SUBJECTIVE MORAL BIASES & FALLACIES: DEVELOPING SCIENTIFICALLY & PRACTICALLY ADEQUATE MORAL ANALOGUES OF COGNITIVE HEURISTICS & BIASES

Mark Herman

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2019

Committee:

Sara Worley, Advisor

Richard Anderson
Graduate Faculty Representative

Theodore Bach

Michael Bradie

Michael Weber

ABSTRACT

Sara Worley, Advisor

In this dissertation, I construct scientifically and practically adequate moral analogues of *cognitive heuristics and biases*. *Cognitive heuristics* are reasoning “shortcuts” that are efficient but flawed. Such flaws yield systematic judgment errors, *cognitive biases*. For example, the *availability heuristic* infers an event’s probability by seeing how easy it is to recall similar events. Since dramatic events like airplane crashes are disproportionately easy to recall, this heuristic explains systematic overestimations of their probability (*availability bias*). The research program on cognitive heuristics and biases (e.g., Daniel Kahneman’s work) has been scientifically successful and has yielded useful error-prevention techniques, *cognitive debiasing*. I attempt applying this framework to moral reasoning to yield *moral* heuristics and biases. For instance, a *moral bias* of unjustified differences in animal-species treatment might be partially explained by a *moral heuristic* that dubiously infers animals’ moral status from their aesthetic features.

While the basis for identifying judgments *as cognitive errors* is often unassailable (e.g., per violating laws of logic), identifying *moral errors* seemingly requires appealing to moral truth, which, I argue, is problematic within science. Such appeals can be avoided by repackaging moral theories as mere “standards-of-interest” (*a la* non-normative metrics of purportedly right-making features/properties). However, standards-of-interest do not provide authority, which is needed for effective debiasing. Nevertheless, since each person deems their own subjective morality authoritative, subjective morality (qua standard-of-interest and not *moral subjectivism*) satisfies both scientific and practical concerns. As such, (idealized) subjective morality grounds a moral analogue of cognitive biases, *subjective moral biases* (e.g., committed non-racists unconsciously discriminating).

I also argue that *cognitive heuristic* is defined by its contrast with rationality. Consequently, heuristics explain biases, which are also so defined. However, this property is causally-irrelevant to cognition. This frustrates *heuristic*'s presumed usefulness in causal explanation, wherein categories should be defined by causally-efficacious properties. As such, in the moral case, I jettison this role and tailor categories solely to *contrastive explanations*. As such, "moral heuristic" is replaced with *subjective moral fallacy*, which is defined by its contrast with subjective morality and explains subjective moral biases. The resultant *subjective moral biases and fallacies* framework can undergird future empirical research.

To my grandmother, mother, and father for all their unwavering support.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support and guidance of many people. First, I would like to thank my committee: Michael Weber, Michael Bradie, Theodore Bach, Richard Anderson, and especially, my advisor Sara Worley, who has been a great teacher, mentor, and friend over these many years. My thanks to Margy Deluca for all her tireless help. I'd also like to thank two special friends and philosophical interlocuters, Peter Jaworski and Scott Simmons. Most of all, I would like to thank my grandmother, mother, and father for all their love and unwavering support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS

CHB	Cognitive heuristics and biases
MHB	Moral heuristics and biases
MBF	Moral biases and fallacies
SMBF	Subjective moral biases and fallacies

SCIRA	Shortcut-vis-à-vis-the-ideally-rational-algorithm
MEIRA	More-economical-than-the-ideally-rational-algorithm
DEVIRA	Deviation-from-the-ideally-rational-algorithm

SCIMP	Shortcut-vis-à-vis-the-ideal-moral-reasoning-procedure
MEIMP	More-economical-than-the-ideal-moral-reasoning-procedure
DEVIMP	Deviation-from-the-ideal-moral-reasoning-procedure

CHAPTER 1: INTRODUCTION

1. Objective: Developing Moral Analogues of Cognitive Heuristics and Biases

Most people agree that racial discrimination is morally wrong. Despite this, there are people who racially discriminate. Why do they do this? To illuminate the project at hand, allow me to temporarily stipulate that racial discrimination *is morally wrong* (and set aside issues of moral truth, etc.—addressed in chapter 4). With this stipulation, we can restate the question as: Why do people perform this *immoral act*? We can then ask the more general question: Why do people perform immoral acts?

