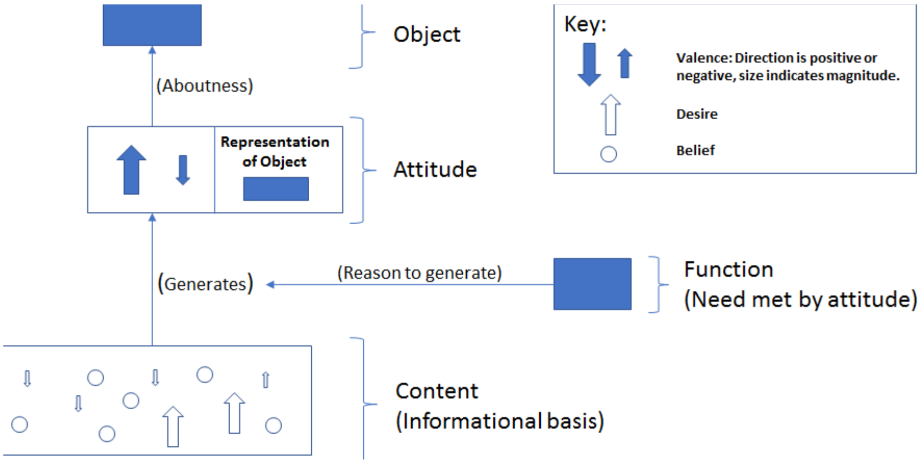

Figure 5.1: The mainstream model.

As a summary evaluation of an object, the attitude possesses an object, which is what the attitude is about or towards, and an evaluation, a positive and negative valence. The informational basis consists in a collection of behavioural dispositions, conative states and cognitive states relating to the object and the attitude is generated from it. For Tanesini, the generation of the attitude proceeds causally from the function that this attitude serves; the need the generation of the attitude meets.

Each of these features is important to the individuation of attitudes. The object and valences of attitude strait-forwardly individuate attitudes from one another, but the informational base and function of attitudes also inform the evaluation of the attitude by determining the scope of behaviour over which the attitude will be influential, as well as the manner in which the attitude may be updated.

Tanesini supplements the discussion of the features of attitudes with a discussion of the key properties of attitudes. These properties are four kinds of strength: as accessibility, as attitude extremity, as centrality to self-conception, and as certainty.

Strength as *accessibility* is the potential for an attitude to influence behaviour. This type of strength is that of the associative link between the representation of the object and the attitude's valence. The stronger the association, the more likely it is that an input which triggers the representation of the object will also trigger the valence (Tanesini, forthcoming, P. 10.; Fazio et al., 1986). This is often understood as the speed of a given association being greater, which enables it to become influential more regularly with less opportunity for confounding factors to come into play. The understanding of these associations having a speed is compatible with, but does not entail, a two-systems approach to attitudes; where some attitudes are slow-system while others are fast-system. It is sufficient for the mainstream account that speed is available on a scale. The greater the speed of the association, and hence the potential for access and influence over behaviour, the stronger an attitude is in terms of accessibility.

Strength as *attitude extremity* refers to the severity of the valence of the attitude. The greater the severity of the valence the greater the attitude′s strength as extremity. This influences the effect that the access of an attitude will have on subsequent behaviour as well as how the attitude responds to an impetus to update. Sam′s attitude that ″salty food is revolting″ has relatively high attitude extremity – revulsion is a relatively extreme affective response. Were Sam to have the attitude that ″salty food is mildly unpleasant″ the attitude would be far less extreme. This is relevant for the mainstream account because we would expect, and we find, that these attitudes have very different behavioural responses to the relevant stimuli.

Strength as *centrality to self-conception* refers to how some attitudes are more closely tied to a person′s self-conception than others. This changes how such attitudes influence behaviour and how they may be changed over time. An attitude that is entirely central to someone′s self-conception will likely be resistant to change and subject to reinforcement mechanisms, in this sense it is said to be strong. Sam′s attitude toward salty food may be entirely peripheral to their self-conception, or it may be a central feature of their self-understanding. These cases may have overlapping, but different, sets of stimuli which will elicit responses as well as different responses which may be elicited.

Strength as *certainty* is the commitment the individual has to their attitude, this degree of commitment determines some of the attitude′s interaction properties, such as the role it forms in knowledge acquisition and how it is changed. How we might understand this dimension of strength is discussed in greater length in section 5.2. Sam has no doubt about their attitude toward salty food. In this sense they might be said to be certain in their attitude. Suppose several decades went by without Sam trying any salty food. Some parts of the content of the attitude have declined in importance and Sam′s attitude has not been reinforced or had a substantial test. They may well still hold their attitude, with the same extremity and other properties, but when asked about their attitude now, in light of the time since they last tried salty food, they may be less certain of their attitude.

The final aspect of attitudes Tanesini raises is the two types of measures of attitudes. Explicit measures are those where the subject is asked to give explicit responses to questions about their attitudes towards an object. Examples of this type include interviews or questionnaires. They prompt a participant with an attitude-relevant query and record and interpret the articulated responses of the participant to the prompt. Implicit measures are those where the attitude of the subject is inferred from some other measure or response, which is taken to measure the attitude indirectly and often seeks to do so without the subject′s deliberative awareness. They prompt a participant with an attitude-relevant circumstance or procedure and record and interpret the non-articulated behavioural responses (broadly construed) of the participant. Tanesini highlights that, while the two types of measure can dissociate, it does not follow from this that the measures ″tap into different constructs″ (Tanesini, forthcoming, p. 11). That is, the mainstream model is not committed to explicit attitudes and implicit attitudes which are each independently accessed by the different types of measure.

Section 5.2: Two *Prima Facie* challenges

This section introduces two prima facie challenges to the mainstream model. The first is that certainty, as a property of the attitude, is construed ambiguously in a way which makes its interpretation problematic. The second is that the resources to distinguish between those attitudes which will be influential under explicit measurement and those which will be influential under implicit measurement, when these measures diverge, are not internal to the mainstream model.

Strength as *certainty* is the commitment the individual has to their attitude. Tanesini offers two distinct notions of attitude certainty; ″The first is clarity which measures the subject′s certainty that a statement expresses her attitude. The second is correctness that refers to the subject′s certainty that her attitude is accurate or correct (Petrocelli et al., 2007). Attitude certainty as correctness is opposed to feelings of doubt about the rightness or truth of one′s attitude.″ (forthcoming, P. 10) This ambiguity runs deeper than Tanesini alludes to.

The certainty as clarity seems not to represent a property of the attitude at all but is rather a commentary on the content of a belief which has the attitude as its object. Meanwhile, certainty as correctness seems unable to distinguish between the unexamined attitude and the examined attitude that is not doubted. These, it seems, have very different strengths and might be changed with greater or lesser ease, yet neither, at the moment, hold any doubts about the incorrectness of their attitude. Insofar as correctness is defined in opposition to doubts, these very different attitudes will be conflated on this dimension. Furthermore, in experimental terms when evaluating a person′s certainty about a given attitude the question ″How certain are you in your attitude X, of valence V toward object O?″ remains open to multiple interpretations by the subject. How the data breaks down numerically along the different lines of interpretation is not clear, or well investigated.

This criticism amounts to a two-tier challenge of definition. The first tier is that attitude psychology must explore how subjects interpret the question of certainty and whether, when regressed upon other measures, these interpretations appear to be identifying a distinct property. The second tier is that of a dilemma where the mainstream model must commit to an explication of attitude certainty in an ontological sense: what is certainty a property of? For example, attitude certainty may describe counterfactual properties: given the chance to examine a given attitude the subject would not doubt its truth. Alternatively, they may claim that the examined and undoubted attitude is the ʹstrongʹ attitude since it has the relevant property of being difficult to change. In this latter case it seems that certainty is a property of a belief about the attitude that acts as a reinforcement mechanism to the attitude itself. These are only examples of plausible commitments that could be made and remain far from exhaustive.

While there are several issues at play in this challenge, it appears to be soluble: with careful conceptual analysis and further empirical study it could be clarified within the context of the model. It is raised as an issue in the hope that this solution will be forthcoming through empirical analysis of the interpretation of such questions by subjects and the commitment of experimenters to an understanding of the model that removes the ambiguity of certainty.

The second challenge is the more problematic. Under some conditions, variance occurs between implicit and explicit measures even when these measures purport to be targeting the same attitude. Since implicitly and explicitly measured attitudes may share an object of evaluation, we may not explain a divergence on this basis. The mainstream model outlines six core functions which may be served by the formation of an attitude. Any of these functions may equally apply to attitudes accessed by implicit or explicit measures and as such the divergence cannot be explained in terms of attitudes serving different functions. Similarly, there is no resource to divide the informational basis between the implicit and explicit within the model. If one were to do so, this appears to commit a division of cognitive systems unsupported by the data. An explanation for the divergence does not lie in

The four kinds of strength offer the remaining conceptual resources wherein or whereof a distinction may be drawn. An implicit association test (IAT) implicitly identifies an association between two concepts by measuring the difference in response speed over a large number of rapid sorting problems. As a result, an attitude detected by an implicit association test must have a strong connection between its object and its valence in order to ensure a high enough probability of delay that it becomes statistically significant. Thus, implicit measures must pick up on attitudes with high accessibility. However, attitudes with high accessibility are also taken to be those which are routinely identified over longitudinal explicit measures of attitude (Holland et al. 2002). As such, an attitude may be strong in this sense, and remain consistently accessible by both kinds of measure. Similarly, both extreme and mild attitudes may be detected by both explicit and implicit measures.

Differences in centrality offer no solutions either. Tanesini references two important cases which are inversions of one another. One where the individuals have high explicitly measured self-esteem but low implicitly measured self-esteem, the other where the individuals have low explicitly measured self-esteem but high implicitly measured self-esteem (Tanesini, forthcoming, p. 19). Here is a prime example of an attitude being closely related to self-conception yet, the strength of that attitude may be measured with either direction showing the stronger response under each measure. Finally, following my criticism of the conception of certainty raised earlier, I do not believe there is any scope for the distinction to be explained in terms of high or low certainty of either kind, assuming that there are two kinds and assuming it is indeed a property of the base attitude. As a result, taking the model as presented by Tanesini (forthcoming), I conclude that the resources for explaining the divergence between implicit and explicit measures are not internal to the mainstream model.

The implicit/explicit distinction challenge is in some ways less tractable than that of the ambiguity of certainty. Because the resources are not currently available, new resources must be introduced or imported from elsewhere. As such, it is not a simple task of conceptual fiat or empirical study.

Section 5.3: The MODE-Mainstream Model

A response that would be open to Tanesini and which, in correspondence, she has inclined to take, is to incorporate the above framework in an overarching MODE model following Fazio (et al., 1986; 1990; Fazio and Olsen, 2014). This section presents the response in a form designed to answer the above challenge. The form of this overarching model is heavily informed and inspired by discussion with Tanesini but does not follow from a particular work of hers. As a result, insofar as the response works, the credit is Tanesini's and insofar as it does not, or misrepresents her view, the blame is mine.

Fazio's MODE model aims to describe the processes by which attitudes affect judgments and behaviour (Fazio, 1990). The model outlines a distinction between spontaneous and deliberative processes of attitude-behaviour association differentiating the two in terms of motivation and opportunity as determinants of which process will be dominant (Fazio and Olsen, 2014). This two-systems approach offers an explanation of the divergence between the findings of implicit and explicit measures as being one of the lack or presence of a motivation to engage critical assessment in conjunction with the lack or presence of an opportunity to do so.

MODE may be modelled (in a simplified form) as the following decision tree (Fazio, 1990; Fazio and Olsen, 2014):

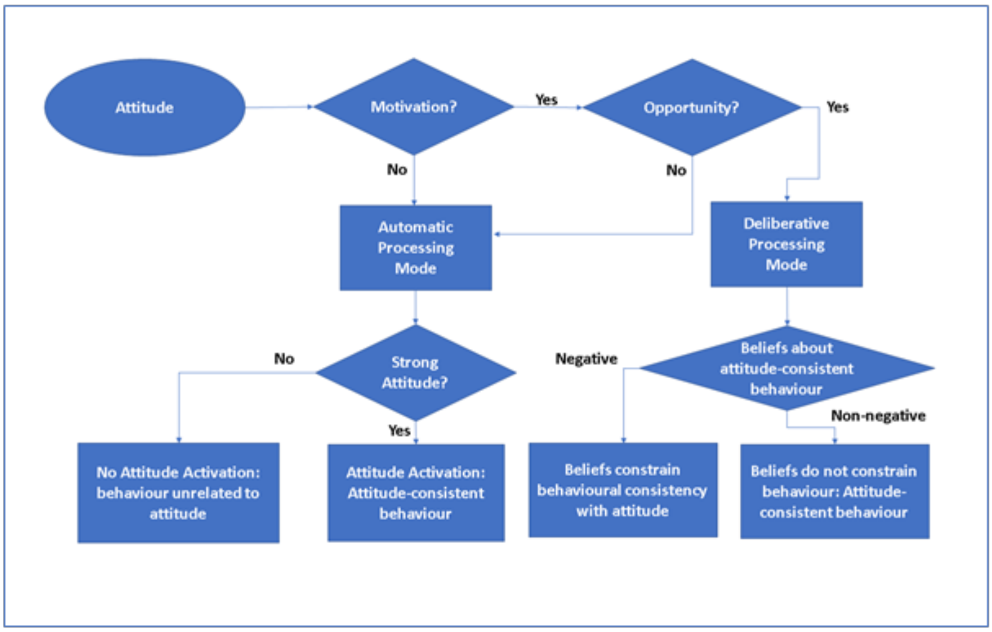
Attitude
Motivation?
Yes
Opportunity?
Yes
No
No
Automatic Processing Mode
Deliberative Processing Mode
No
Strong Attitude?
Negative
Beliefs about attitude-consistent behaviour
Non-negative
Yes
No Attitude Activation: behaviour unrelated to attitude
Attitude Activation: Attitude-consistent behaviour
Beliefs constrain behavioural consistency with attitude
Beliefs do not constrain behaviour: Attitude-consistent behaviour

Figure 5.2: The MODE Model.

MODE explains the divergence between implicit and explicit measures in terms of the presence or absence of the motivation and opportunity to deliberate about the behaviour. In order to acquire the resources to make the distinction between implicit and explicit measures in the mainstream model, we may situate that model within MODE at the ʻattitudeʼ step in the decision tree. This offers the following overall model:

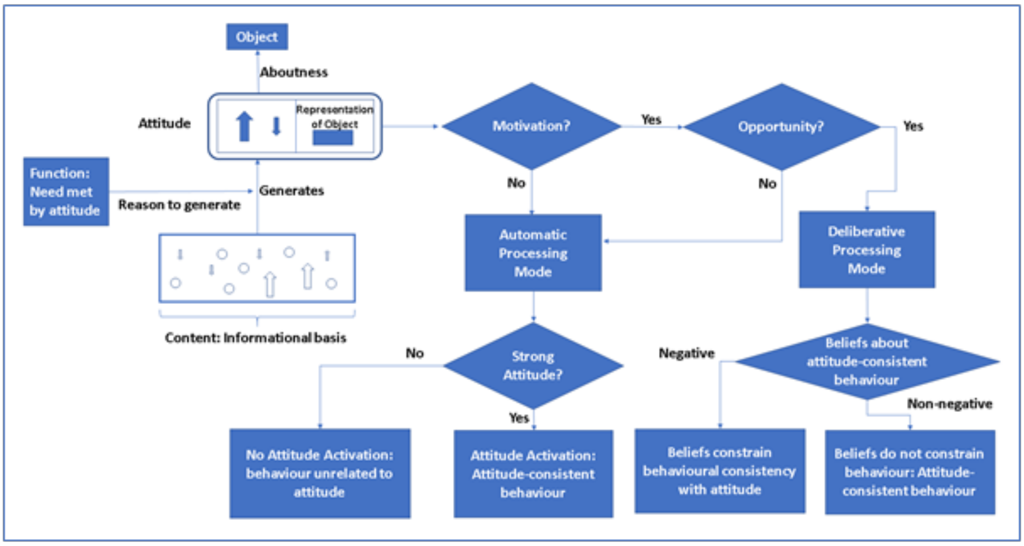
Object
Aboutness
Attitude
Representation of Object
Motivation?
Yes
Opportunity?
Yes
Function: Need met by attitude
Reason to generate
Generates
No
No
Automatic Processing Mode
Deliberative Processing Mode
Content: Informational basis
No
Strong Attitude?
Negative
Beliefs about attitude-consistent behaviour
Non-negative
Yes
No Attitude Activation: behaviour unrelated to attitude
Attitude Activation: Attitude-consistent behaviour
Beliefs constrain behavioural consistency with attitude
Beliefs do not constrain behaviour: Attitude-consistent behaviour

Figure 5.3: The MODE-Mainstream Model.

This overall model imports the ability of MODE to answer the divergence question. Under laboratory conditions it is possible to substantially remove the motivation and opportunity of individuals to deliberate about their behavioural options. If motivation and opportunity to deliberate are both present then, negative beliefs about attitude-consistent behaviour can constrain that behaviour. In the absence of motivation and opportunity to deliberate, a strong attitude will generate attitude consistent behaviour. These are the relevant divergent cases.

The MODE-mainstream model can also explain the typical case of consistent measurement between explicit and implicit measures. In these cases, on this account, there are no negative beliefs about attitude-consistent behaviour which would constrain such behaviour when motivation and opportunity to deliberate are present. There is also scope within the account to explain the, sometimes noisy, findings of these measures, since in the absence of motivation or opportunity for deliberation, the absence of a strong attitude will lead to behaviour that is not systematically related to any given attitude. This does not necessarily mean that the behaviour will be unsystematic, but it may be unsystematic qua a given valence toward a given target object.

By situating the mainstream model within MODE the conceptual resources to explain the divergent cases between explicit and implicit measures are incorporated with relatively little loss in simplicity.

Section 5.4: A *secunda facie* challenge

This simple integration of the models raises a new concern. There appear to be entities which play multiple roles at different points of the model. Beliefs about the appropriate behavioural responses to the object of the attitude appear in both the informational base and in the moderation mechanism of the attitude under deliberation. This duplication of entities has some problematic consequences if it is more than apparent.

Suppose this duplication is indeed more than merely apparent. In those instances where a subject has a belief that a given behaviour is prohibited or taboo, that belief, being relevant to the object of the attitude, would constitute part of the informational basis of the attitude. In cases where that belief would have a substantial influence on decision-making we would expect that same belief to have played a significant part in the formation of the attitude prior to having any effect on behaviour. One of the functions highlighted by the mainstream model is the ″social adjustive″ (Tanesini, forthcoming, p. 22) and in cases where the social sanction for such behaviour is perceived to be significant this function would generate an attitude that took such beliefs about behaviour into account. Furthermore, in instances where there exists a belief that a given behaviour is socially prohibited, we may presume that this belief constitutes a standing motivation not to engage in such behaviour in social situations. This is an example where a single entity seems to play multiple roles, in a way that generates tensions in how we interpret the interesting cases.

In a case where a belief that behaviour is socially prohibited is observed in a subject and the results of implicit and explicit measures of the subjects attitude toward the relevant object diverge. The implicit measure expresses a socially prohibited outcome, and the explicit measure conforms to the social prohibition. Should it be concluded that (1) the subject has a strong attitude that does not become evident in behaviour because of their beliefs about it being socially prohibited, or that (2) the subject does not possess a strong attitude on the matter at all, hence the lack of correlation between attitude and the behavioural outcomes, or that (3) they have an attitude that aligns with their behaviour and the social prohibition but that is not sufficiently strong to bear out in implicit measures? This sort of case, precisely the one that much of the implicit bias literature is interested in, seems indeterminate under this description of the model. Rather than have no answer to how such findings diverge we are left with too many for the interesting cases due to the duplication within the model.

Section 5.5: The construction of explanations

As indicated by term of address of the above challenge, I believe it to be resolvable for the MODE-mainstream model, and that the resolution lies in the clarification of its ontological commitments. This section demonstrates this resolution by applying the ontological critique to the MODE-mainstream model.

The ontological critique presents five challenges to a model that clarify the ontological status of the models claims as well as the specific commitments of the model. The first challenge establishes and clarifies those things which are valuable to the project, which become the sole resource to which arguments about a model′s success or failure may make recourse.

Within the field of psychology more generally, we find abstract concerns such as accuracy, precision, scope, simplicity, fruitfulness, coherence, etc. Within attitude psychology more specifically we may elaborate on accuracy as offering explanations of the relevant phenomena (priming effects, implicit/explicit divergence/existence and change of behavioural patterns/beliefs about those behavioural patterns and how they misfire/etc.). We may elaborate on fruitfulness as offering predictions about the conditions under which particular effects may be elicited or confounded. We may further identify more overarching concerns such as diversity in the research community or the accessibility of results. For specific projects seeking to apply a given model within the research programme of attitude psychology, we may find more specific concerns still such as an account that coheres with existing conceptions of motivation, beliefs and desires, or an account that renders attitudes as plausible candidates to be the psychological grounds of virtues or vices.

In the case of the MODE-mainstream model, researchers are particularly interested in the explanation of patterns of behaviour towards objects of evaluation and the explanation of cases in which patterns obtain and cases in which they are broken. The explanation of these patterns should also give rise to predictions about the kinds of intervention that will be efficacious in interrupting or altering these patterns. The model should also cohere with broader understandings of cognition such that evidence from other, closely related research, may be brought to bear.

The second challenge asks what the objects of the model are; what must exist if the model represents a psychological reality? We find the objects that are the core contributions of the model; either novel posits or pre-exiting posits, novelly expounded. These are the attitude itself, the content of the attitude or its informational base, the function of the attitude, the valence(s) of the attitude, the object of the attitude, the representation of the object within the attitude, the behaviour of the subject, and the situation.

The third challenge seeks to clarify the properties of these objects. The content is the sum of the relevant cognitive, conative, and behavioural states of the person, out of which the attitude is generated. They remain party to deliberation and, given motivation to deliberate and the opportunity to do so, will be influential in the behavioural outcome. In the absence of the motivation or opportunity to deliberate, they are not influential in the behavioural outcome directly. As such they retain the powers normally ascribed to beliefs and desires in conjunction with a constraint on when deliberation takes place, but under this model are further ascribed the ability to act as the formational materials out of which and according to which the attitude is formulated and ′lose′ the capacity to influence high-speed processes. The object of an attitude is that which an attitude is about or toward, and as such almost anything can be the object of an attitude. The object of an attitude defines and constrains the set of external stimuli that will trigger the attitude. The valence(s) of an attitude are the summary evaluation of the object of the attitude derived from the informational base. The valence of the attitude codes the kinds of behavioural responses that the person will tend to give in response to the relevant stimuli given the attitude′s influence on behaviour. The function of an attitude is the need of the person which the attitude meets. It provides the impetus to form the attitude from the informational base as well as structuring the object and valence of the attitude in particular ways to be sensitive or insensitive to particular kinds of inputs. This gives function an aetiological role in the formation of attitudes as well as a role in individuating attitudes from one another. The attitude as a whole has a series of properties which Tanesini outlines as the four ways in which attitudes may be strong: accessible, extreme, central to self-conception, and certain. Further to these properties there are the properties of attitudes as the sum of its parts. It influences behaviour according to its valence following an external stimulus relevant to its object, both of which are derived from the informational base and coded by the function the attitude serves. Finally, it has its interaction properties with the peripheral objects of the model: attitudes are possessed by persons, code behaviour, in response to situations. They do so in conjunction with components of the informational base of the attitude when deliberation is undertaken and independently when deliberation is confounded. Finally, behaviour, persons and situations receive some new gloss under this model. Behaviour is the product of the content and the attitude under deliberation, or only the attitude when deliberation does not occur. Similarly, persons have motivations (both standing motivations which are internal to the person and occurrent motivations which arise from the conjunction of person and situation) which are necessary, but not sufficient, for deliberation and situations which either include or preclude the opportunity for deliberation, which is similarly necessary, but not sufficient for deliberation.