Here are some possible explanations: (1) People perform immoral acts because they have the wrong moral view. That is, they subscribe to a misguided morality that condones actions that in actuality, are immoral. For example, David practices racial discrimination because he believes that racial loyalty, triumph, and dominance are moral virtues and ends. The pertinent feature of this explanation is its appeal to (something like) false moral beliefs.[1]

Another possible explanation is: (2) People perform immoral acts because they do not care enough about acting morally. That is, the weight they assign to moral considerations is insufficient to preclude immoral action. For instance, while Lee thinks that racial discrimination is morally wrong, he does not much care about acting morally; as such, he practices racial discrimination when it is in his self-interest to do so. The pertinent feature of this explanation is its appeal to (something like) moral indifference or

[1] *False moral belief* illustrates the sort of thing that distinguishes this explanation from those that appeal to for instance, moral indifference, moral akrasia, false non-moral beliefs, or poor moral inference-making. Different metaethical views specify this sort of thing in different ways; this dissertation takes no position on what the right specification is. *False moral belief* is invoked merely for illustrative purposes. This dissertation is not committed to the presuppositions of *false moral beliefs* (e.g., that moral claims are truth-apt). Similar qualifiers apply to explanation (2).

morally problematic priorities.

An additional possible explanation is: (3) People perform immoral acts because they succumb to a temptation or aversion—that is, they are weak-willed or *akratic*. For example, a Jim Crow era restauranteur may oppose racial discrimination for moral reasons, but when the opportunity to not discriminate arises, he succumbs to social pressure and discriminates.

There is no single explanation of all immoral acts. Instead, there is a pool of explanations; each explanation in the pool pertains to a different set of cases. A potential explanation that could be added to the three already mentioned is: (4) People perform some immoral acts because they are flawed moral reasoners (i.e., flawed moral-inference makers). For instance, consider racial disparities in sentencing despite controlling for non-racial factors.[2] This constitutes racial discrimination and its constitutive acts are immoral—i.e., constitute moral errors. Yet regarding judges and jurors' commission of such moral errors, appeals to (1) wrong moral views (e.g., endorsed racism), (2) moral indifference, and/or (3) moral *akrasia* do not provide an incontrovertible complete explanation. Flawed moral reasoning is a plausible explanation that warrants investigation. Furthermore, especially for explananda such as this, it is plausible that there are multiple contributing factors (*a la* as expressed in analyses of variance). As such, to warrant investigation, flawed moral reasoning need not even be a plausible best or dominant explanation in any instance. It merely needs to be a plausible significant contributing factor in some cases-of-interest. In other words, the bar for flawed moral

[2] For an overview with copious citations, see American Civil Liberties Union (2014). For a contrary view, see MacDonald (2008). While this example serves narrative functions, ultimately, the arguments of this dissertation do not depend upon it. The plausibility of there being such examples is addressed in §3.

reasoning being explanatorily useful and thus, worth investigating, is relatively low.

So, how can we pursue this potential explanation? Perhaps we can exploit a useful tool for explaining non-moral errors, namely, *cognitive heuristics and biases* (CHB)—or more precisely, the paradigm utilized in the CHB cognitive psychological research program spearheaded by Daniel Kahneman and Amos Tversky (e.g., Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1996; 2000; Tversky & Kahneman, 2003).[3] *Cognitive heuristics* are reasoning shortcuts that are efficient but flawed. Such flaws yield systematic errors—i.e., *cognitive biases*. For example (Tversky & Kahneman, 1982a), (one species of) the *availability heuristic* infers the probability of an event not by directly measuring or assessing probability, but by "taking the shortcut" of assessing the proxy: the ease with which similar types of events can be recalled. As such, this heuristic yields errors when ease-of-recall and probability diverge. Since dramatic events (e.g., airplane crashes) are disproportionately easy to recall, our use of this heuristic yields systematic overestimations of the probability of dramatic events—i.e., yields (a species of) the *availability bias*.[4] In addition to discovering, unifying, and explaining errors, the CHB research program also spurred techniques for preventing errors—i.e., *cognitive debiasing*.

Given CHB's usefulness in explaining non-moral errors, an intriguing proposal is adapting the CHB paradigm to explain some moral errors—i.e., some immoral acts, decisions, and/or judgments.[5] Such adapting would yield moral analogues of cognitive heuristics and biases—i.e., something like *moral heuristics and biases* (ultimately, in

[3] In this dissertation, instances arise when it is not obvious whether to use the term, "paradigm," "conceptual framework," or "model." I will usually use the broadest term, "paradigm," as the risks of excluding important elements (e.g., associated explanatory strategies) outweigh the costs of unnecessary inclusion. Generally, the use of "paradigm" should not be interpreted as having significant implications.

[4] A primer on cognitive heuristics and biases is provided in chapter 2, §2.