These objects with these properties are intended to give rise to the explanations that Tanesini, Haddock and Maio offer. One such explanation is that of the divergence between implicit and explicit measures offered above. We may express the explanation in terms of the ontological critique by highlighting the problem, identifying the relevant objects and properties, and identifying how the outcomes are conceptualised and driven by these objects and properties on the MODE-mainstream model. The problem highlighted was an inability to explain the divergence between implicit and explicit measures of attitude within the mainstream model. Specifically, the divergence sometimes occurs between the valence elicited by an implicit measure toward a target object and a corresponding valence elicited by an explicit measure targeting the same object. These divergences occur often enough to attract significant philosophical and psychological interest, but do not remain stable across multiple different implicit measures.

The relevant objects to the explanation of these divergences on the MODE-mainstream model are object, valence, content, behaviour, and attitude. The outcome of each measure is the behaviour elicited and the object is controlled for in each test. Valence is inferred from the behavioural response. These objects are necessary to frame the problem, while the heavy -lifting of the explanation is done by the content of the attitude and the relationship the attitude has with behaviour, given the presence or absence of particular properties of the person and situation. In the presence of the motivation (standing or occurrent) and opportunity to deliberate, the content of the attitude will take precedence over attitudes. When either the motivation or the opportunity to deliberate is absent attitudes inform behaviour without influence from the content the attitude was formed out of.

By drawing the distinction between those cases in which deliberation does and does not occur, the model is able to explain both the divergent minority of cases and the convergent majority, since we would expect that attitudes on the whole converge with their informational base.

What this application of the first four challenges of the critique affords us is a systematic approach to respond to the *secunda facie* challenges. While the first presentation of the MODE-mainstream model simply inserted the mainstream model into the box labelled ′attitude′ in the decision tree of the MODE model, by applying the ontological critique to the model as a whole the two models are properly integrated. This is done by recognising that the

MODE model describes further properties of entities already party to the mainstream model. This clarification, highlighted in response to the third challenge, dissolves the *secunda facie* challenge of duplication. The two models were not duplicating entities, but rather talking about the same entities possessing properties and capacities which become apparent under different circumstances.

The dissolution of the *secunda facie* challenge offers a further consideration. The presentation of the mainstream model invited the problem of explaining the implicit/explicit prediction. This problem has a solution, available in the current literature and known to researchers who work with the mainstream model. However, the approach taken by Tanesini to identify the features and properties of attitudes fails to draw out the importance of the situation of the mainstream model in MODE for offering its target explanations. Furthermore, the integration of the two models requires careful parsing of the objects and properties of the composite model. This process of criticism and response illustrates the advantage of implementing the ontological critique over the approach adopted by Tanesini. Explicitly connecting our communication of the mechanics of the model to the explanations it offers, and thereby to its place in a progressive research programme, our communication is made more precise, and our models are made more responsive to challenges.

Section 5.6: Predictions and MODE-mainstream

The final challenge that the ontological critique presents is to clarify what it would take to falsify the model and, if this were to occur, what would be at stake. These predictions can be identified by clarifying the populations over which the model expects phenomena to extend.

One example of a potential falsification would be an experiment which could reliably elicit substantive deliberation in subjects in the absence of either motivation to do so, opportunity to do so, or both. If this were to happen, what would seem to be at stake is the commitment to MODE. Since this is a central commitment of MODE-mainstream, it is likely that the auxiliary hypotheses warranting the inclusion of this experiment as relevant to MODE-mainstream would bear the brunt of such a falsification. This would be an extreme case, requiring a direct contradiction of a central tenet of the model.

A less extreme case would be evidence of longitudinal instability of implicit measures of a given attitude. That is, where the implicit measures accessing a given attitude fail to correlate with one another over time. This presents a problem for the model, but one which could be accommodated by altering auxiliary hypotheses. The MODE-mainstream model does not currently directly include an account of changing attitudes which may be sufficient to account for the instability. If the instability is greater than what could be accounted for by simple change, then the model could instead situate attitudes in relation to other psychological phenomena accessed by implicit measures, especially for objects for which a formed attitude is not apparent. The availability of a variety of responses to such a challenge illustrates how the nature of failed predictions underdetermine theory development. Schimmack (2020) presents some evidence that this may be the case for the IAT, but highlights the need for further longitudinal analysis of these effects making it too soon to say whether these amendments to auxiliary hypotheses are avenues which MODE-mainstream need pursue.

Another potential falsification, raised in the previous chapter, would be evidence of substantive divergence between different implicit measures toward a single object. Nosek et al. (2007) presents some evidence that this may be the case. If this is the case then the MODE-mainstream model needs some way to account for how this divergence could occur. As with the first example, it is likely that the auxiliary hypotheses which make this a problematic result for the model will bear the brunt of the falsification. Unlike the previous example, it is unclear precisely how auxiliary hypotheses may be updated or introduced which would mitigate the problem of these results for the core commitments.

In each of these cases, at least one aspect of the model would be called into question and, if the finding is reliable, a non-trivial response to the challenge would be required from the proponents of the model. This list is far from exhaustive but demonstrates some key techniques in identifying potential problems.

Section 5.7: Concluding remarks

The ontological critique is a powerful toolkit for identifying and clarifying certain kinds of conceptual challenge. In this case, the ontological toolkit highlights where models appear to be talking cross-purposes in a manner which could plausibly have been deeply problematic. By rigorous application of the toolkit to the model, the commitments of the model are identified and their relations to one another in building explanations of relevant phenomena are outlined. In the case of the apparent duplication, we have a complete resolution of the problem by exhaustively identifying the objects of the model and their attendant properties. In the case of the implicit/explicit divergence the existing explanation could be further clarified, such that the decision tree of MODE not only explained what made each case what it was, but the model as a whole explains why the divergent cases are in the minority. By linking negative beliefs about attitude-consistent behaviour in the content that the attitude is generated out of to the operation of attitudes in the presence of motivation and opportunity for deliberation, the model explains why divergent cases are minority cases through the dual function of these beliefs.

While this application illustrates the utility of the ontological critique, it also highlights the kinds of cases which will remain intractable to the critique. In particular, the first *prima facie* challenge still stands – the ambiguity of certainty remains. This ambiguity is intractable to the critique since it requires commitment on the behalf of MODE-mainstream researchers to a given understanding of attitude certainty, likely informed by further research into how subjects understand the questions posed to them to gauge their certainty in explicit measures.

The point to note about the ontological critique is its advantage over the approach adopted by Tanesini. By leveraging the construction of explanations as part of the communication of the mechanics of our models, the critique identifies the relevant parts of the model and outlines the resulting properties. This offers the same affordances of Tanesini's approach, while going beyond these parts and properties to linking them to the explanations and predictions which the research programme is directed toward. What the *prima facie* and *secunda facie* challenges represent, in effect, is the advantages of the ontological critique over enumerating the features and properties of the model.

For the MODE-mainstream model, we may conclude firstly that it presents plausible and interesting responses to the apparent challenges it faces and secondly that under the scrutiny of the ontological critique the model clearly illustrates the construction of its target explanations. Third and finally, the responses to the falsification challenge highlights the kinds of challenges which may be set to the model by future experiments as well as potential problems present in the existing literature. Researchers interested in testing and refining the model will likely find fruitful, challenging avenues by designing experiments to test these possibilities and by pushing to refine the MODE-mainstream model in response to these challenges.

Chapter 6: Challenging CAPS

Section 6.0: Introduction

The previous chapter engaged with the mainstream model of attitudes (Haddock and Maio, 2012; 2015; Tanesini, forthcoming). That model presents a functional account of attitudes in terms of the aetiology of those attitudes and the conditions of their implementation in behaviour. This chapter analyses the alternative functional account within the literature: the cognitive-affective personality system (Mischel and Shoda, 1995).

The Cognitive-Affective Personality System (CAPS) is a leading theory in personality psychology, as well as a prime candidate for a model of attitudes in attitude psychology. It has also been deployed to ground the psychological aspect of virtue ethics, through modelling attitudes. Developed by Mischel and Shoda (1995), CAPS is among the most influential and well elaborated psychological proposals for a model of personality. In recent years a great deal of interest has been taken in it by philosophers interested in virtue ethics and virtue epistemology. In particular, CAPS has been used to respond to the situationist challenge to the stable occurrence of virtue, or of broad-track character traits more generally.[50] Much of the contemporary debate about CAPS has centred on its empirical validity. [51] This chapter offers an orthogonal approach to analysing CAPS to these by subjecting it to the challenges of the ontological critique outlined in chapter 2.

Section 6.1 introduces CAPS, as presented by Mischel and Shoda, as well as the core problems it is intended to resolve in personality psychology generally, and in attitude psychology in particular. This section identifies and defines the model′s key postulates and how they are intended to operate. In section 6.2, CAPS as a model of attitudes is subjected to the first four

[50] For the situationist challenge, see Harman (1999, 2000), and Doris (1998, 2002). For responses grounded in CAPS, see Miller (2003), Snow (2009: chapter 1), Webber (2015, 2016), and West (2018).

[51] For the objection to CAPS′ empirical validity, see Doris (2002, pp. 76-80); Miller (2003); Railton (2011); Alfano (2013, pp. 78-79); and Miller (2016). For the response, see West (2018).

challenges of the ontological critique which open the key components of the explanations offered to clearer and more precise scrutiny. This application raises questions about the grounding of the mechanics of CAPS, specifically, it is presumed that the explanations offered by CAPS can be grounded in some form of graph or systems theory. Two candidates for such grounding will be introduced and challenged in the following sections. The first of these candidates, dynamical systems theory, is presented in section 6.3 along with the key challenge such a grounding faces. The explanations offered by CAPS are constructed out of the cognitive-affective units, or attitudes, that constitute the system and analysis in terms of dynamical systems theory does not provide clear warrant for these explanations. Sections 6.4 and 6.5 introduce the other candidate grounding for CAPS, connectionism. Section 6. 4 introduces feed-forward networks which, while implausible as grounds for CAPS on their own, facilitate a precise discussion of the more complex continuous flow networks in section 6.5. Connectionism of both kinds avoids the challenge presented against dynamical systems theory, but each presents its own shortcomings as grounds of CAPS, especially how to ground the explanations of character development that CAPS offers as one of its key aims. Section 6.6 summarises and clarifies the problem presented by these shortcomings for CAPS both *per se* and as a model of attitudes. This summary of the problems leads to the discussion of the final challenge of the ontological critique: prediction. Some general target predictions for CAPS are set out as well as how these relate to the challenges raised in the preceding section. The chapter closes with a call for further critical engagement from CAPS theorists both to refine the theoretical commitments of CAPS in abstraction and to present these as well-defined competing interpretations within the CAPS research programme which may be empirically falsified or vindicated.

Section 6.1: Introducing CAPS

The paper in which Mischel and Shoda introduced CAPS begins with some of the oldest questions in personality studies: how do personality structures interact with features of their situation and how can personality be conceptualised such that it retains general stability, and yet retains plasticity? Their proposal is to model personality on connectionist systems in a way that explains behavioural consistency despite situational variance and explains behavioural change.

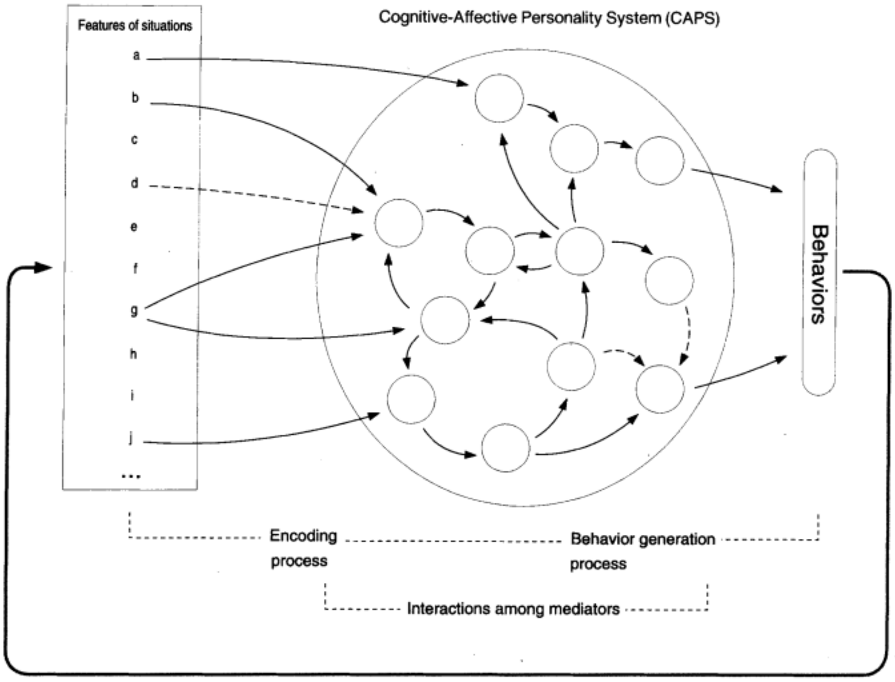
Features of situations
a
b
c
d
e
f
g
h
i
j
...
Cognitive-Affective Personality System (CAPS)
Behaviors
Encoding process
Behavior generation process
Interactions among mediators

Figure 6.1: The CAPS Model. (Mischel and Shoda, 1995, p. 254.)

Mischel and Shoda describe the nodes of the CAPS network as ′cognitive-affective units′, which include ′encodings and affects... expectancies, goals, behavioral scripts and plans′ (1995, pp. 246). I will refer to these collectively as ′mental states′ for the sake of simplicity (following Webber 2015, pp. 1092; 2016, pp. 137-138).

CAPS explains behavioural consistency in terms of the organisation of these mental states, positing that an individual′s mental states are interconnected in a network across which activity flows according to the strength of the connections (Figure 6.1, above). For each connection, activity flows from one node to another; I shall refer to a node as the ʹeffectorʹ when activity flows from it and ʹreceptorʹ when activity flows to it. This network structure has connections to features of the external world. An identical CAPS will produce the same behavioural output from the same situational input.

Holding constant the features of the situation and the structure of the CAPS system, CAPS offers an explanation of behavioural consistency. But because of the internal structure of CAPS, the pattern of activation, and hence the behaviour elicited, will depend significantly on the situational input. Therefore, even minor changes in the situational input cause a different pattern of activation, which will follow different paths across the system. This allows for even small variations in the situation to produce, from the same system, significantly different behaviours. CAPS is therefore designed to explain the empirical evidence of behavioural variation that the situationist challenge to virtue ethics relies on (Harman 1999, 2000; Doris 1998, 2002).

The second problem CAPS seeks to address is that of behavioural change. Specifically, how is it that repetition of a given mental connection causes that mental connection to become stronger, or habituated? CAPS answers this by ascribing the property of strength to each connection. This property determines the way in which the effector and receptor are connected. This strength may be conceived of in several ways. What the different conceptions have in common is that they describe an operator which mediates between the effector and receptor. Given the effectorʹs activation, the strength of the connection determines either the probability of activation, or the degree of activation, of the receptor. The properties of these connections will be discussed in more detail in section 2. With connection strength, it becomes possible to posit that the activation of effector, connection, and receptor cause that connection to strengthen, meaning that, given any future activation of the effector, the receptor has a greater chance of activation or is activated to a greater degree. Thus, over time, the more a given connection is activated the stronger it becomes. This strengthening of connections, this habituation, is what allows CAPS to explain personality change as a gradual process which does not undermine the overall relative stability of the personality system (Mischel and Shoda, 1995: p. 256).

In summary, by positing the existence of connections between mental states which strengthen with activation, CAPS explains behavioural consistency as emerging from similar input conditions following similar mental pathways of cognitive and affective mental states. The cognition that generates behaviour emerges from the activation of these pathways. Since the connections in this system strengthen slowly with activation, this allows an explanation of relative behavioural consistency over time. It simultaneously allows for the possibility of seemingly small changes in the environmental input generating substantially divergent behaviour, since these differences in input may change the pathways activated and thus the behaviour elicited.

By framing the CAPS model in its own terms, we are presented with a plausible picture of the mechanics of these facets of a personality. The next section re-presents CAPS through the first four challenges of the ontological critique, highlighting ambiguity in areas of the model which are usually glossed in prima facie plausibility.

Section 6.2: Ontological Critique.

This section frames CAPS through the first four challenges of the ontological critique. These challenges are:

1. Axiology: What are the standards for this being a successful, or good, model?
2. Objects: What are the objects of the model?
3. Properties: What are the properties of these objects?

4. Explanation: How do these objects and properties give rise to the explanations offered by the model?

Beginning with the axiology challenge we may ask what matters to CAPS? What are the excellences or deficiencies to which we may appeal to advance or criticise the model? This challenge affords answers at several levels; overarching concerns for the discipline, general concerns of the research programme, and specific concerns for a given application of a model.

To begin in the middle, the general concerns for the CAPS research programme were highlighted above; CAPS aims to explain general behavioural stability while accounting for plasticity over time and sensitivity to context. As such, CAPS is successful insofar as it offers explanations of observed patterns in subjects′ behaviour as well as those cases in which behavioural patterns are broken, and the change in those patterns over time. This general value of CAPS is situated within, and conforms to, overarching values within psychology; accuracy, precision, scope, simplicity, fruitfulness, coherence, etc. (Kuhn, 1977). In the context of CAPS, accuracy will mean, among other things, offering explanations of the relevant phenomena that accord with the existing evidence about behavioural signatures and their patterns of change. Fruitfulness will require the model to offer predictions about the conditions under which particular effects may be elicited or confounded. Similarly, scope will be obtained if the model is capable of offering accurate explanations of a wide variety of phenomena. Meeting these demands while remaining no more complex than necessary, and offering the greatest utility to those who must apply it, would afford the model the value of simplicity. We should further identify the overarching concerns of promoting diversity in the research community and the accessibility of results.

Specific projects seeking to apply CAPS offer more specific concerns such as the desire for an account that coheres with existing conceptions of motivation, beliefs and desires, or an account that is compatible with, or explanatory of, related work in ethics or epistemology. The implementability of interventions as well as the availability of clear tools for interpretation are also project-specific concerns. Other concerns will arise depending on the specific demands of a given project.

The complete answer to this first challenge is the set of all properties of the model to which we may make recourse in answering the various questions about how 'good' the model is. Insofar as something is a property to which we may make legitimate recourse in its criticism or defence, it is a partial answer to this challenge.

The objects challenge asks CAPS to identify its component objects. What are the entities posited by the model such that it may offer the explanations it purports to? These may simply be listed [52]:

- Situation
    - Features of a situation
    - Encodings
- Cognitive-Affective Personality System
    - Cognitive-Affective Units
    - Behaviour generating process
    - Attitudes
- Behaviour
    - Behavioural signatures
    - External feedback

The answer to the objects challenge will include all objects to which explanations will refer. This incorporates those objects which are peripheral to the model and to the understanding of which the model offers little or nothing novel, but which are necessary to express the explanations of the model. It also includes those objects which are at the heart of the model which are novel posits of the model or which are extant objects in the field,

[52] This list is broken into something like wholes and parts, though the distinction is not so formal but rather illustrative of the place of these objects in the model. It is done for the sake of clarity. Insofar as a reader finds that this obscures more than it clarifies, they should read the list on one level.

novelly expounded.

The properties challenge asks what capacities and incapacities these objects possess such that they may play their roles in the explanations offered by the model. This third challenge does not afford the brevity of the second.

Situations have the properties they are normally taken to in psychology. They are states of affairs with which persons may interact, in both directions. Situations, under CAPS, have features which are those parts of a given situation which can be salient to the Cognitive-Affective Personality System. Encodings are one of the interactions between situation and person, and describe both the tendency for salience to arise and the actual occurrence of salience in a given situation. By this I mean that, under CAPS, encodings are both the tendency to find X′s and X-like things salient, and the finding of this particular X salient[53]. These properties are largely the same as those afforded to them in other areas of psychology where they are relevant. While the explanations offered by CAPS depends on these properties, CAPS offers little novel exposition of them.

The Cognitive-Affective Personality System is the whole interconnected network which receives the encoded salient features of the situation and outputs the behaviour that arises in response to this input. This system consists in interconnected Cognitive-Affective Units, which are the set of behavioural tendencies, affective content, and cognitive states that make up the whole. These units have connectedness to other such units and degrees of influence upon subsequent units to which they are connected. Via these immediate connections the units have a degree of influence on the behavioural outcomes as a whole. These degrees of influence grow or shrink depending on the habituation of these connections and is sensitive to feedback on the behaviour generated by the system. This change over time is

[53] While little hangs on the distinction between these two, CAPS theorists make recourse to both properties in their explanations. There is not necessarily a problematic overlap between the disposition and the event, but clarity about the use of these terms is useful in identifying which is being done when we are presented with a given explanation.

a relatively slow process but can involve large changes given long enough and enough consistent feedback. Attitudes are those collections of interconnected cognitive-affective units, small or large, which share a more-or-less coherent target object. They have the properties of strength, object, valence, interconnectivity and the capacity to influence behaviour.

The behaviour generating process is the mechanism by which the output of the Cognitive-Affective Personality System becomes, or is realised in, behaviour. This process, together with the patterns of activity across the system give rise to pattens of behaviour in response to given features of a situation. These behavioural signatures make up part of the overall behaviour of the subject and respond to the salient features of the situation. This behaviour as a whole then creates a feedback loop with the situation as the subject alters their relationship with the situation through their behaviour.

The explanation challenge asks how these objects with these properties give rise to the explanations offered by the model. As highlighted in the previous section, CAPS explains behavioural consistency in terms of the organisation of mental states, positing that an individual′s mental states are interconnected in a network across which activity flows according to the strength of the connections. In terms of the objects and their properties, this means that behavioural consistency is explained in terms of relative stability in the tendency encodings, relative stability in the connections between cognitive-affective units and their influence on one another and stability in the behaviour-generating process. These stabilities, holding the relevant features of the situation fixed, bring about relative stability in behavioural signatures in response to these features.

The systematic variability within behavioural signatures is explained with reference to the variability in actual encodings, the variability within the network of cognitive-affective units and the behavioural signatures only making up the input-relevant part of the subject′s overall behaviour. These objects, with these properties, are the grounds of the CAPS response to the situationist challenge.

Finally, behavioural change is explained in the habituations of the connections between cognitive-affective units both through repeated activation and in response to external feedback. This explains how substantive change in behaviour is slow to come about, and responsive to external feedback. By ascribing the property of strength to each connection CAPS explains personality change as a gradual process which does not undermine the overall relative stability of the personality system (Mischel and Shoda, 1995: 256).

The application of the ontological critique to CAPS has the benefit of codifying our analysis of the ways in which CAPS approaches its solutions to these problems, but more importantly it also allows us to ask some precise questions about the mechanisms that constitute CAPS′ explanations. Specifically, the question the following sections shall pose is: what grounds the explanations CAPS is offering to its key questions? What gives CAPS the warrant to suggest that the interconnectivity of cognitive-affective units, or attitudes, can give rise to the sorts of stability and change which it describes?

To explore this question, the following sections explore the two most eminent candidates for such grounding: dynamical systems theory and connectionism. Both have some support in the literature, with dynamical systems theory being the candidate of choice of some philosophers engaging with CAPS[54], and connectionism which appears to be the candidate of choice of Mischel and Shoda (1995, pp. 267-8; 2006; Shoda and Smith, 2004).

[54] The proposal of which I would like to chiefly attribute to two anonymous, independent reviewers of a manuscript version of this chapter.

Section 6.3: Dynamical systems theory as a grounding for CAPS [55]

Dynamical systems theory is a field of mathematics which encompasses those systems which are both non-linear and described by time dependant rules governing the movement of points in vector space. Within this context, non-linear refers to any function which does not have the property of superposition[56]. Rules for mathematical models are generally functional expressions, and to say that a point in vector space is time dependant is to say that the function which describes its location in vector space depends in part on time. Vector space describes the dimensional region within which points exist and move. The system encompasses the complete set of functions describing the time dependant location of points within vector space, the vector space within which they exist, and the complete patterns of development of those functions and that space. The study of these systems expresses the patterns of motion across this space in terms of ordinary differential equations (ODE) or partial differential equations (PDE), expressing the whole system as a resolvable problem for some time (t_1) given relevant knowledge about a previous time (t_0).

A simple example of such a problem is the movement of a pendulum. The amplitude of angle from vertical of the pendulum can be plotted against time for a given starting value in the vector space giving a graph like the following:

[55] This section briefly summarises relevant portions of Glendinning (1994), Wiggins (2003), Lynch (2017) and especially Strogatz (2014) before applying them to the problem of CAPS as a dynamical system.

[56] Superposition can be represented as the conjunction of additivity and homogeneity. A linear function possesses both of these properties. Additivity holds for a function iff $F(n+m)=F(n)+F(m)$ and homogeneity holds iff $\alpha F(N)=F(\alpha N)$, where α is a (real or complex) number. Nonlinear functions, like those described by dynamical systems theory, do not abide by one or both of these constraints.

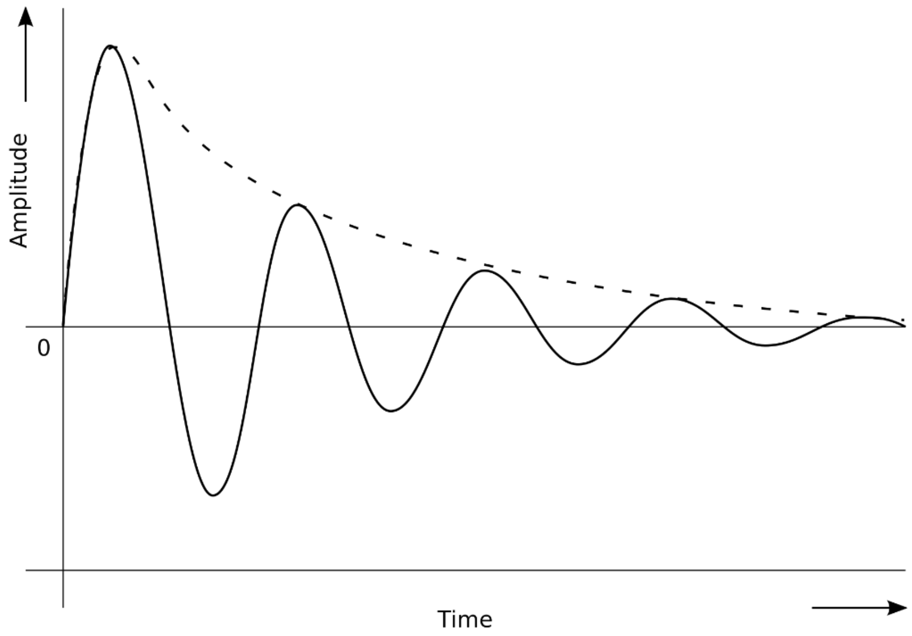

Figure 6.2: The phase-space graph of a dampened pendulum.

By expressing the movement of the pendulum in only one dimension (angle from the vertical) we are able to express its motion in a two-dimensional graph incorporating time. This is what makes the example simple in terms of graphical representation. It does not follow that it is simple in terms of order since the underlying function has terms of several orders (at least 3, including damping effects). This problem, and others like it in dynamical systems theory, can be expressed as the solution to a single overarching equation that describes the movement of a point in the vector space given knowledge about its prior position. More complex problems express the movement of points in time by representing the vectors in that space as arrows.

Analysis of dynamical systems comes in two broad kinds. The first is the analysis of the ODEs or PDEs that instantiate the system. This analysis can be useful if the equations are relatively simple, or display certain clear patterns. For the purposes of this application of dynamical systems theory however, this analysis requires substantially more familiarity with the underlying mathematics of differential equations and their behaviour. Fortunately, the second form of analysis available is the observation of the graphical representations of the system which does not require this background (Strogatz, 2014, p. 9, pp. 125-130, p. 146). This observation begins by noting the key points in the diagram and proceeds to describe the movement in planes between these points.

The kinds of point relevant to these analyses are centres, stable nodes, unstable nodes, hybrid nodes and saddle nodes. Stable nodes are points which vectors travel towards, but not away from (A, below, is a stable, spiral node). Unstable nodes are those points which vectors travel away from but not towards (B, below, is an unstable asymptote). Hybrid nodes are stable when approached from one direction by a vector, but unstable once a vector crosses the node itself[57] (C, below). Saddle nodes are points where incoming vectors divert upon or around some axis centred on the point (D, below). Centres are points around which vectors orbit stably (E, below).

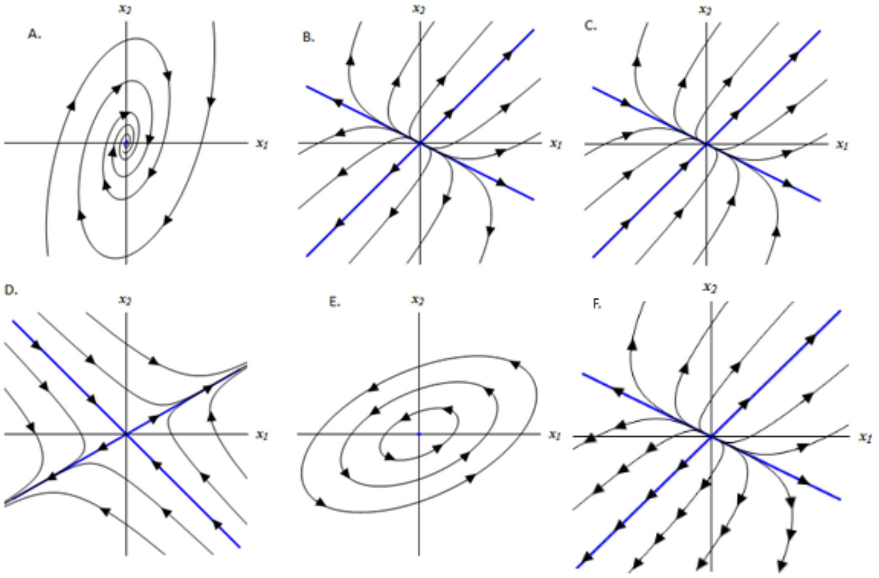

Figure 6.3: Five types of node (A. Stable, Spiral; B. Unstable; C. Hybrid, asymptotic; D. Saddle; E. Cycle; F. Unstable node with varying amplitude and volatility.)

[57] If the stability of the node from the opposite direction is asymptotic, the vector never actually ʻmakes itʼ across the node to the unstable side.

Once we can describe these varieties of point, we may describe the space between them. This has two key dimensions for the purposes of our analysis. The first is the amplitude and the second is the volatility. Amplitude describes the rate at which a vector will move across that part of space. In figure 6.3: F., we have an unstable node at the origin with areas of high amplitude above and to the right of this point, and areas of low amplitude below and to the left. Volatility describes the sensitivity of the eventual path of the vector on its exact location in vector space. The area around the node is highly volatile; small changes in location of a point can entirely change the eventual path of that vector. Further from the origin, this volatility dissipates, and eventual paths become significantly less sensitive to minor changes in location.

These tools collectively allow us to qualitatively analyse the resulting systems.

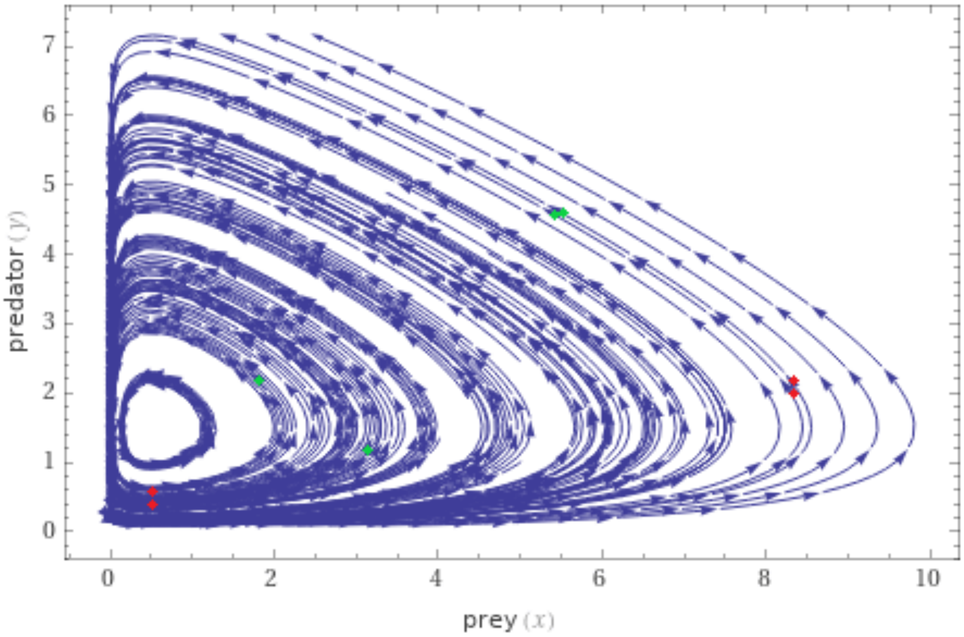

Figure 6.4: The Lokta-Volterra Predator-Prey model (Wolfram Alpha, 2021).

The Lokta-Volterra Predator-Prey model, describes the relationship between populations of predators and populations of their prey as a vector space. Arrows represent the direction of movement of a point in the vector space and spacing of arrows show the magnitude of that movement (greater spacing is greater magnitude and *vice versa*). As populations of prey increase, with non-zero predator values, the population of predators will increase and slowly deplete the quantity of prey. As the quantity of prey tends toward zero, the population of predators begins to collapse, until quantities of prey recover. The graph has a cycle point (0.5, 1.5) and a saddle point at the origin. It also displays slightly more amplitude further from the cycle point. This means that for high populations of prey and predator we would expect slightly larger shifts in population over a fixed time period than for low populations. More significantly, there is a great deal more volatility closer to the origin.

If we are applying this model to understand an ecosystem, we need to be aware of the effect of measurement error if we are close to the origin because this volatility means that small errors can drastically change what the ecosystem will be like in four arrows′ time. The two red points near the origin represent two measurements of the population where the prey population has been measured accurately and identically but there are two divergent measurements for the population of predators. This small variation in four arrows′ time becomes a substantial discrepancy between the resulting green points, not only in what the populations will be, but also their trajectories going forward. By contrast the red points far from the origin have the exact same error in measurement, but the resulting divergence in outcome is even smaller than the initial error in measurement. Knowing where we lie on the topology of the system tells us how things are likely to change, how quickly things are likely to change and how important knowing our exact location is to predicting where we will be in the near future. Graphical representations of such models in vector space allows for nuanced interpretation of how an input will be resolved if we are correct about its location, as well as properties of the space that may inform us about the volatility of our predictions and the effects to which they will be sensitive.

Given the power of such analyses to offer explanations of how structures generate the outcomes they do depending on the space they inhabit and their initial conditions, could dynamical systems theory ground CAPS in such a way that the conclusions drawn from CAPS are warranted?

In order to represent CAPS as a dynamical system, the operation of its parts across vector space would need to be calculated and the differential equations, which describe movement across that space, outlined [58]. With the differential equations outlined we may then map the resulting space, perhaps generating something akin to Waddington′s (1957) depiction of the epigenetic landscape.

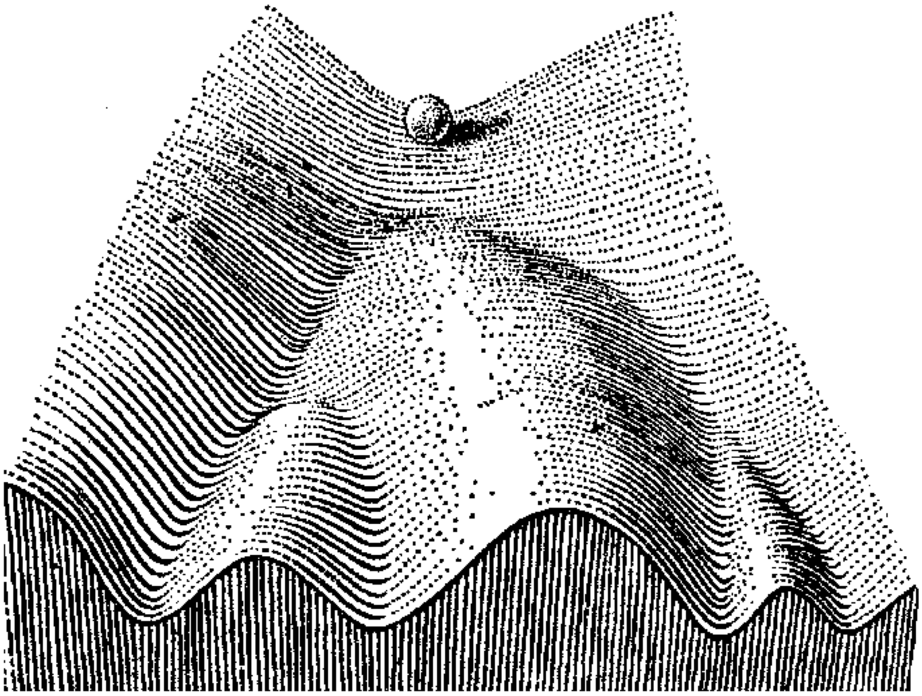

Figure 6.5: The Epigenetic Landscape (Waddington, 1957, p. 29)

[58] I here set aside concerns that doing so is impractical or impossible as similar concerns weigh on any attempt to adequately map a personality. That is, there are no clear reasons why the problem is greater for the representation of CAPS as differential equations than for any other mapping of a personality, though it is perhaps less obvious how this would be done incrementally in the manner of other operations.

In the above figure (Waddington, 1957) the landscape is shaped by the genes that pull it into shape from beneath creating paths which may or may not be followed in the development of a cell (the ball at the top of the diagram). As it moves across this space the cell reaches bifurcation points and takes one of the paths available to it. With enough perturbation from an external factor, it could also 'jump′ from one path to another. Something akin to this structure might be used to describe a CAPS model of attitudes.

The CAPS receives an input (a starting position of a set of beads) which then runs across a vector space, largely following the major valleys in the absence of external perturbations. This offers an output where some beads end at the close of major pathways, some fewer end at the close of minor pathways, and occasionally one ends outside a pathway altogether. The landscape has shaped the output, but its precise outcome remains sensitive to the precise details of the starting conditions, as well as relevant perturbations during the process. This offers a description of how the landscape of the CAPS structures responses to stimuli. With small changes in the landscape, we would generally expect small changes in the outcomes, but the approach also allows for minor changes to entirely alter the outcomes available from a CAPS. This preserves the ability of CAPS to explain both general consistency in behaviour and gradual change as well as those rarer occasions where more radical changes occur.

This grounding for CAPS offers clear resources explanations of the system as a whole. Why did P act in such and such a way when exposed to S? Because the landscape of their cognitive-affective system is structured towards particular sorts of responses to those sorts of inputs. What would it take for P to act in some other way? Substantive changes in their landscape or external perturbations of sufficient magnitude to alter outcomes. Perhaps, with enough information about that landscape, we may even be able to specify where and what changes could be targeted to bring this change about.

If we wish to explain why P acted in such and such way in response to S in terms of the parts of the system, we need to reintroduce the tools outlined above. What are the relevant points in the system which act as one of the various kinds of nodes, and which areas between those nodes have high amplitude or volatility and which low? In figure 6. 5 from Waddington, saddle points occur at the bifurcation of valleys, stable points occur at the base of each valley at the local minima, with unstable points at the ridge between each valley at the local maxima. Some sides of the valley have high amplitude, accelerating the bead toward the local minima rapidly, others have much lower amplitude. Similarly, some areas are volatile as the bead heads for bifurcations or toward large stretches of level landscape (not offered by Waddington's example) where small changes in the momentum or location of the bead have drastic consequences for its path. Other areas offer very little volatility, such as the base of steep valleys, where even quite large changes in location or momentum will have little effect on the path of the bead. These descriptions of the parts of the landscape allow us to answer questions about local effects on the bead rather than referring to the space as a whole.

For a general description of CAPS, this may be adequate grounding – we have an explanation of the key features of stability and gradual change as well as a way, in principle, of determining where radical changes may occur. It also allows a more fine-grained analysis in terms of the relevant nodes and surfaces which structure the responses to relevant inputs. What it does not motivate is the explanation of these properties in terms of cognitive-affective units, or indeed attitudes[59].

[59] For the remainder of this section, I take Webber's (2015, p. 1085) point that cognitive-affective units, and talk of and using them, is the motivation for using CAPS as a model of attitudes at all. The cognitive-affective units are what provide the object-affect relation that attitude psychology studies. Insofar as cognitive-affective units are the relevant conceptual resources CAPS offers that can make it a model of attitudes, then failure to ground these under dynamical systems theory also leaves us without grounds for CAPS as a model of attitudes.

Regarding cognitive-affective units, Mischel and Shoda describe them as:

> ″Situational features are encoded by a given mediating unit, which activates specific subsets of other mediating units, generating distinctive cognition, affect, and behavior in response to different situations. Mediating units become activated in relation to some situation features, deactivated (inhibited) in relation to others, and are unaffected by the rest. The activated mediating units affect other mediating units through a stable network of relations that characterize an individual.″ (Mischel and Shoda, 1995, p. 254)

These units are then related to one another within the system such that:

> ″relationships among the cognitive and affective units guides and constrains further activation of other units throughout the network, ultimately activating plans, strategies, and potential behaviors″ (Mischel and Shoda, 1995, p. 255)

This allows CAPS to incorporate folk psychological understandings of those things which instantiate the units within a novel framework which explains how commonly understood notions can give rise to the full panoply of behaviour with the relevant features of stability and gradual change as well as explaining those cases where our folk-psychological intuitions go awry.

These units are instantiated in five ways:

″1. Encodings: Categories (constructs) for the self, people, events, and situations (external and internal).

2. Expectancies and Beliefs: About the social world, about outcomes for behaviour in particular situations, about self-efficacy.

3. Affects: Feelings, emotions, and affective responses (including physiological reactions).

4. Goals and Values: Desirable outcomes and affective states; aversive outcomes and affective states; goals, values, and life projects.

5. Competencies and Self-regulatory Plans: Potential behaviors and scripts that one can do, and plans and strategies for organizing action and for affecting outcomes and one's own behavior and internal states.″ (Mischel and Shoda, 1995, p. 253, Table 1)

The explanations the CAPS theorist seeks to offer of behavioural patterns makes reference to these units explicitly:

″Person 1 tends to become irritated when she thinks she is being ignored, whereas Person 2 is happier when he is left alone, and even becomes irritated when people tell him personal stories. Suppose also that in Situation A people rarely initiate personal interactions whereas in Situation B such interactions are relatively frequent. Then Person 1 will become irritated in Situation A but not in Situation B; Person 2 will show the opposite if...then... pattern, irritated if B, but not if A. These affects further activate other cognitions and feelings in each situation, following the pathways of activation distinctive for each person. These individual differences reflect the particular acquired meanings of the situational features in terms of the cognitions and affects associated with them, so that even if both people are similar in their overall levels of "irritability" they will display distinctive, predictable patterns of behavioral variability in their if..then... signatures.″ (1995, p. 255)

In the above example, the difference between person 1 and person 2 is a difference in a single affective encoding of a type of situation, either represented as a single cognitive-affective unit or as a small cluster of interconnected units. This unit or cluster then causes the behaviour from the system such that we may meaningfully say that Person 2 felt irritated because they were in Situation B and had the irritability encoding for social situations.

Such an explanation, while natural for CAPS to give, does not seem to be groundable in dynamical systems theory. A CAPS-internal [60] explanation for why P will tend to act in such and such way if they find themselves in circumstance S, if CAPS is grounded in dynamical systems theory, will be

[60] It is as available to CAPS, as to any other model of behaviour or personality, to appeal to external factors altering the outcome, but if CAPS is to be a useful framework its explanations must be capable of offering this sort of explanation if those external factors are held constant.

exhausted by reference to the structure of the vector space or explanations which are reducible to such references.

At minimum, the reduction of cognitive-affective units to properties of the nodes or areas of vector space is non-trivial. This is problematic because the precise details of how one is to do so changes the explanatory power of CAPS in most of the interesting cases it wishes to explain. If one should treat cognitive-affective units as one treats genes in the Waddington epigenetic landscape then these units become the substrata which provide the tension which shapes and distorts the vector space of the CAPS. Or perhaps one should treat cognitive-affective units as the substance of the vector space where moving from (x1, y1, t1) to (x2, y2, t2) involves the movement of influence-over-behaviour from one unit, or cluster of units, to another. It may be that a CAPS theorist may instead treat the language of cognitive-affective units as imprecise and ʻlow resolutionʼ while the dynamical systems representation is much more precise and ʻhigh resolutionʼ. These treatments of the cognitive-affective units warrant different sorts of explanation of why P acted in such and such way in circumstance S, in terms of those units as well as entailing different sorts of predictions about what behaviour we should expect in the short to medium term from P, as well as the relative accuracy and adequacy of these explanations.

More problematically, each and every of these units are supposed to possess content that is realised through their interconnection with other units but which is not reducible to their situation within, or influence upon, vector space. If I plan to A in situation S because I desire X, the content of this plan is not captured by reducing it to the effect it has on the shape of a vector space. This loss of the content of the units is relevant precisely because the explanations CAPS offers moves from the discussion of the connections between these units to the details of behaviour in response to the content of the relevant units.

This is not to say that models of behaviour or cognition grounded in dynamical systems theory all face this worry, though many of them explicitly disavow the internal content of minds in order to pursue the project (Thelen and Smith, 2001, p. 11). The problem arises because CAPS already offers particular kinds of explanations which are seeking adequate grounding. If the proposed grounding in dynamical systems theory does not ground those kinds of explanations, then CAPS must alter substantially to offer explanations for which it has warrant, or find new grounds [61].

Section 6.4: Connectionism: Feed-forward Networks[62]

As highlighted above, the key alternative approach to dynamical systems theory as a ground for the kinds of claims that CAPS makes is connectionism. This section briefly introduces feed-forward neural networks, and explains why these are insufficient to instantiate CAPS. They are introduced and discussed to facilitate the explanation in section 5 of the more complex continuous flow networks which make a far more plausible candidate for CAPS.

A feed-forward net consists in *nodes* and *connections*. The nodes have two properties: a *location* and a *weight*.

The location of a node describes how it relates to other nodes. If it receives input from outside the system but otherwise has no connections directed towards it then it is a node of the input layer. Similarly, nodes which have connections directed towards them but no connections directed away from them constitute the output layer. Those nodes with connections directed towards and connections directed away are part of the hidden layer or layers. More formally, no member node (N) of a given layer (L) has an output (N(O)) which depends upon, or is depended upon by, the output of any member of that layer, including itself. Furthermore, no node (N) has any inputs (N(I)) which depend upon the output of that node (N(O)). This definition of location is sufficient to ensure that a network is feed-forward

[61] As highlighted in an earlier footnote, the use of CAPS as a model of attitudes depends on the grounding of the cognitive-affective units, and as such the conclusion regarding cognitive -affective units applies *mutatis mutandis* for attitudes.

[62] This section is based on Gurney 1997: chapters 1-3.

since it guarantees that all layers are synchronous and that there are no internal loops: activity flows across the network in only one direction. These layers are illustrated in figure 6.6.

The second property of nodes is their weight. There are different ways of representing weights and their interactions. Mischel and Shoda, in their appendix, represent these relations of weight as summing across the system (1995: 267-8). When represented as addition, these weights have a value between −1 and 1. Nodes whose weights are negative values function as inhibitors of incoming signals, those whose weights are positive values are amplifiers of incoming signals, and those whose weight is zero merely pass on signal as it is received, though their presence may still shape the overall output of the network due to the structure of the connections. More complex equations may be used to derive the nodes′ output values from their inputs, which may change the range of these values, but the basic principle holds.

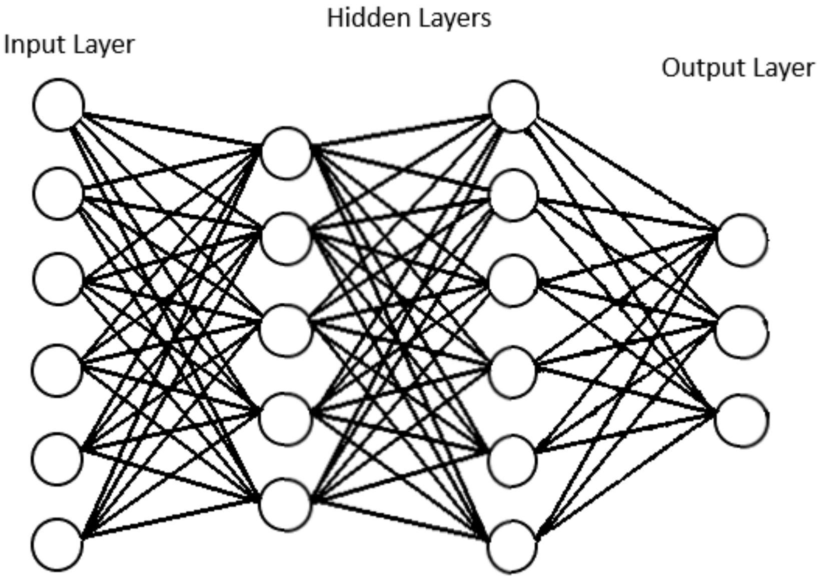

Figure 6.6: A feed-forward net with two hidden layers.