[5] "Acts, decisions, and/or judgments" and forthcoming, "acts/decision/judgments," are stand-ins for the more precise, but less accessible, "cognitive outputs." This is addressed in chapter 2, §3.

chapters 3-5, ‘moral heuristic’ and ‘moral bias’ will be replaced with/specified as ‘subjective moral fallacy’ and ‘subjective moral bias’). Such analogues hold scientific and practical promise. They offer (1) to illuminate various immoral acts/decisions/judgments by identifying and unifying them as instantiations of systematic moral biases, (2) explain many such biases with moral heuristics (or moral fallacies), and (3) yield techniques for reducing the incidence of immoral acts/decisions/judgments—i.e., yield *moral debiasing* techniques (elaborations and additional benefits in §4). Given this potential usefulness, the task of this dissertation is to develop moral analogues of cognitive heuristics and biases. This mostly involves developing meanings for the concepts, ‘moral heuristic’ and ‘moral bias’ (later replaced with/specified as ‘subjective moral fallacy’ and ‘subjective moral bias’). It also involves addressing issues that stem from those meanings, such as implications for the accompanying explanatory strategy. In general, the task of this dissertation is to make progress on the philosophical foundations of a moral psychological paradigm that features these analogues and can undergird fruitful empirical research.[6]

[6] It should be noted that the terms, “moral heuristics and biases,” have been employed elsewhere, including in empirical research (e.g., Lindstrom, Jangard, Selbing & Olsson, 2018; Petersen, 2009; Wilkinson-Ryan & Baron, 2009). In this respect, the notion of moral analogues of cognitive heuristics and biases is not novel. This dissertation is distinguished from such uses by the combination of the following. (1) This dissertation provides a philosophically rigorous foundation for the concepts that yields precise and justified definitions. Many uses simply state a plausible definition without justification and do not wrestle with the issues confronted in chapters 3 and 4. This is not meant pejoratively, as per the division of scientific labor, psychologists should not be expected to provide foundational philosophical defenses of new concepts—especially since withholding new concepts pending such defenses would be counterproductive. (2) The basis of the moral analogues includes ‘cognitive heuristic’ per the classical notion of the Daniel Kahneman and Amos Tversky school (Ch. 2, §2.3). This is distinct from Kahneman and Shane Frederick’s subsequent notion of heuristics per attribute substitution (Kahneman & Frederick, 2002, p. 53-60). While at bottom, the issues raised in this dissertation bear upon both the classical and attribute-substitution versions, the classical notion is more useful for the purposes of adaption—explicated in Ch. 3, n. 33. Uses of “moral heuristic” coming out of the attribute-substitution school include Bruers (2013) and Dubljević & Racine (2014). Another distinct conception of ‘cognitive heuristic’ is per the Gerd Gigerenzer school (e.g., Gigerenzer, Todd, & the ABC Research Group, 1999). Gigerenzer (2010; Fleischhut & Gigerenzer, 2013) rejects “moral heuristics” *per se*, as he interprets the concept as entailing moral-domain-specificity; nevertheless, there are uses of “moral

A sensible method of developing moral analogues of cognitive heuristics and bias by adapting the CHB paradigm to explaining moral errors begins with taking those CHB concepts and regarding their meanings' cognitive-domain-regarding (or domain-general-regarding) features, replacing them with moral-domain-regarding-features to yield *prima facie* MHB concepts. Given that the central meaning of 'cognitive heuristic' is reasoning shortcut, such replacement yields a *prima facie* concept of '*moral* heuristic' with the meaning: *moral* reasoning shortcut; given that the central meaning of 'cognitive bias' is systematic error, such replacement yields a *prima facie* concept of '*moral* bias' with the meaning: systematic *moral* error.[7] To complete the adaption, these *prima facie* concepts

heuristic" that appeal to this school (e.g., Cosmides & Tooby, 2006; Petersen, 2009). (3) (If I may be so bold) the foundation I provide yields a paradigm that is scientifically and practically adequate (though perhaps merely per one conception of such). Cass Sunstein (2005) provides a groundbreaking treatment of 'moral heuristic,' but one that is highly problematic (Ch. 4, §4.2; Ch. 4, n. 27). It is worth noting that uses of "moral heuristic" that lean upon Sunstein's paradigm may be vulnerable to entanglement in the paradigm's problems (e.g., Fischer, 2016; Sheehan & Lee, 2014). An exciting new approach -though one that suffers from some of the adequacy issues addressed in chapter 4- is provided by Hanno Sauer (2018) (while it will be a topic of future research, it is beyond the scope of this dissertation). As far as I can find, there is not a treatment of moral analogues of cognitive heuristics and biases that also conforms to 1-3. That said, I ultimately replace 'moral heuristic' with 'subjective moral fallacy' (Ch. 3; Ch. 5) and replace 'moral bias' with 'subjective moral bias' (Ch. 4); this will also distinguish the material provided in this dissertation. There are various relations that the philosophical foundations provided in this dissertation may hold with past and future uses of "moral heuristic and bias" and other moral analogues of cognitive heuristics and biases. These include (1) entailing adjustments in such uses (e.g., adjusting extensions), (2) offering philosophical buttressing that can be seamlessly slid under such uses, and (3) being irreconcilable with such uses and thus, yielding either competing conceptions or orthogonal concepts. An orthogonal concept result would support using distinct terms. Such relations will likely vary between different uses.