Connections have two properties: a *direction* and being the *vector*[63] of the nodal weights. The direction of a connection describes how it relates two nodes. Being a vector describes how a connection links the two nodes it connects. The weight of the effector node travels along the connection and, in CAPS as Mischel and Shoda describe it, is added to the weight of the receptor node. This combined weight is what travels along further connections when the receptor node itself then acts as an effector node. The direction of a connection describes this relationship between the effector and receptor: the direction of flow of weights.

These two parts, and their four properties, allow feed forward networks to take a variety of inputs and render a final numerical value for each node in the output layer. With given values of the input nodes and the operation along each connection across the hidden layers, each output node will have a resulting value. This allows the net to operate as a sorting program. Each output node is taken to represent some category and whichever node has the higher value is taken to be the category which the input is sorted into. With random values assigned to the hidden layers of the algorithm the outputs for any given input will, in essence, be random. This is an example of what is called the ′propagation′ of the network, specifically, its ′forward propagation′. In the first instance, with random starting values, the forward propagation results in random output values.

There are two broad manners of improvement available to feed-forward nets: supervised and unsupervised learning (Gurney 1997, pp. 97-132, Hinton, 2010). Supervised learning requires the input to come with some form of meta-informational tag identifying the input′s category. These tags allow the sorting by the net to be judged correct or incorrect. The term ′supervised′ refers to this supervision of the network by the tagged dataset. A supervised learning process provides an input with a meta-informational tag to the network and then compares the output sorting with the value of the

[63] Here used in the epidemiological sense of being the propagator of some transmitted property rather than as possessing direction and magnitude as used in dynamical systems theory. Apologies for any terminological confusion that results; the terms are not my own.

tag. When the network ʻcorrectlyʼ sorts the input, no changes to the network are made. When the network ʻincorrectlyʼ sorts the input, the network needs to be changed in order to ensure it makes the correct sorting. This is most commonly done by a process called ʻback-propagationʼ.

Back-propagation is a simple statistical approach to learning which treats learning as an optimisation problem, specifically, the minimisation of an error quotient. Backpropagation has six steps:

1. Identification of an error in output, both its existence and its magnitude.
2. Calculation of an error derivative, expressed as a gradient of error (EΔ).[64]
3. This error derivative is then run back across the net. [65]
4. Calculate the new weights: For each node the new weight (W_{n+1}) will be a function of the previous weight (W_n), the connections between the node and the output (C[N⇔O]), and the error derivative (EΔ): $W_{n+1} = F(W_n, C[N \Leftrightarrow O], E\Delta)$
5. Update the weights to the newly calculated weights ($W_n \rightarrow W_{n+1}$).
6. Re-propagate: Run the original input across the network to check the ʻcorrectʼ output is now generated.

Updating in this manner means that the network will eventually correctly sort the given input and is repeated until it correctly sorts the entire training set. This process ensures that the minimal amount of correction is applied to the network weights to achieve a correct sorting. When repeated over a large sample of inputs, backpropagation will derive the minimum overall error quotient for the network structure in sorting input items of a given type into the given output categories. Assuming the category is

[64] The exact formula for calculating the error derivative will vary based upon different approaches to backpropagation and different structures of the network, but the principle remains the same regardless.

[65] Using the same weights as when the input was fed forward (propagated), the error derivative is run backward across the network with the weights operating on the error derivative in the same manner as on the initial inputs.

obtainable the network will become reliable in sorting inputs in a systematic manner.[66]

As an example, imagine each node on an input layer refers to a pixel of a black-and-white portrait. Each input has a value between 0 (pitch black) and 1 (white). The nodes in the hidden layers then each have a random value starting weight ($-1 < W_0 < 1$) assigned. The network has two outputs, labelled male and female. As the first portrait is fed into the network a random output is generated. If this output correctly identifies the portrait as a portrait of a male or female person, then the next input is fed in. If the output is incorrect, then the magnitude of error is found (i.e. the minimum amount that the output figures would have to change by in order to render the correct answer). The derivative of this error is then calculated and backpropagated, which establishes the new weights. This entire process is then repeated across hundreds or thousands of inputs.

Once the sample set is completed the network will be able, with relatively high accuracy, to identify whether a black-and-white portrait is of a man or a woman. This demonstrates the power of a supervised feed-forward net in tackling specific, well-defined problems. For less well defined or unlabelled datasets, however, supervised training is not possible. In these cases, unsupervised learning is necessary.

Unsupervised learning describes any learning algorithm which involves the self-organisation of the inputs into clusters, or the modelling of the

[66] A category is categorically obtainable if there are no local maxima/minima, and the categorisation reliably refers to some property of the input. If there are local maxima/minima the categorisation is conditionally obtainable. Whether such a categorisation in fact obtains will be dependent on the starting point of the algorithm (e.g. an algorithm sorting pictures of faces into happy and sad may become ′stuck′ dividing the faces based on brow line, which tracks, but is a local maximum of, the properties of a sad/happy face). If there is no property of the input that the categorisation refers to then the category (insofar as there can be said to be a category) will not obtain, though some correlation may still be generated (e.g. asking an algorithm to sort novels into blue and red to track how a group of English literature professors sorted them, may produce some correlation, but is unlikely to actually track any category of ′blue novels′ or ′red novels′).

probability density of inputs. In essence, this involves the categorisation of the dataset by statistical programmes rather than by human 'tagging'. The exact form this takes varies widely depending on application. These statistical analyses all identify key clusters in the dataset and treat these clusters as categories. With this categorisation the network is able to assign output nodes to the categories and run backpropagation as with the supervised learning. While there are noticeable differences between supervised and unsupervised learning, their basic structure remains broadly the same.

Finally, feed-forward networks may be recursive in a carefully constrained sense. Aguiar, Dias and Field (2019) discuss and introduce several proofs of dynamic synchrony[67] for feedback loops in feed-forward neural networks. Specifically, for loops which connect the entire output layer to the entire input layer such that every node in the input layer receives at least one connection from the output layer and every node in the output layer is connected to at least one node in the input layer (2019: 22). Their proofs show that synchrony may be maintained for these neural networks with what I shall call *external loops*. This proof of the preservation of the synchrony of layers proves that there is no contradiction in the inclusion of such a loop in a feed-forward network. This allows the output of a given node to, without contradiction, contribute to its own input.

Loops have what is called a 'transverse', which refers to the minimal set of nodes and connections that are necessary to instantiate the loop. In the loop ABCBCD, the transverse of the loop would include the nodes BC and the connections *BC* and *CB.* The proof offered by Aguiar, Dias and Field (2019) does not hold for loops which connect nodes in one layer to nodes in an earlier layer when the layers being connected are not the output layer and the input layer. In the case of such an *internal loop*, the transverse of the loop would not necessarily preserve synchrony within itself, nor would it necessarily preserve synchrony with the rest of the network (2019: 34). Such

[67] The property of a network or transverse such that each layer may be entirely calculated simultaneously. This is necessary in order that the next layer may be calculated. Without synchronicity, propagation and backpropagation fail to operate.

internal loops generate networks which cannot be synchronously calculated and, as a result, are not stable.

The instability of these equations means that for any given propagation, there does not exist a single stable solution. The outputs will possess either multiple values or no value at all. Because of how the output is used in feed-forward networks, this likely means that the network will fail to offer a sorting solution. Despite this, an unstable sorting solution may still be possible and useful from a single forward propagation. If every member of the set of values for the one option is greater than every member of the set of values for another, the network has successfully sorted despite offering multiple solutions. However, what must then be established is the degree of error and how to backpropagate across such a network. Doing so will offer multiple solutions to the new weights, and no way to choose any one solution over another. As a result, for the purposes of a network that must be both forward propagated and backpropagated, such unstable internal loops are impossible.

These proofs allow for the inclusion of external loops in feed-forward networks demonstrated by the application of so called Recursive Neural Networks and other external-loop, or memory-based networks.[68] This is important, structurally, because they highlight how a network may operate in constant interaction with its environment while its own output shapes that environmental input. This is important as we move to apply such connectionist networks as the grounds of CAPS.

With feed-forward networks described we may now turn to their similarities and commonalities with the CAPS model. Feed-forward networks have two kinds of entities, nodes and connections. The nodes have two properties: location which sorts the nodes into input, hidden, and output; and weight, which determines the node's influence on the output of the network as a whole. Similarly, the connections have two properties:

[68] These kinds of recursion characterise Elman Networks (Elman, 1990, 1991), Hopfield Networks (Hopfield, 1982), and gated networks such as Long Short-Term Memory (Hochreiter and Schmidhuber, 1997).

direction; and being the means of operation of the nodal weights. Feed-forward networks have a further notable property of the network as a whole: the possibility of external feedback loops. These external feedback loops are illustrated above in figure 6.1 and in figure 6.7 later in this chapter, both from Mischel and Shoda (1995).

CAPS draws analogies to each of these features and properties. Input nodes are analogous to perceptual inputs, identifying features and codifying the world around the individual. Similarly, output nodes are analogous to behavioural responses, available in different modalities to varying degrees. The hidden nodes and the connections between them are taken to be analogous to the cognitive processes which give rise to behaviour. The second property of nodes is weight, this represents something like the influence a given node has on the overall behavioural outcome or subsequent dynamics of the CAPS system. In input/perception nodes this represents salience detection: the importance of faces, the identification of danger, social cues, etc. In hidden/cognition nodes this represents the relative influence some beliefs or desires or sets of beliefs or desires have on behavioural outcomes. These weights are reflected in the behavioural outcomes being shaped by particularly well-connected beliefs or desires.

The connections between nodes capture a broad category of interactions between the mental states and processes represented, from perceptual uptake to inference to association to motor control. These connections represent the process by which external inputs activate mental states, which in turn activate further mental states and, eventually, behaviour. This represents the first property of connections: being the means of operation of the nodal weights. The second property of connections is direction. In CAPS, this is represented by the potential for asymmetry in connections: thinking of coffee may be followed by thinking of alertness but this does not entail that thinking of alertness will be followed by thinking of coffee, the connection strengths may change independently of one another.

Furthermore, CAPS relies on the ability to construe the network as having external loops to explain the phenomenon of our behaviour informing our perceptual experience of the external world as well as the way in which we may codify, or weight, those experiences. This dynamic, expressed as external loops, is incorporated by Mischel and Shoda in order to frame ʻthe person not as reacting passively to situations, nor as generating behaviour impervious to their subtle features, but as active and goal-directed, constructing plans and self-generated changes, and in part creating the situations themselvesʼ (1995: 252).

By drawing these parallels, CAPS is able to represent behavioural stability as emerging from inputs being processed by the same network with the same calculation weights which sort similar inputs into similar categories and responds accordingly. This overall stability is the source of the stability of feed-forward nets: their ability to consistently sort inputs. Simultaneously, because of how the network sorts inputs, feed forward networks can be thrown by seemingly innocuous features of the input. The manner in which a feed-forward net assigns weight to features of the input, or to connections between nodes, can lead to it tracking features which correlate with the relevant category in a wide class of cases but not in another smaller set. In this second set the network will produce substantially diverging outputs despite there not being a category-relevant change in the situation. The network was tracking something which correlates with the category, but which is not in fact category-relevant.

The second phenomenon that CAPS seeks to explain is how personality may be relatively stable while allowing for personality change over time. By grounding our explanations of CAPS in feed-forward nets, CAPS may explain change as back-propagation. When an individual encounters a negative reaction to their behaviour or fails to meet some success condition, CAPS may represent their learning as a backpropagation of the error (or a derivative of the error) in response. We may represent questions such as ʻwhy did I do that?ʼ, ʻwhat went wrong?ʼ, ʻhow can I do better in future?ʼ, or ʻwas I wrong about X?′ in terms of backpropagation. The identification of error, the use of the same associations and connections with the same strength to determine where the error originates from, to what degree, and to update to a new value, which is tested in future situations, can all be modelled following the same steps as backpropagation outlined in section 3.

Despite these similarities between feed-forward networks and CAPS, however, there remains an important difference between the two: CAPS requires a groundwork that allows for internal recursive loops.

The argument for this claim begins by considering the phenomenology of thought, then offers several explanations which CAPS may offer, concluding that the one which is least problematic is that recursive loops must occur within the network.

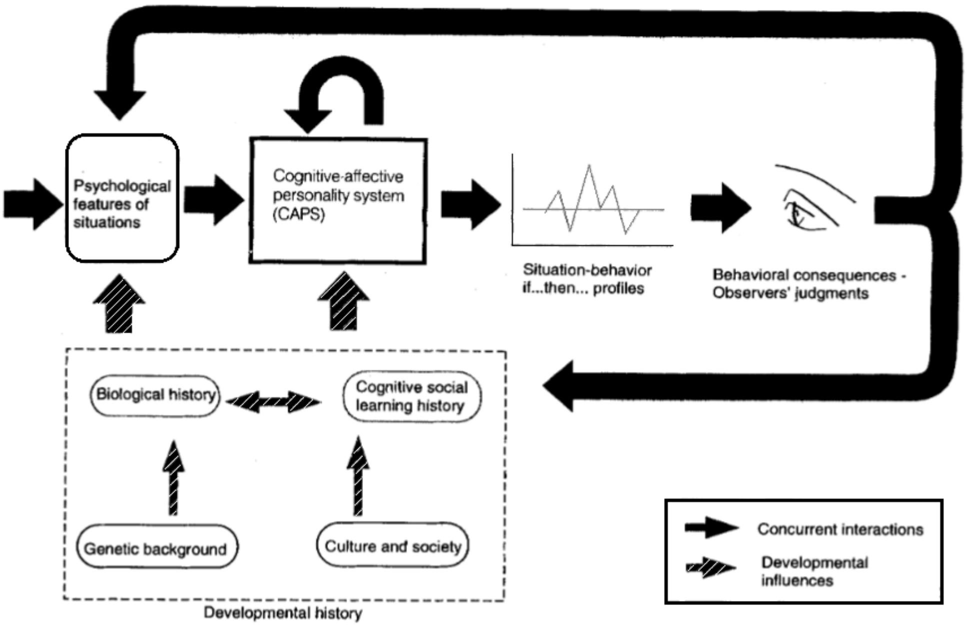
Psychological features of situations
Cognitive-affective personality system (CAPS)
Situation-behavior if...then... profiles
Behavioral consequences - Observers' judgments
Biological history
Cognitive social learning history
Genetic background
Culture and society
Developmental history
Concurrent interactions
Developmental influences

Figure 6.7: CAPS in relation to concurrent interactions and developmental instances (Mischel and Shoda 1995: 262).

Mischel and Shoda (1995) are ambivalent about the need for internal recursive loops in CAPS. In their diagram (figure 6.7, above), there are concurrent interaction loops leading from the CAPS back into the CAPS. This implies some commitment to the existence of recursive loops. Similarly, in figure 6.1 there are loops of connections between nodes represented which do not go via the behaviour output. Furthermore, they state that CAPS ′is continuously activated by its own internal feedback system through chronic activation of cognitions and affects and their interactions within the system′ giving as examples ′long-term planning and sustained goal pursuit′ and ′such activities as fantasy, ruminations and daydreaming′ (1995: 262). However, the majority of their discussion focuses on those cognitive structures that run via the world, especially behavioural feedback from observers and criteria for goal success or failure, and the extended example they give in their appendix does not incorporate any internal loops. [69] I will argue that CAPS *must* include internal recursive loops and therefore may not be represented as a feed-forward network.

Suppose I asked you for your opinion on the current state of democracy in western Europe. Suppose I also asked you to consider your answer for five minutes before answering. What happens in those five minutes? Some of the thoughts that occur to you might involve: the increased rise of populism in Europe, presumably with some affect depending on your view on such a rise; your knowledge of the debt imbalance within the eurozone and the lack of an exchange mechanism to relieve such an imbalance; general real wage stagnation even in the wealthiest European countries; loss of a widespread European identity being co-opted into increased reliance on national

[69] In their later work, both Mischel and Shoda explicitly commit to the necessity of internal feedback loops (Shoda, LeeTiernan and Mischel, 2002: 318; Shoda and Smith, 2004: 157; Shoda and Mischel, 2006: 443-448; Shoda, et al. 2013: 555). However, in all instances of modelling behaviour or simulating situations in their later work, feed-forward non-recurrent networks are used (Shoda, LeeTiernan and Mischel, 2002: 319; Shoda et al. 2013: 556). Moreover, each instance of the latter appears on the page directly following an instance of the former: each feed-forward simulation or linear modelling technique is prefaced by its disavowal. This seems to demonstrate the depth and persistence of the CAPS theorists′ ambivalence about internal feedback loops and how they may be modelled, though it is plausible that the complexity of such recurrent modelling is what motivates this apparent ambivalence. If it is the modelling complexity which motivates the apparent ambivalence, this section seeks to argue that such complexity is intractable.

identities; which resolves back into the thoughts about the rise of populism in Europe.

I take this to be an uncontroversial, highly plausible description of the phenomenology of rumination. I also take this to be an apt description of a diverse category of thought necessary for a wide range of cognitive and behavioural tasks. Thoughts occur which lead to further, usually related, thoughts. Often, a single thought-path will incorporate at least one thought more than once as an issue is returned to, perhaps in conjunction with some new context, perhaps not.

How are we to explain such a phenomenon in the context of CAPS? It is easy to see how a single idea might recur in a thought process by means of an external loop. For example, one might make notes on a whiteboard to facilitate one′s reasoning. Here the thoughts cause behaviour which creates a chronic environmental cue causing the thought to recur. This example cannot explain cases where a single idea recurs in a thought process that does not use such an external prop, as in our rumination about democracy in western Europe. However, not all external loops require props in the world. Rather than writing on a whiteboard, suppose you instead spoke out loud. Further, suppose you merely mentally rehearsed such speech. This imaginative action could be an output of the CAPS system that then acts as an input, creating an external loop entirely ′within the skin′. Mischel and Shoda do explicitly state that such activities constitute an input to the CAPS system (1995: 251).

But this kind of loop still does not cover all the cases of recurrent ideas. There remains a class of cases where no such deliberate mental activity is undertaken and yet a single thought recurs. Indeed, when one is kept awake at night by the dogged recurrence of a small set of thoughts, this seems precisely to be what occurs. How might such recurrences be explained, if not by loops internal to the CAPS system itself?

Might the apparent recurrence in fact be a case of being multiple tokens of the same node type in the system? This might seem to allow for the phenomenology without internal feedback loops. However, it would require a CAPS system of infinite size. This is a basic point of graph theory. The suggestion is that the recursive phenomenology of thought processes is represented in the CAPS model by a linear progression across a system that includes multiple tokens of the same type. This is equivalent to translating a directed cyclic graph into a directed acyclic graph. This cannot be done, except by making the resulting directed acyclic graph infinitely long (Wilson 1996: 26-42). The CAPS proposal is to capture the complexities of behavioural stability, change, and situational variation in a simply expressed model. Such a model cannot be infinite in size.

The remaining explanation of such a phenomenological description is that the apparent recurrence of the same mental state is modelled as being precisely that. The flow of activity through the CAPS system includes internal loops. This is not to assume that the flow of activity through the CAPS model must match the phenomenological flow of thought. The point is rather that the CAPS model could not explain that phenomenology in any other way. This necessity would explain why Mischel and Shoda sometimes indicate that CAPS includes internal loops.

Feed-forward nets, however, flow in a single direction from input to hidden layers to output, or, in the case of backpropagation, in reverse. They do not, and cannot, incorporate internal feedback loops. This is because the calculation of a feedforward net depends upon synchronicity across network layers. Nodes in a feed-forward network cannot calculate their output until they have received all of their inputs. If a node (N_1) in layer (A) depends for calculating its output on a node (N_2) in layer (B) which itself depends upon the output of N_1, N_1 will fail to calculate its output in the first instance because it has not received its input from N_2, N_2 will fail because it has not received its input from N_1, and this section of the network will never activate. More simply, if there are internal feedback loops in a network then there is at least one node whose output would partially depend upon its own output. As a result of the necessity of internal recursive loops, it is not possible to fully represent CAPS as a feed-forward network.

We cannot accept a version of CAPS that leaves internal loops unmodelled. Such a model would not be able to track a significant proportion of the relevant phenomenological and behavioural evidence and would systematically give false predictions where the magnitude and kind of error would remain unknown. It could scarcely be said to be a model of the target phenomena at all. As such feed-forward networks cannot ground the explanations CAPS seeks to offer of its target phenomena. However, there is another kind of connectionist network that does allow for internal loops: continuous flow networks.

Section 6.5: Connectionism: Continuous flow networks[70]

The previous section introduced the two entities of feed-forward networks and their properties. A continuous flow network also consists of nodes and connections, its nodes have the property of nodal weights, and its connections have the properties of connection weights, being weight vectors, and having a direction. One significant change to these properties concerns node location.

Nodes in a continuous flow network do not belong to layers in the same way that nodes in feed-forward networks do. As with feed-forward networks, some nodes will receive input from outside the network, and so may be

[70] The literature sometimes refers to networks that can recur collectively as 'recurrent networks'. I use the term 'continuous flow' to differentiate the class of networks that can receive continuous input. This excludes those networks that may recur but may only receive sequential inputs and are constrained to reach equilibrium (e.g. Hopfield, 1982), higher order recursion (Elman, 1990, 1991), as well as gated networks such as Long Short-Term Memory (Hochreiter and Schmidhuber, 1997). These networks are constrained to external recursive loops discussed above, to avoid some, though not all, of the problems with back-propagation raised towards the end of this section. The category of recurrent or continuous flow networks includes Continuous Time Recurrent Neural Networks (Beer, 1995), and Liquid State Machines (Maass and Markram, 2004; Hazan and Manvitz, 2012), among others. This section summarises the shared features and emergent properties of this class of networks. For a discussion of the substantial merits of the application of these networks to specific problems in psychology, see Jordan (1997).

regarded as input nodes. However, unlike in a feed-forward network such nodes may also receive inputs from within the network. For example, in figure 6.8 the top two nodes may both receive input from outside the system. The top-left node receives only an external input, while the top-right node receives an external input and input from two other nodes within the system. As a result, while they are both ʻinput nodesʹ, there is no ʻinput layerʹ since one of these nodes depends upon the other.

Output from a continuous flow network may operate in two ways. The first is that certain nodes have connections to an output function making them output nodes. These output nodes will govern the output function. I will refer to this kind of network as a node-output network. Alternatively, the whole network activation pattern may be holistically monitored, and the pattern of activity within the network continuously generate an output. I refer to this kind of network as a holistic-output network. Either kind of output seem *prima facie* plausible for a continuous flow grounding of CAPS, and everything which follows is applicable to either kind. The decision about which kind provides the best fit for CAPS does not bear weight in the overall argument.

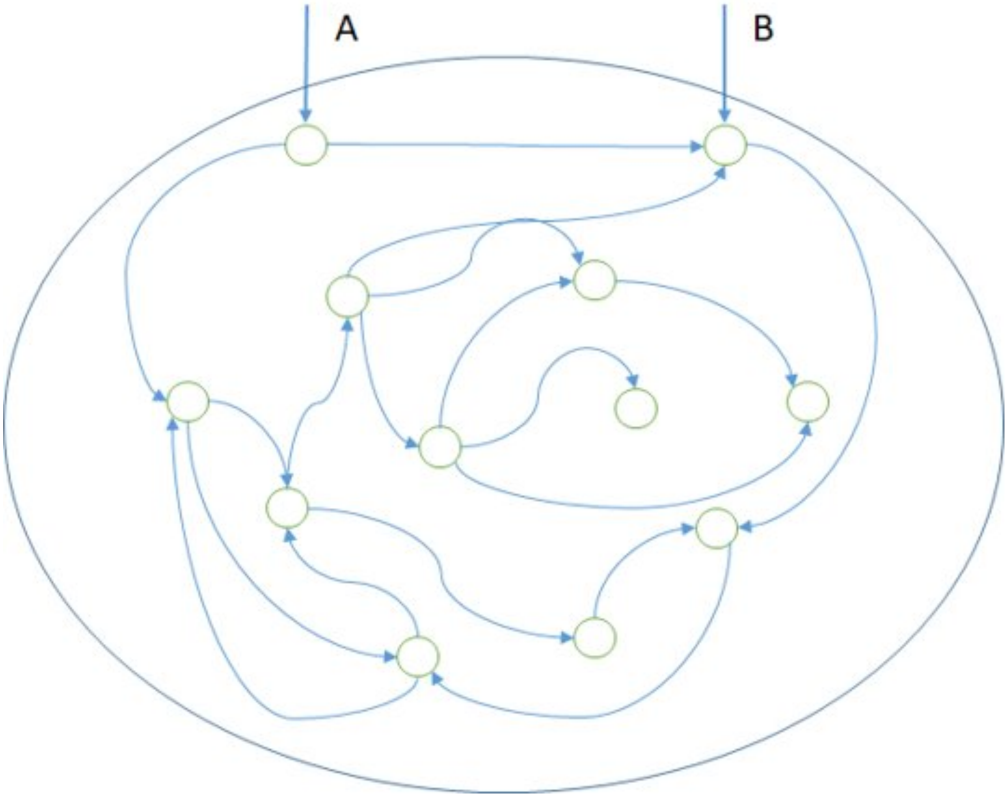

Figure 6.8: A continuous flow network with two inputs, A and B.

The entities that constitute continuous flow networks have a further property, which allows them to operate continuously rather than sequentially. Nodes in a continuous flow network have storage. This records, for a fixed period, the inputs the node has received. This storage defines the set of inputs that the node receives, which are factors for the node′s output. By relying on storage with a time-frame, nodes no longer require their complete set of inputs before generating an output. The output is instead generated continuously, rather than as a set value. This change in input requirement allows them to operate without layer synchrony and thereby to have recursive loops within the network.

These differences in the properties of the constituents change how the network operates as a whole. The network does not necessarily run to a definite conclusion; it is possible for a finite input to have an infinite output. This lack of a definite conclusion also changes how the output may be represented. In a feed-forward network the output nodes each reach a static value. In a continuous flow network, each output is a continuous waveform. As a result, the output of the system as a whole is more like the output of a heart monitor than the numerical value of a blood pressure reading.

Continuous flow networks make a good analogy for CAPS for four reasons. First, as with feed-forward networks, continuous flow networks have input and output systems that mirror the perceptual input and behavioural output structure that CAPS outlines, along with the intermediary, interconnected nodes that represent the subject′s mental states. Second, these nodes have weights which represent the influence on behaviour of any given mental state. Third, these nodes are connected to one another by connections which possess a connection weight and a direction, representing both the closeness of the cognitive association between two mental states and the potential for asymmetry in such relations. Finally, the storage capacity of each node represents the duration over which inputs are processed to calculate the continuous output, representing the way that inputs are only finitely efficacious on their recipient nodes – once activated they are not influential forever.

To illustrate how continuous flow may represent some relevant mental phenomena consider the following cases: Andrew and Bethan. Andrew is sitting a mathematics exam. He reads the question at the start of the paper (perceptual input), which leads to his recall from memory of many beliefs he has about the topic in question, along with affects about being in an exam setting, and desires to answer the question quickly and move on (intermediary mental states). As it happens, the desire to answer the question quickly is more influential than many of the relevant beliefs and leads to Andrew writing out some formulae he remembers that are relevant to the topic, out of panic (behavioural output). Seeing these formulae (perceptual input), Andrew′s memories of the topic (intermediary mental states) are better connected and more influential on behaviour and he is able to collect the relevant thoughts and formulate an answer to the question (behavioural output). This illustrates an external loop, where behaviour causes changes in perceptual input, which allows for different associations to become influential, which causes further behaviour.

Bethan is also sitting an exam, hers is in politics. She turns over her exam paper to read the first question on the 2015 Greek election and the subsequent actions of the Troika (perceptual input). This connects to a wide variety of beliefs and affects about the morality of the actions of the Troika (an intermediary node or cluster of nodes), the causes of the election swing (further intermediary nodes), the current financial and political situation in Greece (further intermediary nodes), which leads back to the beliefs and affects about the morality of the actions of the Troika (the first intermediary node or cluster of nodes). After ruminating on these thoughts for some time, she writes a plan of her answer (behavioural output). This illustrates an internal recursive loop and how it operates in a continuous flow CAPS. The continuous flow description also allows us to explain why behaviour emerging from an internal recursive loop will be different to that emerging via the same nodes, but without recursion, since the influence of those nodes which are repeatedly represented will be greater on the behavioural output and thereby shape the outcome to a greater extent.

These examples demonstrate the ways in which grounding CAPS in continuous flow networks resources the sorts of explanation which CAPS wants to offer in these sorts of cases. However, a significant resource offered by feed-forward networks remains unaccounted for in continuous flow networks; improvement by backpropagation. This mechanism explained error correction in behaviour as a response to feedback in the environment that operates by correcting the weights assigned to the connections and nodes that caused the behaviour. It also explained how those corrections could become specialised and attach to the wrong features of situations, either by misidentification of the problem or by over- or under-correction in relation to that problem.

Continuous flow networks cannot be back-propagated, since they do not stop being forward-propagated, but rather run continuously. Furthermore, because the effect that an input has on a node depends on the timing of that input being within the storage capacity of the node, running the same information back across the same system will not follow the same routes, let alone mirror the same weights in reverse, in the way that feed-forward networks do.

Backpropagation offered grounds for one of CAPS′ key explanatory virtues; the explanation of personality change. So, if CAPS is a continuous flow system, and thereby does not have the option of backpropagation, how can we ground improvement and personality change?

By far the most prominent candidate for a continuous flow network is Hebbian learning (Hebb, 2005). Hebbian learning describes how the connections between neurons change with their coactivation. While the basic Hebbian theory has been superseded by more refined versions within neuroscience, they retain the key problems as grounds of CAPS. As will be clear, these problems are peculiar to the abstract model CAPS presents but not for their applications in neuroscience. After presenting a basic version of Hebbian learning and outlining the solution such learning offers a continuous flow system, this section closes by illustrating why the explanations of improvement and development offered by CAPS cannot be grounded in Hebbian learning systems.

Hebb (2005) describes how, when the firing of one neuron causes the firing of another, the connection between those neurons strengthens. This is Hebb′s ′neurophysiological postulate′:

′When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.′ (2005, p. 62)

This change, or process of growth, is represented by a greater propensity of the connection to cause the activation of the receptor node given the activation of the effector node. In continuous flow models, this is represented by an increase in the connector strength of the connection *AB* between two nodes, A and B, when A causes the activation of B. This strengthening is sensitive to the direction of the connection, meaning that the strengthening of *AB* does not entail the strengthening of *BA*. Furthermore, it is dependent on the causal relationship between the activation of A and the activation of B. Mere co-occurrence, for Hebb, is insufficient to strengthen the connection between the nodes.

The network in figure 6.8 is now governed by a basic Hebbian learning algorithm. The output of the system is a holistic representation of the patterns of activation, which continuously represents the active nodes and their weights. In such a network, when one node activates and, thereby, causes the activation of a recipient node, the connection between the two nodes becomes stronger such that the propensity of the effector node to activate the receptor node increases. As inputs are introduced to the network at A, the activity flows across the network causing a pattern of output. This pattern represents the weights of the nodes that are active as well as the order and timing of their activation. When the same input is introduced a second time, the pattern will be different. Each of the nodes that were connected in the first pass of the input will have been strengthened by their activation. In the second pass this is represented by the increased propensity to follow the previous path over close alternates. With many, varied inputs, the system will strengthen those connections that are well connected to the relevant inputs over those with weak connections.

For example, take the case of Andrew from the previous section. During his mathematics exam, Andrew writes out the formulae which he associates with the question topic. This helps Andrew to answer the question due to the close connection between seeing the formulae written out and the relevant mental states needed to answer the question. This may be explained by looking at Andrew′s revision strategy, which focussed on how to utilise the relevant formulae over repetition of past papers. By doing so, the connections between the formulae and the relevant beliefs were strengthened in a way that the connections between reading a question and the relevant beliefs was not.

This offers an explanation of how a continuous flow network may develop through Hebbian learning, as well as how such an application may make sense of a in the context of CAPS. However, there are two problems with the application of Hebbian learning to CAPS. The first is the source of connections, the second is content density and sensitivity to such content.

Hebbian learning describes how *AB* may strengthen following A′s activation causing B′s activation. However, since Hebbian learning requires the causing of B by A, *AB*must already be established before it is possible to strengthen the connection. This is not a problem for neurons, which can rely on their physical proximity to one another to establish a connection which may then be strengthened. Similarly, feed-forward nets connect each node in one layer to every node in the next layer. The established layer structure of feed-forward nets is analogous to the physical proximity for neurons.

For an abstract model like a continuous flow CAPS, however, there must be some other governing mechanism for the establishment of connections. This is because CAPS lacks a mechanism which would be analogous to the role of physical proximity in the neurological systems Hebbian learning was designed to explain. Moreover, it does not allow for connections to be forged by experience or by reasoning. As a model of personality, CAPS is supposed to explain how inputs from the environment are processed and how the processes of reasoning occur. Since it explains those processes in terms of the connections that are already present between nodes, the presence of connections between nodes cannot be a result of either of those kinds of process.

CAPS aims to explain an individual′s pattern of behaviour in terms of the connections between nodes. Without an explanation of the origin of those connections, a Hebbian CAPS ultimately cannot explain why an individual′s pattern of behaviour is one way rather than another. Furthermore, making predictions about changes in behaviour requires an account of how connections are established, in order to prescribe the events that would establish the necessary connections or to control for such events while existing connections are strengthened or weaken.

The second problem is that CAPS is offering explanations of informationally dense interactions. When a particular behaviour receives negative feedback, CAPS is meant to be able to adjust the weights of the connections and the nodal weights accordingly (Mischel and Shoda, 1995: 262). This requires a sensitivity to the content of the nodes and connections which should, perhaps at minimum, be a sensitivity to the negation of the content of a given node. We should be able to explain how it is possible that the belief that P is sometimes in some way sensitive to the experience that −p. Hebbian learning is not sensitive to such content. Figure 6.7, for example, requires something much stronger and more informationally rich than Hebbian learning offers.

In the absence of a sensitivity to negation we would need to explain why feedback that negated the connection between two nodes has the effect of relatively weakening the strength of that connection. A student, Michaela is asked about the cause of the start of World War 1. She responds with her belief that it was the murder of Archduke Franz Ferdinand that started World War 1. Michaela is corrected by her history lecturer who states that the murder was simply one event among many which caused the war and, absent the murder, the war would have occurred nevertheless. What has happened to the connection between the student′s beliefs about the murder of Franz Ferdinand and the start of World War 1?

A Hebbian account would see the extant connection between these beliefs, a connection where one node, the start of World War 1, caused the activation of the other, the murder of Franz Ferdinand, via that connection. Following Hebb′s principle, this connection would then strengthen. This strengthening would make it more likely that, in a similar situation, asked a similar question, Michaela would offer the same response. To demonstrate the absurdity of this conclusion, suppose the lecturer had instead agreed and praised Michaela. Hebbian learning would predict that the effect on that connection would be the same: they fired together, so they wired together. This is what is meant by a lack of sensitivity to negation, or more generally to the content of the nodes and the connections. It should be possible to weaken connections that fired together based upon negative feedback, but this is not possible on a Hebbian account.

This challenge extends to more sophisticated Hebbian learning. The challenge bites against basic Hebbian learning because of the lack of sensitivity to content inherent in a learning mechanism which simply strengthens on the basis of activation. More complex Hebbian-style learning mechanisms, such as spike timing dependent plasticity, do allow for the weakening of connections between neurons (Bi and Poo, 2001). However, this weakening only occurs when the receptor node is has recently ′spiked′. This means that a connection between nodes A and B will weaken if the connection *AB* fires shortly after the activation of B. This condition fails to capture the kinds of weakening of connections that sensitivity to content would require: the Michaela example is expressed in terms that fit this more nuanced learning mechanic. Insofar as more sophisticated articulations of Hebbian learning retain the basic principle that the connections strengthen with activation, i.e. that they retain some version of Hebb′s ′neurophysiological postulate′, they will fail to capture these cases where content sensitivity is essential to the level of explanation CAPS seeks to offer.

Without an adequate mechanism of change, or refinement, continuous flow systems fail to offer the grounds for CAPS′ core explanations: general behavioural stability and gradual change.

Section 6.6: Concluding Remarks

The preceding sections proceeded from the application of the first four challenges of the ontological critique. CAPS presented the explanation of general behavioural stability alongside the explanation of gradual change as its key aim. By interrogating the grounds CAPS purports to be able to rely on in offering its explanations, I demonstrated that the resources to support those explanations in a single framework can be found neither within dynamical systems theory nor within continuous flow connectionist systems. The former offers no clear way to ground explanations which utilise the cognitive-affective units, and thereby attitudes, that CAPS′ explanations rely upon. The latter can ground the use of such tools, but cannot offer an explanation of gradual change and refinement that CAPS′ key explanations depend upon.

Several conclusions should be drawn from this exploration. First, CAPS′ explanations are not as robust or coherent as they appear to be upon initial inspection. Second, the analysis presented above is not an exhaustive engagement with the possible ways CAPS could be grounded. Third, the utility of the ontological critique is demonstrated twice over through the preceding explorations. Initially by identifying the questions we need to put to CAPS to clarify the ontology that CAPS purports support its explanations and later by applying those same techniques to the analysis of how we might ground CAPS in the two primary candidates. CAPS offered us the aims, answering the first challenge, allowing us to proceed to identify the relevant objects, their properties and how the conjunction of these purports to offer the explanations CAPS requires.

These conclusions drawn from the first four challenges lead us to the final challenge of the ontological critique: prediction.

CAPS offers several candidates for predictions which should be taken forward for experimental testing. These include, but are not limited to, the role of attitude strength in structuring behaviour, the introduction and subsequent effects of novel cognitive-affective units and connections between them, the range of stability of behavioural signatures, and others. What they have in common is that the details of the prediction depend on questions and challenges already raised.

If we are to experimentally test the claim that heightened attitude strength can alter behaviour in such and such way then we need to decide what the mechanic for proceeding from an influential attitude to behaviour is. If we wish to ground our understanding of CAPS in dynamical systems theory this may require us finding a way to conceptualise the cognitive-affective unit talk in terms of the properties of nodes and areas, while grounding our understanding of CAPS in connectionism requires that we have a clear conception of how behaviour relates to changes in the system.

This is not to say that no such predictions can be made on the basis of CAPS, rather it is to highlight the importance of making such theoretical commitments, explicitly, and subjecting the resulting refined model to potential experimental disconfirmation. As identified in chapter 2, when we do so we must be clear about the prediction being made, what it would take to falsify said prediction, and what would be at stake if such a falsification were to occur. This applies equally to the analysis of CAPS *per se* and to the analysis of CAPS as a model of attitudes within attitude psychology.

This chapter has not demonstrated that CAPS is incoherent or even that it is a bad model of attitudes. What it has shown is that there remain significant areas in which work must be done to refine the claims CAPS makes to bring those in line with the claims for which it has warrant, or to solidify the grounds of CAPS to expand its warrant to meet its claims. Some of this work is theoretical, some experimental. Lasting progress for the research programme will require both.

Chapter Seven: Where we go from here.

Section 7.0: One last introduction

The ontological critique represents an approach to theory improvement in psychology informed by, and in the spirit of, recent work in social epistemology. It takes the large, abstract problem of developing a more robust theoretical framework for social psychology and presents a toolkit for breaking down the justifications of the scientific judgements made within the literature. While it does so in a relatively procedural, formal manner, the approach is intended to do the work needed in social psychology not by resolving problems with theory by its mere application, but by unpacking the construction of conclusions to afford the transparency needed for precise, constructive criticism. It is hoped that this tool will help drive the creative development of theory in a field in crisis. This chapter clarifies the scope of the critique, by drawing together the case made for its implementation across the preceding chapters and by situating it in relation to other relevant projects in the literature. It closes by highlighting areas for future research and two substantive problems for the thesis which remain unresolved.

Section 7.1 draws together the practical case for the critique made in chapters 4, 5, and 6. Section 7.2 makes the corresponding theoretical case for the critique from chapters 1, 2 and 3. Section 7.3 relates the ontological critique to the ongoing replication attempts in the literature as a drive toward making social psychology a more progressive research programme. Section 7.4 relates the critique to some of the ongoing work in the literature on the theoretical accounts of replication in psychology, particularly Irvine (2021), demonstrating the fruitfulness of the critique as a constructive response. Section 7.5 discusses another attempt to improve the theoretical framework in psychology by adopting a theory from a related field. The ontological critique is contrasted with this approach and some grounds to prefer the approach of the ontological critique as a strategy are presented.

The ontological critique is about identifying and clearly stating the warrant we have for our conclusions and the relevant populations of extension over which those conclusions remain warranted. In this spirit, sections 7.6 and 7.7 present two *mea culpa*, two problems for the critique as it is presented and argued for within this thesis. Section 7.6 presents what I term the objects and properties problem raised during the application of the critique to CAPS. This problem presents us with reason to be concerned that the critique imposes ontological categories on theories which may be inappropriate or misleading. This is particularly worrying for theories which have non-WEIRD origins. Section 7.7 presents the Lakatos problem, where responses and criticisms of Lakatos′ approach to the philosophy of science, especially from Feyerabend and Kuhn, are relevant and potentially problematic to the approach adopted throughout this thesis because of its Lakatosian roots. Each of these problems is presented with a partial response to the problem, but each response requires further research to make robust. The chapter closes the thesis with section 7.8, summarising the key conclusion of the thesis.

Section 7.1: Practical case for the Ontological Critique

The question of whether or not we have good reason to think that the ontological critique succeeds in its aims can be addressed both practically and theoretically. The practical question asks whether we have enough evidence of the practical efficacy of the critique. This section partially answers this question by summarising the findings of the preceding three chapters, before finessing the question and asking what it would take to offer a more robust practical response.

Chapters 4, 5 and 6 each addressed particular models within attitude psychology. In doing so they collectively make a practical case for the efficacy of the ontological critique by, in each instance, applying the critique in a way which was either fruitful in identifying challenges, or diagnosing challenges, or resolving challenges, or indicating the direction in which solutions seem likely to lie, or some combination of the above. Each of the three models were outlined in their own terms and then broken down through the lens of the ontological critique. Doing so identified key concerns with each of the three models and indicated avenues for heuristic progress. These chapters focus on the practical problems within research programmes which the ontological critique offers us traction in addressing. They illustrate implementations of the critique with a variety of emphases highlighting the flexibility of the ontological critique – from offering simple clarification (chapter 4), to challenging how we present and understand our models – especially when our models are composite (chapter 5), and to philosophically explore the roots of the epistemic warrants our models purport to offer (chapter 6).

Chapter six applied the ontological critique was to the Cognitive-Affective Personality System (CAPS) model (Mischel and Shoda, 1995; Webber, 2015; 2016). This application required clear conceptual analysis of precisely what was offered to the empirical literature by the CAPS model. What CAPS is trying to explain, and what conceptual means it employs in doing so. The initial outline of CAPS highlights how, framed in its own terms, CAPS presents a prima facie plausible account of its explanations of phenomena, and of our justification in making further predictions on this basis. This implementation highlighted challenges to grounding the epistemic warrant for the explanations offered by CAPS in either dynamical systems theory or connectionism. In the case of the former, it is not clear that the foundations of the model in a dynamical system could give us warrant for the explanations in terms of cognitive-affective units which CAPS purports. For the latter, there is not a clear candidate for grounding the updating mechanism which CAPS relies on in its most central explanations. These challenges, while not necessarily insurmountable, are non-trivial and the manner in which we try to resolve them was shown to have significant consequences for our explanations derived from CAPS.

This implementation of the ontological critique represents a relatively theory-heavy approach to model analysis. It specifically highlighted the grounding of explanations and explored the assumption that adequate grounds would exist within connectionism, or within another strong contender for such grounds, dynamical systems theory. This illustrates one kind of implementation of the critique: do our conceptual resources do the work for us, which they seem to at first glance? The ability to explore these questions with clarity and nuance is one way in which the transparency afforded by the critique can help advance our models and bring about heuristic progress within our research programmes.

The ontological critique gives us a framework and structure to allow us to identify strengths and weaknesses of models. In this case, as in others, the ontological critique is not essential to the criticisms made, but it offers us a toolkit for clarifying the goals of the model and the manner in which the model purports to meet them. The application of the ontological critique to CAPS emphasises its utility in identifying the many parts of these, often very complex, models in a way which enables us to clearly state our questions and challenges.

Chapter five addressed the mainstream model of attitudes (Haddock and Maio, 2012; 2015; Tanesini, forthcoming) and its situation within a broader MODE model (Fazio et al, 1986). Following Tanesini (forthcoming), the mainstream model was presented through the features of attitudes and the properties of attitudes. This presentation of the model invites some *prima facie* challenges which are readily addressed by situating the mainstream model within an overarching MODE account of attitudes (Fazio et al, 1986). The application of the ontological critique focussed on the relationships within the composite model and apparent duplications of function.

Applying the critique to MODE-mainstream highlighted that the integration of one model within another is not always a straightforward process and emphasised the utility of the ontological critique in two ways. The first was the utility of the ontological critique in identifying where the models were 'talking past' one another – where differences in their frames of reference made the implementation of the composite model more complex than strictly necessary. The second utility lay in highlighting the incorporation of the prima facie challenges in the operation of the ontological critique – the prima facie challenges therefore represent an advantage of the critique over the features/properties approach used by Tanesini (forthcoming).

With the critique applied, chapter 5 identified the mainstream model as a causal model of the aetiology of an attitude, while the MODE model is a decision tree which describes the effect of such an attitude on behaviour under different circumstances. The two models provide answers to closely related, but distinct, questions in different manners.

The utility of the ontological critique illustrated in this case lies in offering greater scope to clarify a model in a way which heads off prima facie challenges, and which explicitly situates the model in a broader theoretical framework. The former provides us with a structured way to provide clarity in our model. The latter enables us to utilise composite models more readily in offering our explanations while avoiding the confusions the combination of these models sometimes generates.

Chapter four engaged with Machery′s arguments for the trait picture of attitudes, as well as his argument against the Freudian picture. By clarifying the commitments of the Freudian picture which were party to the successful *modus tollens*, chapter four identified the target of the negative programme as the commitment to the individuation of implicit attitudes on the basis of high automaticity and low introspectability. This commitment is relatively widespread within the philosophical literature on attitudes and is almost endemic to the psychological literature.

Applying the ontological critique to Machery′s positive programme, the chapter further demonstrates that the strength of the trait picture is that it offers a model which avoids the commitment to the individuation of implicit attitudes on the basis of high automaticity and low introspectability. The challenges of the ontological critique identify how the trait picture constructs explanations that avoid this pitfall while also interrogating the potential of the picture to offer falsifiable predictions.

The exploration of the falsification challenge further refines and clarifies Holroyd′s (2016) and Holroyd, Scaife and Stafford′s (2017) criticisms of Machery by identifying the areas in which the model lacks precision and highlighting the directions for heuristic progress for the model. Interrogating its possible falsifications makes clear that the trait picture requires the introduction of auxiliary hypotheses that enable us to make predictions which may be tested.

These three applications highlight various utilities of the critique. The chapters together go some way to demonstrating that while such criticisms or recommendations could be made for the target models without the ontological critique, its use enables and calls for a more rigorous and thoroughgoing examination of our models. Furthermore, it can do so in a way which is transparent to readers – by identifying our goals in utilising a model and the construction of our explanations and predictions constructive, precise criticisms are invited. This answers the easier of the practical versions of the ′does it work?′ challenge: here are some cases where it seems to have done so.

The harder version of the question may be phrased something like: do we have good reason to think that the ontological critique brings about heuristic improvement in research programmes when employed by practitioners in those research programmes? This version does not admit of a full answer in this thesis, for two key reasons. First, it is an empirical question which would require well designed, large-scale, empirical studies to adequately address with a practical answer. Secondly, the Lakatosian framing of the thesis makes this a historical question (addressed further below) which can only be answered by the long-term movement of the research programme toward greater explanatory and predictive power. In the absence of a good practical response to this challenge, chapters 2 and 3 offer a theoretical response; that we have good theoretical reasons to adopt the critique.

Section 7.2 Theoretical Case for the Ontological Critique

This more abstract, theoretical case for the ontological critique was expressed in chapters 2 and 3. These chapters addressed a variety of responses to contemporary problems in psychology and identifies where these responses succeed and where they miss their mark. This part of the thesis made the case for existence, and importance, of a gap in our available tools for evaluating and improving the theoretical frameworks of psychology and presented a candidate which fills the gap: the ontological critique.

Chapter three unpacked Holroyd′s (2016) and Holroyd, Scaife and Stafford (2017)′s approach to model choice: desiderata for a model of implicit cognition. The chapter began by identifying that the desiderata presented in the two papers in fact reduce to two, with the advantage that these two desiderata also accord with, and express project-specifically, the desirability of explanatory and predictive power in our models. The chapter went on to present a challenge to the desiderata approach – undecidability between cases.

Undecidable cases arise when models offer reasons to exclude some apparent evidence from the set of relevant evidence. The first desideratum, project-relevant explanations, then fails to operate as expected because the target dataset is no longer shared. This is compounded by the fact that simply treating ′explaining away′ and ′explaining′ as relevantly equivalent fails to capture the way that, if we ought to ′explain away′ some datapoint, the explaining of that datapoint by a model counts against, rather than for, the model.

The chapter closes by demonstrating that such divergences in explanation occur within the literature – an easy case between MODE-mainstream and CAPS (addressed using the ontological critique in chapters 5 and 6 respectively) and a hard case of the trait picture (addressed in chapter 4). This does not rule these and similar desiderata out as useful in improving our understanding of psychology, but it does highlight that they are inadequate tools for tackling the current challenges faced in the field and that an approach which can accommodate these hard cases is, to that extent, to be preferred over a desiderata approach.

Chapter two presented two statistical approaches to improving the reliability of claims in psychology. The urgency of the problem was framed by addressing the option of disregarding social psychology as providing us with novel evidence for its claims and conclusions. This framing highlighted the importance of finding or creating tools which improve our ability to rely on social psychology. Specifically, how can we be sure that a given finding in the extant literature and the findings of research going forward is established by a reliable method?

The first approach the chapter discussed was selecting for p<0.005 and treating only those studies as providing reliable evidence for their stated conclusion. This approach is closely related to, and inspired by, Benjamin et al.'s (2018) suggestion that authors and editors redefine statistical significance as obtaining p-values bellow 0.005 rather than the currently accepted p<0.05. The suggestion operates on the same principle, that selecting in this way improves the positive predictive value of the results on which we rely. That is, selecting studies which obtain p<0.005 means that we are more likely to be relying on true effects than selecting for p<0.05, even for low prior probabilities of the test hypothesis.

Presenting selecting for P<0.005 as a tool for readers invited a novel challenge. While the strategy was recommended because of its efficacy in raising the PPV of the studies on which we rely, the approach reduced the available dataset of findings on which we may rely so substantially that it may be prohibitive of some areas of study. In practice, this means that for many research questions in social psychology, adopting the strategy for readers is little different in practice to disregarding social psychology as providing novel evidence for its conclusions, an approach which is undesirable if it is avoidable. While the heuristic is both epistemically undemanding and efficacious, this cost makes the strategy prohibitive of many interesting and important research questions.

To avoid the contraction of the literature available to engage with, where possible and reasonable, while maintaining the same consideration of maximising the positive predictive value of the findings on which we rely, the chapter went on to propose the implementation of PPV/prior curves by readers and authors. Good practice in applying these curves was outlined and the relatively high epistemic cost of the approach was raised as a concern for its implementation by readers. For authors the approach is relatively undemanding compared to the existing statistical approaches employed in psychology. For readers, it is significantly more demanding especially where post-hoc power calculations are required. This epistemic burden was argued to be worth adopting in order to pursue those research questions where p-values do not reach the more stringent criterion as it enables researchers to continue to explore these important research questions with transparency about the strength of the evidence for their claims. The approach also allows for nuanced interpretation of the strength of the data offered by a publication and enables us to be sensitive to Bayesian criticisms and concerns without abandoning our existing statistical practices.

These approaches, while doubtless useful and perhaps essential to improving psychology′s reliability, only target the strength of the statistical basis of the conclusions offered in a publication. They neglect the theoretical framework within which that conclusion is tested and expressed. They address only one part of the reliability of the overall method for establishing conclusions. The chapter closed by introducing the ontological critique as a candidate approach to interrogating our theoretical frameworks; by emphasising the transparency of the theoretical construction of the explanations and predictions of the model the ontological critique targets a dimension of conclusion construction which is both missed and occluded by the statistical approaches to improving reliability. This is presented as a companion tool to the statistical approaches which increases the reliability of another part of our overall method for arriving at, and evidencing, a given conclusion.

Together, these chapters offer us theoretical grounds for thinking first that we need a tool which targets the area of conclusion generation which is missed by the statistical approaches, and second that the ontological critique offers us such a tool. This is not so much an argument that the critique works, but an argument that we need some systematic approach and that the ontological critique offers us such an approach.

In combination with the evidence of the practical utility of the ontological critique in chapters 4, 5 and 6, these theoretical grounds give us good reason to be optimistic about the prospects of the approach if it were to receive somewhat widespread adoption. Specifically, we may be optimistic about its potential to fuel and facilitate heuristic improvement in the research programme of social psychology.

Section 7.3 Replication attempts and the Ontological Critique

The thesis began by framing concerns of reliability in psychology through the lens of the replication crisis in psychology, and especially in social psychology. This framing focussed initially on the case of Bem (2011) and the ensuing controversy and response among psychologists. Some responses were raised which were important to address argued that we had reason to think that Bem′s method was anomalous, and that problems with that method did not generalise to social psychology *per se*. While each of these responses has merit, and deserves consideration, they were each found to either address problems for method in psychology generally, or offered considerations which would make Bem′s method anomalous, but which significantly understated the problem.

The first chapter went on to make a more direct case against the method of social psychology by addressing the findings of the Open Science Collaboration (2015), and exploring criticisms of the findings from Gilbert et al. (2016) These findings display a worrying rate of replication failures in social psychology, over a more representative sample than that of Bem (2011). Moreover, responses from Gilbert et al. (2016) do not undermine the conclusion that we ought to be concerned about the reliability of the conclusions drawn in social psychology. Rather, they refine the conclusion to particular concerns about specific, systematic sources of unreliability.

In order to highlight these specific consequences of failed replications, the first chapter introduced and outlined Machery′s (2017) argument for a resampling account of replications. Under this account an experiment X is a replication of experiment Y if X samples some component(s) of Y from the same population(s) that Y sampled. This account is shown to have several advantages over Schmidt′s (2009) taxonomy.

First, the resampling account offers clear prescriptions for experimenters looking to design a replication by identifying the target claim and distinguishes the ways in which these generalisations operate within our conclusions. The account clarifies that a replication is testing a generalisation from the original sample to a given population as part of the construction of the original conclusion. Furthermore, it distinguishes between the different ways in which this is done to target generalisations across populations of experimental units, treatments, measurements, and settings. Second, it clarifies the dimensions of extension for an experiment′s conclusion. A conclusion may be extended by extending an existing generalisation to a new population and testing the new extension. For example, a conclusion which applies to university graduates may be extended to include professional course graduates, and an extension study is testing this novel generalisation. Alternatively, a conclusion may be extended by changing a fixed factor for a new fixed factor or treating what was a fixed factor as a random factor to be sampled. For example, an experiment which used a given dosage of a trial drug may test a new dosage to see what effect that has, or it may test a variety of dosages to try to generalise to a distribution of effects across a population of treatments.

Applying this account of replications highlights a further advantage of the resampling account: it identifies the theoretical burden of a failed replication for the original finding. In the case of the replication of Schnabel and Nadler (2008) as part of the Open Science Collaboration (2015), Gilbert et al. (2016) highlighted experimental differences which they took to undermine the legitimacy of the second experiment as a replication. On the resampling account, not only is the second experiment clarified to be a replication, but the consequences of the response from Gilbert et al. for the original conclusion is identified as the requirement to curtail the generalisations which were made by Schnabel and Nadler. I note that such curtailment is undesirable, and perhaps absurd, given the hypothesis Schnabel and Nadler were testing. Given this undesirability, or perhaps absurdity, this forecloses the avenue of response from Gilbert et al.

The original framing of the thesis as a response to rates of replication failure in social psychology requires some situation of the subsequent argument in relation to contemporary projects in the literature on replication, and of replications. This ongoing work includes practical replications of findings in the literature, purely theoretical work on how we should engage with and understand replications, and work which interweaves the two. This section closes by discussing the relevance of the practical work of large-scale replication attempts and the hybrid work in replication markets. The relation with theoretical work on replication is the topic of section 7.4.

The practical work especially includes those large-scale multi-lab replication attempts that follow in the vein of the Many Labs Project (Klein et al., 2014) and the Open Science Collaboration (2015) discussed in chapters 1 and 2. Some contemporary projects include the more recent Many Labs 2 (Klein et al., 2018) which studied the individual variability in sampling experimental units. Many Labs 3 (Ebersole et al., 2016) tested the variability within student populations (widely used in psychology) of several relatively well understood effects. Many Labs 4 (Klein et al., 2019) attempted to establish whether a large-scale replication of the mortality salience effect would be influenced by the involvement of the original author (author advised labs) compared with the control (ʻin-houseʼ labs). Unfortunately, all labs failed to replicate the effect leaving the effect of original author involvement unclear. Many Labs 5 is currently investigating the effect of pre-data collection peer review as an intervention to increase replicability (Ebersole et al., 2020).

These projects and others like them highlight the importance of a resampling account of replications to identify where generalisations in original studies are, or are not, well founded, as well as to identify common confounding effects. The resampling account allows us to identify each of the experiments of Many Labs 4 as a replication. In each case, the labs were resampling the experimental units, attempting to use the same treatments and measurements, with some labs also deliberately resampling the setting. Simultaneously the account clarifies that the overall project represents an extension of the original experiment to try to isolate the effect of original author involvement on replication rates.

The projects also further accord with the implementation of the ontological critique to the underlying hypotheses which are being tested. By clarifying the proposed mechanism of the effect and its operation within, and interaction with, experimental treatments and measures, the dimensions of generalisation which are the target of replications can be clearly identified. Greater clarity about the generalisations we are testing allows us to more precisely state the populations which are licit and illicit to sample and give principled justifications for these bounds and respond to criticisms made to the results on these grounds.

Another strand of practical replication attempts in the contemporary literature is the attempt to identify the predictability of replication successes and failures through replication markets. This work attempts to run parallel replication attempts and replication markets, where the ability of participants in the latter to identify studies in the former which will replicate, for financial incentive, is tested alongside the actual replication of the study (Dreber et al., 2015).

Replication markets represent an interesting challenge for the ontological critique, especially as I have argued for its implementation from a position of the epistemic problem of replication failures in social psychology. Dreber et al.′s findings indicate that it is possible for psychologists to predict replication rates of studies with significantly better than random accuracy, though still quite imperfectly. That is, given the incentive and explicit drive to consider the replicability of the research in question, psychologists will correctly predict whether a replication will elicit or fail to elicit an effect in the same direction (p<0.05) about 70% of the time (Dreber et al., 2015, p. 15344). Without this incentive and drive to consider replicability psychologists continue to cite studies that do not replicate. As mentioned in footnote 27, Yang et al. (2020), present their somewhat unsettling finding that, in line with Dreber et al.′s finding replication can be predicted, by prediction markets, expert surveys and machine learning algorithms designed around the prediction market behaviour:

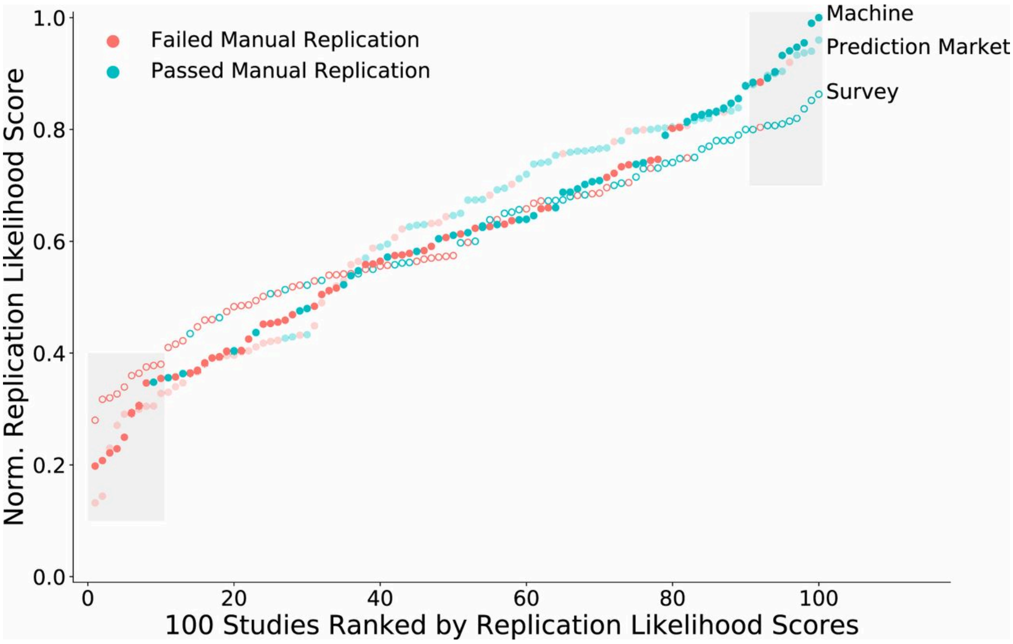

Figure 7.1: Performance of prediction markets, surveys, and machine learning algorithms (Yang et al., 2020, p. 10765).

Psychologists continue to cite studies which will not replicate at near identical rates to studies that will:

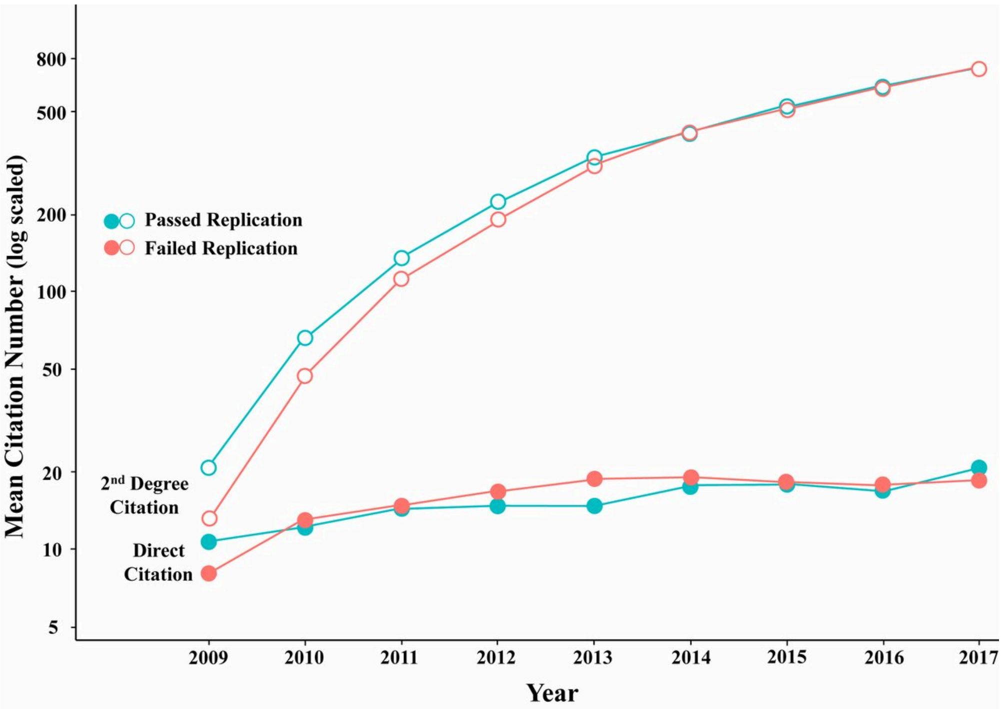

Figure 7.2: Direct and second-degree citation rates of replicating and non-replicating studies. (Yang et al., 2020, p. 10763)

One interpretation of these findings, which accords with the case I offer for the ontological critique, is that the demand and incentive to consider a given study in terms of its replicability first and foremost enables experts to distinguish reliable and unreliable studies in a way which they do not absent this incentive and impetus. The proposal of the ontological critique does not claim that there is anything particularly arcane or obscure about identifying the claims for which we have warrant. Rather, it claims that by asking the right questions and being careful about how we construct our conclusions and identifying what we do and do not have good warrant to claim, the ontological critique offers a relatively straightforward approach to outlining the relevant information from which judgements about reliability can be made.

On this interpretation, the answer to the related question of whether we even need the ontological critique, given that psychologists are able to make these judgements without it, is that the critique aims to provide the impetus and conditions which drive users to undertake these considerations in a careful and transparent manner. The critique makes the underlying working of these judgements explicit to those making the judgements allowing them to identify and work to remedy errors or shortcomings. Implementing the critique as part of publication or registrations practices further allows for the clear communication of the justifications of judgements to readers.

Section 7.4: Replication Theory and the Ontological Critique

Relating the ontological critique to the practical replication attempts underway in the literature invites us to explore how the ontological critique fits with our theoretical understanding of replication attempts. This section highlights one insight the ontological critique offers replication theory by presenting a challenge to the understanding of replications in psychology from Irvine (2021). By reframing the understanding of the value of replications in terms of the resampling account of replications (Machery, 2017), this section concludes that replications are not only useful for theory generation but are essential, but that Irvine is right that a far richer understanding of our broader theoretical position is needed. This conclusion echoes calls from Machery for authors to specify their generalisations and the populations to which these generalisations are taken to be valid.

The ontological critique is framed as a response to both of these challenges: specifying our generalisations over relevant populations and offering clear theoretical justifications for the conclusions we draw and the replications we conduct.

Irvine (2021) argues that significantly more theory development is required in a field before ʹgoodʹ replications can be carried out than is currently possible in psychology. For Irvine, this concern is compounded by a worry that even ʹgoodʹ conceptual replications offer very little theoretical pay -off. These two points together lead to the conclusion that aiming at replication in psychology is misplaced, with replications better characterised as explorative studies than as an attempt to comment on, or re-address, a previous study. This has implications for current practices and language used around replications, which Irvine draws out.

Irvine (2021) follows Schmidt′s (2009) taxonomy of replications into direct and conceptual. Following Stroebe and Strack′s (2014), Irvine also takes all replications in psychology to be conceptual replications, since the materials or operationalisations of an original study were designed and developed for a target population or context, and may not generate the same phenomenon when these are varied. Fabrigar and Wegener (2016) recommend that our analysis of replication studies instead focused on degrees of psychometric invariance, where some studies try to recreate the same psychological conditions as the original study, while others deliberately extend the original experiment to a new population.

To preserve psychometric invariance in a study, the designers need a robust idea of how differences across populations might affect the efficacy of the original stimuli in a new sample. Given the sheer variety of potentially relevant differences and the rate at which confounding factors exacerbate one another ″there is often no well-developed and widely accepted set of background theory that can be easily referred to in order to confidently and precisely inform the design of good direct replications.″ (Irvine, 2021, p. 3)

Similarly, replications using new operationalisations in new settings require further understanding of the measurement procedures and operationalisations involved. In order to utilise some new procedure, one must have a good understanding of how to intervene on the target phenomenon and asses the outcome of the new intervention. This requires more than simply understanding the differences to which the effect may be susceptible, including some understanding of the general causal profile of the target phenomenon. What is hypothesised to cause it and how this relates to other key factors of the setting, measurement and treatment.

Finally, it is necessary to know the level of variance between the original study and the replication to know if one is looking at a relatively psychometrically invariant replication or not, and hence how to evaluate the results obtained. In order to know the level of causal independence, a robust understanding of the causal profile of the phenomenon is necessary. As Irvine highlights, this is significantly ″messier than counting mere changes in experimental design″ (2021, p. 4) as two studies with different populations, treatments, measurements and settings may well be psychometrically invariant − successfully targeting the same psychological phenomenon − because of the differences between them rather than in spite of them.

These concerns about the level of theory needed before one can conduct ′good′ replications are compounded by a relatively low theoretical benefit of replications. Irvine highlights that relatively psychometrically invariant replications, when successful, simply demonstrate that the procedures do something in a reliable way, though it is not clear from that repeatability what that is. Similarly, when variance is deliberately introduced Irvine cites Meehl′s comment on the theoretical understanding needed to interpret findings:

> ″For example, [say] Meehl′s Mental measure correlates .50 with SES in Duluth junior high school students, as predicted from Fisbee′s theory of sociability. When Jones tries to replicate the finding on [Mexican American] seniors in Tucson, he gets r = .34. Who can say anything theoretically cogent about this difference? Does any sane psychologist believe that one can do much more than shrug?″ (Meehl, 1978, p. 814)

Irvine concludes that ″even good conceptual replications, in themselves, do not support very strong claims about the adequacy of relevant theory.″ (Irvine, 2021, p. 850) That is, without far richer theoretical frameworks than are currently available, even good replications tell us very little about the relative strength or weakness of the underlying theories or hypotheses.

Given a lack of this kind of robust theoretical framework from which we can run good replications, and the relative poverty of the theoretical payoffs of such replications, Irvine concludes that treating replications as exploratory rather than confirmatory better recognises their role in theory production and changes expectations around successful or failed replications to account for this. If Irvine is right, this means that the large-scale replications discussed earlier in this chapter tell us little to nothing about the original experiment since the field lacks the broad theoretical understanding necessary to preserve the psychometric invariance needed to retest the original hypothesis.

I propose that adopting a resampling account of replication heads off the concern about the contribution to theory of replications and thereby offers a reason to be more optimistic about theoretical progress and the fruitfulness of replication. The concern Irvine raises is that without robust theoretical understanding of the phenomenon, we cannot adequately justify generalising any given finding across a reasonable population. This is a concern for replications because in order to design a potentially successful replication[71], an understanding of where relevant confounding differences may arise is essential. On a resampling account, this is articulated as a challenge of having enough theoretical understanding of the phenomena that we can reasonably state the bounds of the sampled population and be able to justify our priors about the structure of the relevant distribution of the efficacy of effect within that population.

What framing this challenge through the resampling account clarifies is that this is not a challenge to replications as replications but to the generalisation from data to phenomenon. While these generalisations are essential to replications, they are also essential to the conclusions drawn within psychology publications. These publications do not, and arguably should not, conclude ′treatment X increased measure Y in population Z relative to the control, we have no reason to believe that this says anything about any given class of which X, Y or Z may be members′.

To illustrate this claim, in the current issue of Psychological Science generalisations include moving from a sample of 50, mostly young adults (mean age = 21.2 years) to the conclusion ″some types of music can disrupt

[71] An experiment which has a reasonable chance of eliciting data which well evidences the same actually extant underlying phenomenon as the original experiment.

night-time sleep by inducing long-lasting earworms that are perpetuated by spontaneous memory-reactivation processes″ (Scullin, Gao, and Fillmore, 2021, p. 1), generalising experimental units, treatments, and settings to conclude about a general musico-psychological phenomenon. Another paper generalises to the conclusion that ″after real social interactions, evaluative feedback about oneself that violates ones self-view modulates all processing stages with an early negativity and a late positivity bias″ (Schindler et al., 2021, p. 1) describing a phenomenon of social cognition in humans generally from a sample of 46 native German speaking, right-handed adults, with an average age of 23.7. Another publication generalised from 18 dogs (average age 4.6 years) observing their owners′ jealousy framed interaction with either a fake dog or a fleece cylinder to the claim that ″dogs display jealous behaviour″ and that the results ″provide the first evidence that dogs can mentally represent jealousy-inducing social interactions.″ (Bastos et al., 2021, p. 646)

If Irvine is right that without a far more well-developed theoretical account than we currently possess of these phenomena we are not justified in making any of the above generalisations –– that is, if all researchers are reasonably able to do with their data is report a correlation coefficient between some measure and a treatment for the population sampled, in a strict interpretation of the experimental setting –– then it is not clear that any understanding of any underlying phenomena could be gained and therefore that any science is being done.

I propose that we can distinguish between the level of theory needed to make these generalisations and the level of theory at which we have justification for believing all of our relevant generalisations to be robust. The former describes the current state of psychological theory – and explains how the above publications come by the generalisations they make. The latter describes the state of psychological theory which we would expect to systematically stand up to replication attempts – both experimental unit resampling and resampling of other components, as well as offering a reasonable degree of certainty in the likely outcomes of extensions.

This raises the question for the resampling account of what we are testing when we run a replication, given the current state of theory in psychology. When we are resampling a given experimental component we are acquiring new data on the distribution of the previously observed effect across a different sample of the original target population. This new data will either accord with the original sample, or not, to varying degrees and in varying ways. Novel finding of a replication is not, therefore, the repeatability of a given experimental procedure, but the relative robustness or fragility of the generalisations made in the original publication across the relevant populations. Replications are testing one or more of the generalisations of the original conclusion rather than re-testing the original hypothesis.

This characterisation allows us to state the value of replication attempts in an incomplete – even far from complete – science. They identify which generalisations are more robust and which are less, they may even go so far as to exclude some generalisations and vindicate others given enough replication attempts of the same experiment. This answers the second part of Irvine′s challenge, but what can be said about the first part – how can we design good replication studies?

Machery provides a response to this challenge – ″Experimentalists should make explicit, possibly in pre-registrations, whether the experimental components are fixed or random factors, and in the latter case they should describe the relevant populations.″ (2017, p. 565) By explicitly stating the populations over which the experimenters consider generalisation legitimate they give explicit license to replicators to resample those populations to test the generalisations they make. At minimum, the legitimate generalisations will include all those which the experimenters rely upon in formulating their conclusions.

Given the poor state of the theory surrounding many of these experiments it is likely that these generalisations are unreliable, but it is also likely that the practice of making these generalisations is difficult and challenging under these sorts of uncertainties. I propose that what the ontological critique offers is a framework by which experimenters can outline the theoretical framework within which their experiment is situated from which a clearer boundary of the relevant populations can be drawn. This aids in meeting Machery′s proposal, but also more generally in formulating robust, well evidenced conclusions.

Finally, it is necessary to emphasise that I take Irvine (2021) to be correct about the theory success conditions of systematic reliability of a given field under replication. Theoretical frameworks which outline the relevant features of the phenomenon in question and its proposed interaction with relevant populations of treatments, measures and settings are required for us to formulate a replication with a high chance of succeeding, but also for formulating a conclusion which has a high chance of surviving such a replication attempt.

Irvine′s argument emphasises the importance for psychology of improvement in the current state of theory in order to improve its reliability under replication. Machery (2017) provides resources for a response about the role of replications in testing particular reliabilities. Finally, the ontological critique provides a systematic approach to outlining and justifying our generalisations and the populations to which they extend – offering a pragmatic incremental way forward from the problem of systematic replication failures.

Section 7.5: Ontological Critique and other approaches

The previous section presents the ontological critique as a pragmatic solution to an ongoing problem with the theoretical framework of psychology, a problem which makes replications complex to run and conclusions difficult to justify. This section contrasts the ontological critique with other means of making progress toward a more robust theoretical framework for psychology. One avenue which deserves mention is the adoption of a well-tested theoretical framework from another discipline to act as a new bedrock from which psychology can try to rebuild a theoretical understanding of its subject matter – either by refuting particular aspects of the framework as it applies to psychology, or by presenting another competing framework which can measurably outperform the borrowed framework.

This approach is proposed by Muthukrishna and Henrich (2019) and was briefly mentioned in chapter 2. The approach emphasises the reliability of a given model to produce conclusions which are replicable now – that is, if today we swap our patchwork of theoretical commitments for the dual inheritance theory, we can hope to begin running experiments that will be more likely to replicate at an appreciably higher rate.

Muthukrishna and Henrich (2019) emphasise the aforementioned need for well-specified theoretical frameworks for a field to expect its findings to replicate relatively reliably. As they note, in contrast to textbooks from other sciences ″psychology textbooks are largely a potpourri of disconnected empirical findings on topics that have been popular at some point in the discipline′s history, and clustered based on largely American and European folk categories.″ (2019, p. 221)

Given this poverty, they propose the adoption of dual inheritance theory – which emphasises the formal modelling of genetic and cultural evolutionary processes with consideration of how these each influence and are influenced by our psychology. They propose that adopting such a model within psychology would offer a theoretical foundation which would allow researchers to make meaningful predictions about phenomena rather than offering post hoc explanations. In particular the ′WEIRD people problem′ is well tackled by adopting this framework because the framework identifies that the relevant tests involve identifying the predicted patterns of cross-cultural and lifespan variation rather than taking a hodgepodge of psychological phenomena and arbitrarily exploring their extension across less WEIRD populations based on access. (2019, p. 223)

Setting aside the specifics of adopting dual inheritance theory for psychology, there is a significant cost associated with adopting a novel theoretical framework wholesale. Any given experiment has theoretical understanding of the population of units which the experiment targets, the mechanism of the treatment and the relationship between the treatment, potential measurement outcomes and the target phenomenon. These are all framed with respect to the relevant features of the experimental setting which is further theorised – if only as being irrelevant. The experiment is designed to engage with a phenomenon only once the world has been theoretically ordered in such a way that would make the potential outcomes interpretable as representing some aspect of the target phenomenon. Changes to this theoretical framework do not make this prior work uninterpretable (the strong claim to which Kuhn is sometimes taken to have committed), rather there may be occasions on which the translation of the original research into the new programme may be problematic to the point that preserving the relevant sense of a conclusion, claim or finding may prove practically impossible (Kuhn, 1970, p.268). However, adopting an entirely novel theoretical framework presents a translation problem orders of magnitude greater than that normally faced when the theoretical framework of a field advances, even during archetypical Kuhnian "scientific revolutions".

Moving from Newtonian physics to Einsteinian requires a fundamental re-understanding of the ways in which entities relate to one another – not just to the way we understand objects but to the structure of space and time in which those objects are situated and understood. This makes translation of the findings of experiments conducted within the assumptions of Newton complicated, but largely doable, with enough time and effort. Imagine this shift were to occur but Newtonian physics was not expressed mathematically, meaning that the two theoretical frameworks do not share that common language. This translation exercise becomes a herculean task.

Were we to see the widespread adoption of a completely new theoretical framework in psychology, we should expect the task of translating existing research findings and claims into the new framework to require a staggering quantity of time and effort. It may be that, now that we have a more robust theoretical framework, we are better off abandoning the previous work to focus on where psychology goes from here, especially given that we know that some significant portion of that work is unreliable. By contrast, what the ontological critique presents is an approach which preserves the possibility of utilising the much of the existing work in psychology while building a richer theoretical framework from within the field. As well as preserving invested resources, this has the further advantage that the architects of the theoretical advance of the field be those working within it.

While the argument for the importance of the adoption of robust theory in psychology is commendable, if it is possible to move from where psychology is currently to where it has a richer theoretical framework to draw upon without a step change where the work that has been done hitherto in psychology is set aside in favour of a new understanding, then that is to be preferred. This is arguably the case even if the only benefit is the partial preservation of the vast resources already invested – researcher time, attention, ideas, creativity, to say nothing of funding.

Insofar as the ontological critique is effective in improving the theory surrounding our experimental findings, and offering principled explanations of the scope of our generalisations, utilising the ontological critique allows us to move, in a non-incommensurable way, away from the challenges of the existing literature. What this means is that, by interrogating the underlying claims behind our existing research we can continue to build on that research where it provides the foundation for us to do so. Where it does not, we have good reason to believe there was not robust existing research to build upon. In contrast, moving wholesale to a new paradigm especially a paradigm with so little in common with many of social psychology′s extant theoretical commitments, leaves us unable to continue to utilise previous work . Insofar

If it is not possible there remains a modified role for the ontological critique: the precise adaptation of the premises of any given adoptive model are, given their source, not already mapped onto the relevant problems and phenomena. If we want to use a novel model to explain a phenomenon which is outside its original remit, the implementation of the critique clarifies the mechanics of the model and the explanations offered by the model. Outlining these components allows us to make principled predictions. Finally, though it is expressed throughout this thesis 'long form' the critique is as amenable to breaking down the mechanics of mathematical models, or of the interpretations of statistical generalisations of those models[72].

Section 7.6: Objects and Properties in the Critique.

This section raises the first of two major theoretical challenge for the thesis for which only partial answers are presented. This problem arose during the application of the critique to CAPS. One of the components of a CAPS system is the connection between nodes. When articulating the ontological critique for CAPS this component highlighted a problem – is a connection between two nodes an object in itself or a property of the two objects it connects? This highlights the potential for slippage between the target of the 'objects' challenge and the 'properties' challenge in the ontological critique.

[72] For example, the interpretation of the decision points used by AlphaZero to play chess are not available from a mathematical examination of its algorithm, but they can be described using the ontological critique to offer explanations and make predictions about tendencies to move given ways in given positions. AlphaZero has a tendency to play h4!? in situations where more conventional engines evaluate this to be a poor move or even a mistake. Some commentators (Mittal, 2020) describe this as investing in space on the kingside in a way which risks one pawn but in turn increases the cost of short castling for black. This description can be formalised through the critique and, with further detail, be tested across the games in which it does, or does not, make this move.

In the chapter I decided, somewhat pragmatically, to opt for the former rather than the latter because of the ease it afforded in talking about how the network operated. Connections could then be afforded their own properties, rather than piling, potentially iterative, properties on nodes.

In this case, I take the solution I arrived at to be adequate to the task of the ontological critique for CAPS – offering a robust analysis of the existing model which is both critical and fruitful. However, I also take this problem case to be a member of a class of cases which will be ambiguous under the application of the critique. This presents two further questions: First, how should we handle such cases going forward, as a class? Second, does this represent a closing off of theoretical space within the critique – are there metaphysical assumptions built into the structure of the critique which will favour some theoretical structures or approaches over others?

At present the best response available to this problem is pragmatic. It is to apply the critique bearing this problem in mind, dealing with each problem case as it arises and handling these cases in whatever way ‘works’. It may be that doing so offers a pattern or principle of addressing problem cases, or indicates an avenue for improving the critique, but it may be that these cases remain problematic.

A given researcher or group of researchers that is applying the critique will possess aims internal to their project. These will sometimes offer pragmatic reasons to treat a given ambiguous case one way or another. If we are very fortunate, there will be a systematic preference for sorting these cases in the same direction. Perhaps ambiguous cases tend to be more readily treated as objects to facilitate clarity in describing the relevant interactions. If not, as I believe more likely, then each individual case may be sorted on its own merits, but this would emphasise the second, more concerning question.

If there are cases which do not systematically fall neatly as a target for the objects challenge or the properties challenge this means there are relevant cases within our utilised theoretical space which do not readily fit into the ontological critique′s structure. These cases, when sorted one way or another, are being given a particular ontological gloss by the application of the critique. We are not simply clarifying a pre-existing theoretical claim, we are making a novel one through applying the critique. This ontological gloss adds content to a model which is then party to the subsequent analysis. Given the role of the critique in facilitating judgements of the current state and future direction of theories this gloss could plausibly lead to a bias towards particular kinds of models which has no basis in them being better at offering explanations or predictions in line with the axiology challenge.

The ontological critique is designed to be, as far as possible, ontologically neutral in its approach to clarifying the construction of explanations and predictions. Given this aim, and given that any ontological gloss that biases the operation of the critique in this way is undesirable, we should regard the critique itself an unfinished tool. Insofar as it is possible to maintain the efficacy of the approach without the gloss, it should do so. Insofar as it is not possible to eliminate gloss and remain efficacious, its application should come with a warning – to be careful of biasing our theories in order to conform to the simplicity of the approach rather than to conform to the best available understanding. Given the structuring of the critique within an anglophone, analytic philosophy approach it is probable that this problem is more likely to occur for models constructed outside, or without utilising, WEIRD-centred patterns of thinking or theorising, and theorists should be especially wary of drawing strong conclusions about such models.

Given my background and current thinking, I am not in a position to make a relevant improvement to the critique which would mitigate these concerns. For this reason, the critique is presented as an unfinished tool. By this I mean that the approach is intended to be further improved through its application by a diverse community of practitioners who, by applying it to a greater variety of target theories, will gradually offer more nuanced and ontologically neutral approaches for capturing the relevant properties of the diverse range of theories in the available theory-space.

Section 7.7: The Lakatos Problem

The Lakatos problem is the problem that the framing of this thesis and the understanding it presents of science, and of progress for social psychology or for the models of attitude addressed, is construed within a Lakatosian framing which is widely challenged within the philosophy of science. This framework is the justification for the treatment of social psychology as a research programme, aiming at ever greater explanatory and predictive power, and which is evaluated through its historical movement towards the same.

There are two avenues of this problem which I will address here. First is my reading of Lakatos, and second is the challenge offered to Lakatos by Feyerabend. The bones of a response to both are sketched after the points are both outlined. Some key points of my position and their relation to some of what I have argued are clarified, indicating an avenue for future research with the aim of applying the reading of Lakatos I present here to resolving the broader theoretical challenge presented to Lakatos by Feyerabend.

My reading of Lakatos, here and throughout is more in line with the Lakatos-as-quiet-Hegelian (Hacking, 1983) than as the Lakatos-who-simply-employs-Hegelian-rhetoric. This is somewhat less popular in anglophone philosophy of science than the alternative (Musgrave and Pigden, 2021), and construes Lakatos as committing to a view of science whereby we cannot know which programmes are and are not *science* at a given time. We must instead evaluate the historical position of each research programme which will be offering ever greater explanations or predictions and so be progressive, or it will not and be degenerating. Where this becomes more problematic is in the stronger claims that the rationality of science inheres in the historical dialectic of its practice rather than in any given decision, discovery, or theory of its participants.

This second, stronger claim about the rationality of science is necessary if we are to vindicate the concept of the potential for individual researchers to persist in a degenerating research programme in the hope of its eventual improvement and vindication. The account I have offered of social psychology indicates that social psychology is just such a research programme. Furthermore, the recommendation I offer for social psychologists to adopt the critique and proceed piecemeal in improving their theory rather than importing something tried-and-tested from elsewhere is just such a recommendation – that it is rational, not just permissible or pragmatic, for social psychologists to attempt to turn the ship around rather than to change the research programme entirely.

One further note on this reading and application of Lakatos, as is apparent in the way I address the aptness of models to evidence and data within this thesis, I take it that Hacking (1983) goes too far to suppose that Lakatos rejects the correspondence theory of truth (Musgrave and Pigden, 2021). The rationality of science emerging from its historical position, and hence what is or is not science depending on where a given programme will ′end up′, is readily construed as an epistemic, rather than ontological claim. We cannot know if what we are studying in this programme, right now, is science or pseudoscience, but we can have good reasons for optimism or pessimism in this regard. This is consistent with the correspondence theory of truth, and gives the Hegelian approach a non-trivial role in understanding the operation of science.

The challenges presented to Lakatos′ picture of the history of scientific research programmes, and hence his account of science, come from Feyerabend (1975; Motterlini, 2000). He argues that when Lakatos is labelling a given research programme as progressive or degenerating, this presents a prescription to working researchers. One ought to work on progressive research programmes; one ought to abandon degenerating research programmes. One ought not to work on alchemy, one ought to work on high-energy physics. One ought not to work on evaluative priming, one ought to work on dual inheritance theory.

Feyerabend′s challenge takes the form of a dilemma − either this is indeed what Lakatos′ account entails for working researchers or, in Feyerabend′s words, ′anything goes′. That is, either Lakatos′ account entails a prescription to abandon degenerating research programmes and to join progressive programmes or Feyerabend is right about epistemic anarchism and the only rule which will capture everything that we want to call science and excludes everything we do not is that anything goes. Science is an anarchic enterprise, not a rule-governed nomic one. (Feyerabend, 1975, Motterlini, 2000)

Lakatos wants to reject the claim that the labels of progressive and degenerating are prescriptions, but preserve the force of the labels in order that science is, overall, rule-governed. He claimed both publicly and in correspondence (Motterlini, 2000) to have a response to Feyerabend′s challenge but died suddenly of a heart attack in 1974, at the age of 51, before delivering on this claim. The closes Lakatos comes to presenting a response to this challenge is found in an in-text note: ″One may rationally stick to a degenerating programme until it is overtaken by a rival and even after. What one must not do is to deny its poor public record. Both Feyerabend and Kuhn conflate methodological appraisal of a programme with firm heuristic advice about what to do. It is perfectly rational to play a risky game: what is irrational is to deceive oneself about the risk.″ (1970, p. 104) That is, for Lakatos, the normative labels of degenerating and progressive inform us about the ′risk′ of the game we are playing. This determines how we might go about playing that game. What the methodology of scientific research programmes provides an injunction against is treating a risky research programme as risk-free.

The relationship between these two problems is that for Lakatos to respond to Feyerabend′s challenges, as he claimed to be able to, Lakatos′ understanding of the enterprise of science seems committed to a more-than-rhetorical understanding of the claim that the ″owl of Minerva spreads its wings only with the coming of dusk″ (Hegel, 1991, p.24). That is, the process of scientific research does not leave us in a position to make judgements about it until long after the event. However, we are in a position to respond with urgency to our current crises and confidence to current growth. For the normative force of our judgements about the rationality of good science, this is sufficient, and justifies the avoidance of treating these judgements as heuristic advice about which programmes researchers should continue to pursue.

Quite how this reading of Lakatos should be squared and how doing so may leverage Lakatos′ position in responding to Feyerabend′s criticisms, requires further work to justify. The bones of that justification can however be briefly sketched.

The backbone is the idea that Lakatos is right that ′progressive′ and ′degenerating′ do not prescribe or proscribe the projects that scientists should participate in. This is rejecting the first horn of Feyerabend′s dilemma. The normative force of these labels is that they do prescribe what participants in a given research programme should be doing, as a member of that research programme. If a researcher has good reason to believe they are in a degenerating research programme, this is and should be cause for concern. While the rational response to this concern is underdetermined by the problem itself, we may sketch out some of the avenues.

First, more radical changes to the underlying hard-core commitments must be on the table. Participants should be willing to consider doing more than simply fiddling with a few peripheral auxiliary hypotheses in their work. Second, the scope of the problem of degeneration should be examined as part of that research programme. The causes of failures should become a part of the research programme′s target phenomena. Finally, the aim of this work is not simply to resolve minor theoretical concerns but to recognise the importance of improving the long-term robustness of the research programme. Not merely addressing the problem by bailing water, but also, perhaps thereby, trying to get back to port. For participants in progressive

This preserves the normative function of Lakatos′ labels to describe the way in which the operation of researchers in research programmes conforms to an overall normative framework towards the rational construction of a science. These injunctions are not inherently problematic to Lakatos because they are entailed by the constitutive aims of each research programme as a science. It is notable that this is not true if the programme is not a science, being labelled as non-science is not pejorative on Lakatos′ account.

These prescriptions, by preserving the injunctions on the behaviour of researchers within research programmes, also shows why not ′everything goes′. Those programmes whose internal norms do not conform to this dynamic, where problems for the research programme require that researchers try to change that course, are not science. Thus, we have an account of science where not everything goes, but it is also entirely rational to persist in a degenerating research programme. It is only irrational to recognise that one persists in a degenerating research programme and to not seek to change that trajectory.

This account leverages constitutive norms of research programmes in order to arrange the diagnose the labels as offering analogous guide to other projects. Just as there are no normative injunctions that one should play chess *per se*, there is no normative injunction that one should participate in the research programme of social psychology, or any other research programme. Yet there are injunctions to play for checkmate once one is playing chess, as there are injunctions for social psychologists to do something to improve the reliability of their field. There is no injunction to open with 1. e4 e5, or any other opening and defence in chess, but there is an injunction to deliver checkmate when one can, or to defend against it when one must. There is no injunction to adopt a given statistical or theoretical approach, but participants should try something that shows promise in moving the programme toward progress.

While several aspects of this account require further work, this limited version offers the avenue for that work theoretically and mirrors that in this thesis′ practical recommendations to researchers in social psychology. That is, the theoretical work of vindicating the interpretation of Lakatos and of offering a response to Feyerabend goes hand-in-hand with the practical work of actually engaging with the ongoing research programmes in their various historical positions and using this to inform our theoretical perspective.

Section 7.8: Summary

The thesis has argued that the proposed statistical responses to the replication crisis in social psychology are necessary but insufficient for the challenges presented by social psychology and proposed a supplementary 'ontological critique', a systematic series of challenges that clarify the claims a psychological model is entitled to make and hence the interpretations of data for which we have warrant. This tool was then demonstrated by applying it to some prominent models in attitude psychology with interesting and fruitful results. This conclusion drew together these findings, identifying areas for further research and highlighting the relationship that this thesis has to other projects in the field, highlighting the importance of systematically clarifying our generalisations, and the role of practical engagement with ongoing research programmes for theory in the philosophy of science. The thesis is a thesis in philosophy on psychology, which finds a collaborative path forward for both fields.

References:

- Aguiar, M.A.D., Dias, A., and Field, M., 2019. Feedforward Networks: Adaptation, Feedback, and Synchrony. *J Nonlinear Sci,* 29, pp. 1129 – 1164. https://doi.org/10.1007/s00332-018-9513-7
- Alfano, M., 2013. Identifying and Defending the Hard Core of Virtue Ethics. *Journal of Philosophical Research,* 38, pp. 233 – 260.
- Baranski, E., et al., 2020. Many Labs 5: Registered Replication of Schnabel and Nadler (2008), Study 4. *Advances in Methods and Practices in Psychological Science*, 3(3), pp. 405 – 417. https://doi.org/10.1177/2515245920917334
- Bar-Anan, Y., and Nosek, B., 2012. A Comparative Investigation of Seven Implicit Measures of Social Cognition. *Social Science Research Network.* http://dx.doi.org/10.2139/ssrn.2074556
- Bastos, A., et al., 2021. Dogs Mentally Represent Jealousy-Inducing Social Interactions. *Psychological Science, 32(5),* pp. 646 – 654. https://doi.org/10.1177/0956797620979149
- Bakker, M., et al., 2012. The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science,* 7(6), pp. 543 – 554.
- Beer, R. D., 1995. On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3, pp. 469-509.
- Bem, D. J., 2011. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), pp. 407–425. https://doi.org/10.1037/a0021524
- Bem, D. J., Utts, J., & Johnson, W. O., 2011. Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101(4), pp. 716–719. https://doi.org/10.1037/a0024777
- Benjamin, D.J., et al., 2018. Redefine statistical significance. *Nat Hum Behav.* 2, pp. 6–10. https://doi.org/10.1038/s41562-017-0189-z
- Bi, G. Q., Poo, M. M., 1998, Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength,

and postsynaptic cell type. *J Neurosci*, 18, pp. 10464 – 10472.

- Brannon, L., et al., 2007. The Moderating Role of Attitude Strength in Selective Exposure to Information. J*ournal of Experimental Social Psychology,* 43(4), pp. 611 – 617. https://doi.org/10.1016/j.jesp.2006.05.001.
- Bressan. P., 2019. Confounds in "Failed" Replications. *Frontiers in psychology*, 10:1884. https://doi.org/10.3389/fpsyg.2019.01884
- Button, K., et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci,* 14, pp. 365 – 376. https://doi.org/10.1038/nrn3475
- Carlin, B., and Louis, T., 2008. *Bayesian Methods for Data Analysis*. London: Chapman and Hall.
- Catalini, C., Lacetera, N., and Oettl, A., 2015. The Incidence and Role of Negative Citations in Science. *PNAS,* 112(45), pp. 13823 – 13826.
- Clarkson, J., et al, 2009. Does Attitude Certainty Beget Self-Certainty? *Journal of Experimental Social Psychology,* 45(2), p. 436 – 439.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145–153. https://doi.org/10.1037/h0045186
- Devezer, B., et al., 2020. The case for formal methodology in scientific reform. *Royal Society Open Science*. 10.1098/rsos.200805
- Doris, J., 1998. Persons, Situations, and Virtue Ethics. *Noûs*, 32(4), pp. 504-530.
- Doris, J., 2002. *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.
- Dreber, A., et al., 2015. Using Prediction Markets to Estimate the Reproducibility of Scientific Research. *PNAS,* 112(50), pp. 15343 – 15347.
- Duhem, P. M. M., 1991. *The Aim and Structure of Physical Theory*. Riviere, M [translator]. Princeton, NJ: Princeton University Press.

- Ebersole, C. R., et al., 2016. Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, pp. 68–82. https://doi.org/10.1016/j.jesp.2015.10.012
- Ebersole, C. R., et al., 2020. Many Labs 5: Testing pre-data collection peer review as an intervention to increase replicability. *PsyArXiv. 10.31234/osf.io/sxfm2*
- Efron, B., and Tibshirani, R., 1986. Bootstap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science,* 1(1), pp. 54 – 75. DOI: 10.1214/ss/1177013815
- Efron, B., and DiCiccio, T., 1996. Bootstrap Confidence Intervals. *Statistical Science,* 11(3), pp. 189 – 228. DOI: 10.1214/ss/1032280214
- Elman, J., 1990. Finding Structure in Time. *Cognitive Science,* 14(2), pp. 179 – 211. https://doi.org/10.1207/s15516709cog1402_1
- Elman, J., 1991. Distributed Representations, Simple Recurrent Networks, and Grammatical Structure. *Machine Learning,* 7, pp. 195 – 225.
- Engber, D., 2017. Daryl Bem Proved ESP is Real: Which Means Science Is Broken. *Slate.* [online] Accessible at: [https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html]
- Fabrigar, L. R., & Wegener, D. T., 2016. Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, 66, pp. 68–80. https://doi.org/10.1016/j.jesp.2015.07.009
- Fazio, R. H., 1990. Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in experimental social psychology,* 23, pp. 75-109.
- Fazio, R. H., & Olson, M. A., 2014. The MODE model: Attitude-behavior processes as a function of motivation and opportunity. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind*. New York: The Guilford Press. pp. 155–171.

- Fazio, R. H., and Olsen, M. A., 2003. Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, 54, pp. 297–327.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M.C., Kardes, F.R. 1986. On the automatic activation of attitudes. *J Pers Soc Psychol*., 50(2) pp. 229 – 38. doi: 10.1037//0022-3514.50.2.229.
- Feyerabend, P., 1975. *Against Method.* London: New Left Books.
- Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. 2012. Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, 103(6), pp. 933 – 948. https://doi.org/10.1037/a0029709
- Sritharan, R., & Gawronski, B. 2010. Changing implicit and explicit prejudice: Insights from the associative-propositional evaluation model. *Social Psychology*, 41(3), pp. 113 – 123. https://doi.org/10.1027/1864-9335/a000017
- Gendler, T., 2008a. Alief and Belief. *The Journal of Philosophy*, 105(10), pp. 634-663.
- Gendler, T., 2008b. Alief in action (and reaction). *Mind & Language*, 23(5), pp. 552–585. https://doi.org/10.1111/j.1468-0017.2008.00352.x
- Gendler, T., 2011. On the epistemic costs of implicit bias. *Philosophical Studies*, 156(1), pp. 33-63.
- Gilbert, E., 2015. Replication of "A Needs-Based Model of Reconciliation: Satisfying the Differential-Emotional Needs of Victim and Perpetrator as a Key to Promoting Reconciliation" (Shnabel & Nadler, JPSP, 2008). OSF Storage [online]. Available at: https://osf.io/fuj2c/
- Gilbert, D., et al., 2016. Comment on "Estimating the Reproducibility of Psychological Science". *Science,* 351(6277), pp. 1037. DOI: 10.1126/science.aad7243
- Glendinning, P., 1994. *Stability, Instability and Chaos*. Cambridge: Cambridge University Press.
- Greco, J., 1999. Agent Reliabilism. *Philosophical Perspectives,* 13, pp. 273 – 296.

- Greco, J., 2009. Knowledge and success from ability. *Philosophical studies,* 142(1), pp. 17 – 26.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K., 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* , 74(6), pp. 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R., 2009. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* , 97(1), pp. 17–41. https://doi.org/10.1037/a0015575
- Greenwald, A. G., Banaji, M. R., and Nosek, B., 2015. Statistically small effects of the implicit association test can have societally large effects. *J Pers Soc Psych,* 108(4), pp. 553 – 561.
- Gurney, K., 1997. *An Introduction to Neural Networks.* London: UCL Press.
- Hacking, I., 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.
- Haddock, G., and Maio, G., 2010. *The Psychology of Attitudes and Attitude Change*. London: Sage.
- Haddock, G., and Maio, G., 2012. *Psychology of Attitudes; Volumes 1-5* . London: Sage.
- Haddock, G., and Maio, G., 2015. *The Psychology of Attitudes and Attitude Change*. [2nd Edition] London: Sage.
- Hassel, S., & Ridout, N., 2018. An Investigation of First-Year Students' and Lecturers' Expectations of University Education. *Frontiers in Psychology*. 8. 10.3389/fpsyg.2017.02218.
- Harman, G., 1999. Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society*, 99, pp. 315-331.
- Harman, G., 2000. The Nonexistence of Character Traits. *Proceedings of the Aristotelian Society*, 100, pp. 223-226.

https://doi.org/10.1111/j.0066-7372.2003.00013.x

- Hazan, H., and Manvitz, L., 2012. Topological Constraints and Robustness in Liquid State Machines. *Expert Systems With Applications,* 39, pp. 1597 – 1606.
- Hebb, D., 1949. *The Organisation of Behaviour*. London: John Wiley and Sons.
- Hegel, G. W. F., 1991. *Elements of the Philosophy of Right*. Nisbet, H. [Translator]. Cambridge: Cambridge University Press.
- Hinton, G., 2010. Learning to Represent Visual Input. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences,* 365, pp. 177-184. 10.1098/rstb.2009.0200
- Hochreiter, S., and Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation,* 9(8), pp. 1735-1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735
- Holland, R., et al., 2002. On the Nature of Attitude-Behaviour Relations: The Strong Guide, The Weak Follow. *European Journal of Social Psychology,* 32(6), pp. 869 – 876.
- Holroyd, J., 2016. VIII—What Do We Want from a Model of Implicit Cognition?, *Proceedings of the Aristotelian Society*, 116(2), pp. 153–179. https://doi.org/10.1093/arisoc/aow005
- Holroyd, J., and Kelly, D., 2016. Implicit Bias, Character and Control. In: Masala, A. and Webber, J., (eds.) *From Personality to Virtue Essays on the Philosophy of Character*. Oxford: Oxford University Press. pp. 106-133.
- Holroyd, J., Scaife, R., and Stafford, T., 2017. What is Implicit Bias? *Philosophy Compass,* 12(10), e12437. https://doi.org/10.1111/phc3.12437
- Hopfield, J. J., 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *PNAS,* 79(8), pp. 2554 – 2558.

- Ioannidis, J., 2005. Why Most Published Research Findings Are False. *PLoS Med,* 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124
- Irvine, E., 2021. The Role of Replication Studies in Theory Building. *Perspectives on Psychological Science,* 16(4), pp. 844 – 853. doi: 10.1177/1745691620970558.
- Kelly, D., and Roedder, E., 2008. Racial Cognition and the Ethics of Implicit Bias. *Philosophy Compass,* 3(3), pp. 522 – 540.
- Jacowitz, K., and Kahneman, D., 1995. Measures of Anchoring in Estimation Tasks. *Personality and Social Psychology Bulletin,* 21(11), pp. 1161 – 1166. doi:10.1177/01461672952111004
- John, L., Loewenstein, G., and Prelec, D., 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), pp. 524-532. doi:10.1177/0956797611430953
- Jordan, M. I., 1997. Serial order: A parallel distributed processing approach. In J. W. Donahoe & V. P. Dorsel (Eds.), *Neural-network models of cognition: Biobehavioral foundations.* Amsterdam: North Holland. pp. 471–495.
- Jeffreys, H., 1961. *Theory of Probability*. [3rd Edition] Oxford: Clarendon Press.
- Kahneman, D., 2012. A proposal to deal with questions about priming effects. *Nature: Open Letter.* Available at: https://www.nature.com/news/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf
- Kerr N., 1998. HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), pp. 196-217. doi:10.1207/s15327957pspr0203_4
- Klein, R. A., et al., 2014. Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), pp. 142-152. http://dx.doi.org/10.1027/1864-9335/a000178

- Klein RA, et al., 2018. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4), pp. 443-490. doi:10.1177/2515245918810225
- Klein, R. A., et al., 2019. Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement. *PsyArXiv*. https://doi.org/10.31234/osf.io/vef2c
- Kriegel, U., 2012. Moral motivation, moral phenomenology, and the alief/belief distinction. *Australasian Journal of Philosophy,* 90, pp. 469–86.
- Kuhn, T., 1970. Reflections on my critics. In: Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Vol. 4. Cambridge: Cambridge University Press. pp. 231 – 278.
- Kuhn, T., 1977. *The Essential tension*. Chicago: University of Chicago Press.
- Lakatos, I., 1970. Falsification and the Methodology of Scientific Research Programmes. In Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Vol. 4. Cambridge: Cambridge University Press. pp. 91 - 196.
- Lakatos, I., 1971. The History of Science and its Rational Reconstructions. In: Buck, R. C., and Cohen, R. S., (eds.), *PSA 1970: Boston Studies in the Philosophy of Science*, 8. Dordrecht: Reidel, pp. 91–135.
- Lakatos, I., 1978a. Philosophical Papers Volume 1: *The Methodology of Scientific Research Programmes.* J. Worrall and G. Currie (eds.), Cambridge: Cambridge University Press.
- Lakatos, I., 1978b. Philosophical Papers Volume 2: *Mathematics, Science and Epistemology*. J. Worrall and G. Currie (eds.), Cambridge: Cambridge University Press.

- Lakens, D. and Caldwell, A., 2021. Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, 4(1), https://doi.org/10.1177/2515245920951503.
- Lakens, D., et al. 2017. Justify Your Alpha. *PsyArXiv.* https://doi.org/10.31234/osf.io/9s3y6
- Lakens, D., 2020a. *Distribution of Cohen's d, p-values, and power curves for an independent two-tailed t-test*. Application available at: https://lakens.shinyapps.io/p-curves/
- Lakens D., 2020b. *P-Curves code*. Accessible at: https://github.com/Lakens/p-curves
- Larvor, B., 1998. *Lakatos: An Introduction*. London: Routledge.
- Levy, N., 2015. Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Nous,* 49(4), pp. 800 – 823.
- Lynch, S., 2017. *Dynamical Systems with Applications Using Mathematica*. Basel: Birkhauser.
- Maass, W., and Markram, H., 2004. On the computational power of circuits of spiking neurons. *Journal of computer and system sciences,* 69(4), pp. 593 – 616.
- Machery, E., 2016. De-Freuding Implicit Attitudes. In: Brownstein, M., and Saul, J., 2016. *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology.* Oxford: Oxford University Press.
- Machery, E., 2017. What is a Replication? *Philosophy of Science,* 87(4), pp. 545–567. 0031-8248/2020/8704-0001
- Meehl, P. E., 1970. Some methodological reflections on the difficulties of psychoanalytic research. In Radner, M., Winokur, S., (Eds.) *Minnesota studies in the philosophy of science: IV. Analyses of theories and methods of physics and psychology*. Minneapolis, MN: University of Minnesota Press. pp. 403–416.
- Meehl, P. E., 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting*

and Clinical Psychology, 46, pp. 806–834.

- Meehl P. E., 1990. Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, 66(1), pp. 195-244.
- Miller, C., 2003. Social Psychology and Virtue Ethics. *The Journal of Ethics,* 7(4), pp. 365 – 392.
- Miller, C., 2016. Does the CAPS model Improve our understanding of personality and character? In: Masala, A. and Webber, J., (eds.) *From Personality to Virtue Essays on the Philosophy of Character* . Oxford: Oxford University Press. pp. 155 – 185.
- Mischel, W., & Shoda, Y. 1995. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), pp. 246–268. https://doi.org/10.1037/0033-295X.102.2.246
- Mischel, W., & Shoda, Y., 2006. Applying Meta-theory to Achieve Generalisability and Precision in Personality Science. *Applied Psychology*, 55(3) pp. 439 – 452.
- Mittal, A., 2020. AlphaZero at the Candidates 2020. [online] Available at: https://www.chess.com/blog/vinniethepooh/alphazero-at-the-candidates-2020
- Motterlini, M., 2000. *For and Against Method*. Chicago: University of Chicago Press.
- Munafo, M., et al, 2017. A Manifesto for Reproducible Science. *Nat Hum Behav.,* 1, 0021. https://doi.org/10.1038/s41562-016-0021
- Musgrave A., 1976. *Method or Madness?*. In: Cohen R.S., Feyerabend P.K., Wartofsky M.W. (eds) *Essays in Memory of Imre Lakatos. Boston Studies in the Philosophy of Science, vol 39* . Springer, Dordrecht. https://doi.org/10.1007/978-94-010-1451-9_27
- Musgrave, A., and Pigden, C., 2021. Imre Lakatos. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), Accessible at: <https://plato.stanford.edu/archives/sum2021/entries/lakatos/>

- Muthukrishna, M., and Henrich, J., 2019. A Problem in Theory. *Nat Hum Behav.,* 3, pp. 221 – 229.
- Myowa-Yamakoshi, M., Scola, C. & Hirata, S. Humans and chimpanzees attend differently to goal-directed actions. *Nat Commun,* 3, 693. https://doi.org/10.1038/ncomms1695Nosek et al 2002
- Nosek, B., et al., 2007. The Implicit Association Test at age 7: A Methodological and Conceptual Review. In: Bargh, J. A., (Ed.) *Automatic processes in social thinking and behavior.* New York: Psychology Press. pp. 265 – 292.
- Nosek, B., et al., 2021. Replicability, Robustness, and Reproducibility in Psychological Science. *PsyArXiv*. 10.31234/osf.io/ksfvq
- Nosek, B., and Gilbert, E., 2016. Mischaracterising Replication Studies Leads to Erroneous Conclusions. *PsyArXiv. 10.31234/osf.io/nt4d3*
- Open Science Collaboration, 2015. Estimating the Reproducibility of Psychological Science*. Science*, 349(6251).
- Oppenheimer, D., et al, 2009. Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology*, 45(4), pp. 867 – 872.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E., 2013. Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* , 105(2), pp. 171–192. https://doi.org/10.1037/a0032734
- Pashler, H., and Wagenmakers, E., 2012. Editors′ introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science,* 7(6), pp. 628 – 630.
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D., 2005. An inkblot for attitudes: Affect misattribution as implicit measurement. Journal of Personality and Social Psychology, 89(3), pp. 277–293. https://doi.org/10.1037/0022-3514.89.3.277
- Peck T. C., et al. 2013. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Conscious Cogn*., 22(3), pp. 779-87. doi:

10.1016/j.concog.2013.04.016.

- Petrocelli, J., et al, 2007. Unpacking Attitude Certainty: Attitude Clarity and Attitude Correctness. *Journal of Personality and Social Psychology,* 92(1), pp. 30-41.
- Popper, K., 1959. *The Logic of Scientific Discovery*, translation by the author of Logik der Forschung (1935), London: Hutchinson. Republished 2002, London & New York: Routledge Classics.
- Popper, K., 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge.
- Pritchard, D., 2004. Epistemic Luck. *Journal of Philosophical Research, 29, pp. 191 – 220.*
- Pritchard, D., 2012. Anti-Luck Virtue Epistemology. *The Journal of Philosophy,* 109(3), pp 247 – 279.
- Quine, W. V. O., 1951. Two Dogmas of Empiricism. *The Philosophical Review*, 60(1), pp. 20–43. doi:10.2307/2181906
- Railton, P., 2011. Two Cheers For Virtue: Or might virtue be habit forming? *Oxford Studies in Normative Ethics,* 1, pp. 295 – 330.
- Rees, C., 2016. A Virtue Ethics Response to Implicit Bias. In: Brownstein, M., and Saul, J., (eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics.* Oxford: Oxford University Press. pp. 191 – 214.
- Robert, C., 2001. *The Bayesian Choice*. New York: Springer.
- Rosenthal, R., 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), pp. 638–641. https://doi.org/10.1037/0033-2909.86.3.638
- Rosenthal, R., & Fode, K. L., 1963. The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, 8(3), pp. 183–189. https://doi.org/10.1002/bs.3830080302
- Rosenthal, R., & Rubin, D. B., 1978. Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1(3), pp. 377–415. https://doi.org/10.1017/S0140525X00075506

- Rydell, R. J., & Gawronski, B., 2009. I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion*, 23(6), pp. 1118–1152. https://doi.org/10.1080/02699930802355255
- Schaller M., Park J. H., Mueller A., 2003. Fear of the Dark: Interactive Effects of Beliefs about Danger and Ambient Darkness on Ethnic Stereotypes. *Personality and Social Psychology Bulletin*. 29(5), pp. 637-649. doi:10.1177/0146167203029005008
- Schimmack, U., 2012. The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17(4), pp. 551–566. https://doi.org/10.1037/a0029487
- Schimmack, U., 2021. The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science*, 16(2), pp. 396-414. doi:10.1177/1745691619863798
- Schimmack, U., 2020. A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie canadienne*, 61(4), pp. 364–376. https://doi.org/10.1037/cap0000246
- Schimmack, U., and Bartos, F., 2020. Z-Curve 2.0: Estimating Replication Rates and Discovery Rates. *PsyArXiv*. 10.31234/osf.io/urgtn
- Schimmack, U., and Brunner, J., 2018. Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. *Meta-Psychology,* (4), https://doi.org/10.15626/MP.2018.874
- Schindler S., Höhner A., Moeck R., Bruchmann M., and Straube T., 2021. Let′s Talk About Each Other: Neural Responses to Dissenting Personality Evaluations Based on Real Dyadic Interactions. *Psychological Science,* 32(7), pp. 1058 - 1072. doi:10.1177/0956797621995197
- Schmidt, S., 2009. Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. *Review of General Psychology*. 13. 90-100. 10.1037/a0015108.

- Schwarz, N., et al., 1985. Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly* , 49(3), pp. 388–395. https://doi.org/10.1086/268936
- Schwitzgebel, E., 2010. Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief, *Pacific Philosophical Quarterly*, 91: pp. 531–553.
- Scullin M. K., Gao C., Fillmore P., 2021. Bedtime Music, Involuntary Musical Imagery, and Sleep. *Psychological Science,* 32(7), pp. 985-997. doi:10.1177/0956797621989724
- Sherman, S., et al., 2003. Implicit and Explicit Attitudes Toward Cigarette Smoking: the Effects of Context and Motivation. *Journal of Social and Clinical Psychology,* 22. pp. 13-39. 10.1521/jscp.22.1.13.22766.
- Shnabel, N., & Nadler, A., 2008. A Needs-Based Model of Reconciliation: Satisfying the Differential Emotional Needs of Victim and Perpetrator as a Key to Promoting Reconciliation. *Journal of personality and social psychology,* 94, pp. 116-32. 10.1037/0022-3514.94.1.116.
- Shoda, Y., et al., 2012. Cognitive-Affective Processing System Analysis of Intra-Individual Dynamics in Collaborative Therapeutic Assessment: Translating Basic Theory and Research into Clinical Applications. *Journal of Personality*, 81(6), pp. 554 – 568.
- Shoda, Y., & Smith, R. E., 2004. Conceptualizing personality as a cognitive-affective processing system: A framework for models of maladaptive behavior patterns and change. *Behavior Therapy*, 35(1), pp. 147–165. https://doi.org/10.1016/S0005-7894(04)80009-1
- Shoda, Y., LeeTiernan, S., Mischel, W., 2002. Personality as a Dynamical System: Emergence of Stability and Distinctiveness from Intra and Interpersonal Interactions. *Personality and Social Psychology Review,* 6(4), pp. 316-325. doi:10.1207/S15327957PSPR0604_06
- Simmons, J. P., Nelson, L. D., Simonsohn, U., 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science,* 22(11), pp.

1359-1366. doi:10.1177/0956797611417632

- Simmonsohn, U., 2014. P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science, 9(6), pp. 666 – 681.*
- Snow, N. E., 2009. *Virtue as Social Intelligence: An Empirically Grounded Theory.* Abingdon: Routledge.
- Sosa, E., 1985. Knowledge and Intellectual Virtue. *The Monist,* 68(2), pp. 226 – 245.
- Sosa, E., 1993. Abilities, Concepts, and Externalism. In: Heil, J., and Mele, A., (eds), *Mental Causation*. Oxford: Oxford University Press.
- Sterling, T., 1959. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--Or Vice Versa. *Journal of the American Statistical Association*, 54(285), 30-34. doi:10.2307/2282137
- Stroebe, W., and Strack, F,. The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*. 9(1), pp. 59-71. doi:10.1177/1745691613514450
- Strogatz, S., 2014. *Nonlinear Dynamics and Chaos*. 2nd Edition. Boulder CO: Westview Press.
- Tanesini, A., forthcoming. Attitude Psychology and Virtue Epistemology: New Framework. In: Ballantyne, N., and Dunning, D., (eds) *Reason, Bias, and Inquiry: New Perspectives from the Crossroads of Epistemology and Psychology.* Oxford: Oxford University Press. Page numbers refer to the Feb. 2018 draft, available at: https://tanesini.wordpress.com/work-in-progress/
- Thelen, E., and Smith, L., 2001. Development as a dynamic system. *TRENDS in Cognitive Sciences,* 7(8), 343 – 348.
- Waddington, C. H., 1957. *The Strategy of the Genes*. London: George Allen & Unwin Ltd.
- Wagenmakers, E. J., 2007. A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review,* 14, pp. 779–804.

https://doi.org/10.3758/BF03194105

- Wagenmakers et al., 2011. Why psychologists must change the way they analyze their data: The case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology,* 100(3), pp. 426 – 432.
- Webber, J., 2015. Character, Attitude and Disposition. *European Journal of Philosophy,* 23(4), pp. 1082 – 1096. https://doi.org/10.1111/ejop.12028
- Webber, J., 2016. Instilling Virtue. In: Masala, A. and Webber, J., (eds.) *From Personality to Virtue Essays on the Philosophy of Character* . Oxford: Oxford University Press. pp. 134 – 154.
- West, R., 2018. Virtue Ethics is Empirically Adequate: A defense of the CAPS response to situationism. *Pacific Philosophical Quarterly,* 99(1), pp. 79 – 111.
- Wiggins, S., 2003. *Introduction to Applied Nonlinear Dynamical Systems and Chaos.* New York: Springer.
- Wilson, R., 1996. *Introduction to Graph Theory*. 4th Edition. Harlow: Longman Group.
- Wilson, T., Lindsey, S., and Schooler, T., 2000. A Model of Dual Attitudes. *Psychological Review,* 107, pp. 101 – 126.
- Wolfram Alpha LLC. 2021. Wolfram|Alpha: Lokta-Volterra Predator-Prey model. Accessible at: https://www.wolframalpha.com/input/?i=predator-prey+model&lk=3
- Yang, Y., et al., 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *PNAS*, 117(20), pp. 10762 – 10768.
- Zunick, P., et al, 2017. The Role of Valence Weighting in Impulse Control. *Journal of Experimental Social Psychology,* 72, doi: 10.1016/j.jesp.2016.11.014

Appendix:

Sample form: Ontological Critique

The axiology challenge: What do we care about?

In the research programme in general? *In this project in particular?*

An ideal answer identifies the dimensions of excellence or deficiency which are legitimate recourse for advocating or criticising the model.

The objects challenge: What are the proposed objects?

An ideal answer will identify all those entities to explanations of phenomena will refer. Central objects are those which are novel to the model or given a novel exposition. Peripheral objects are those which are not novel to the model, but which are necessary to offer the model's explanations.

The Properties challenge: What are the proposed properties of these objects?

An ideal answer will enumerate all those properties of the objects outlined in the previous challenge that are necessary for the explanations offered by the model.

The Explanation challenge: How does the conjunction of the objects and their *properties give rise to the observed evidence?*

An ideal answer clarifies what role the ontological commitments play in our explanations.

The Falsification challenge: What would it take to falsify these explanations?

What would have to occur for the explanation to be falsified?

What would be falsified if this occurred?

An ideal answer identifies some potential results that would falsify the model. Given such a result, it identifies the features of the model we take to be at stake in such a falsification.