[7] The method of adapting described above involves developing the moral analogues by developing their concepts' meanings—or more precisely, intensions. In this respect, such adaption takes an internalist-oriented approach to the semantics of 'moral heuristic/fallacy' and 'moral bias.' Some may object to this. In its defense, the approach stems from 'cognitive heuristic' and 'cognitive bias' depending upon descriptivist means of reference. That is, 'cognitive heuristic' and 'cognitive bias' *do not* -or at least, *should not*- directly refer. In other words, employing traditional Kripke-Putnam-style causal reference and externalist natural-kind semantics would be inapt (Kripke, 1980; Putnam, 1975); more precisely, reflecting my revisionary posture, such reference and semantics would not be the best option or "policy" in this case. This stems from the kinds, *cognitive heuristic* and *cognitive bias*, not constituting either (traditional) natural kinds (*a la* kinds that "carve nature at its joints") or even kinds that carve *specific systems* at their causal joints—i.e., *system-specific objective kinds* (SOK) (Ch. 3, §3). This reflects cognitive heuristics and biases' implicit, though essential contrast with ideal-rationality (Ch. 3, §1, §5), and this essential property's -or central property's- causal-inefficaciousness (Ch. 3, §2, §5). (At least *prima facie*) not constituting a (traditional) natural kind (or even an SOK) frustrates causal reference, which frustrates externalist semantics, which supports internalist semantics for 'cognitive

will be subjected to scrutiny and where warranted, revised. This is done in chapters 3 through 5.[8]

heuristic' and 'cognitive bias,' which supports an internalist-oriented approach to developing 'moral heuristic' and 'moral bias.' While my approach is internalist-*oriented*, the intensions are open to significant revision; if this presupposes some utilization of externalist semantics, I am fine with that, as I am open to hybrid approaches. Considering alternatives to my approach, perhaps a full-throated externalist semantics can be grounded in a promiscuous-explanatory view of natural kinds that can apply to *cognitive heuristic* and/or *cognitive bias* (my thanks to Theodore Bach for this suggestion). However, I suspect that the descriptive content that I identify with these concepts' intensions would prove indispensable—i.e., even if externalist-oriented views could forego intensions *per se*, I suspect that the descriptive content would still have to play a vital role (e.g., in reference *a la* Stanford & Kitcher, 2000). In other words, I doubt a *full-throated* externalist semantics is workable for *cognitive heuristic* and *cognitive bias*, and some "hybridization" would be necessary. If this is right, adherents of externalist-oriented views should find my focus on descriptive content worthwhile, even if ultimately, incomplete. Otherwise engaging externalist views/options for 'cognitive heuristic' and 'cognitive bias' is beyond the scope of this dissertation. Nonetheless, perhaps there is a view that (a) renders *cognitive heuristic* and *cognitive bias* natural kinds, (b) achieves reference without any appeal to descriptive content, (c) overcomes canonical challenges (e.g., the *qua problem*—e.g., Mallon, 2017, §3), (d) yields a full-throated externalist semantics of 'cognitive heuristic' and 'cognitive bias,' (e) befits cognitive heuristics and biases more than alternative views, and (f) renders an intension-inclusive approach to 'cognitive heuristic' and 'cognitive bias' neither useful, nor interesting. If so, a rejection of such a view (e.g., per rejecting promiscuous-explanatory views of natural kinds) -or more specifically, a rejection of the view *being employed for* 'cognitive heuristic' and 'cognitive bias'- may be an assumed premise of this dissertation and a subject for future research.

[8] The concepts this adaption ultimately yields (see chapter 5) may be opaque at this point. 'Moral heuristic' will be replaced with 'moral fallacy' (Ch. 3). This is due to problems with cognitive heuristic's central property, shortcut-ness, or more fully expressed: shortcut vis-à-vis the ideally rational algorithm (SCIRA). Those problems especially stem from SCIRA's causal-inefficaciousness within the human cognitive system (Ch. 3, §2). That stems from SCIRA's constituents' explanatory cross-purposes, including serving non-causal explanations. Those constituents are: deviation from the ideally rational algorithm (DEVIRA), more economical than the ideally rational algorithm, and natural assessment process (revised as intuitive process). Ultimately, only DEVIRA (contrastively) explaining cognitive biases is scientifically useful and survives the adaption of 'cognitive heuristic' to the moral domain. This yields a moral analogue with the central property, deviation from the ideal *moral reasoning procedure* ('DEVIMP'), which befits the term, "moral fallacy" (Ch. 3, §7). Per DEVIRA and DEVIMP's contrastive (vs. narrow causal) explanatory function and *moral fallacy* constituting a practical kind, there is no need to unify *moral fallacy* through identification with a distinct cognitive mechanism—a mistake made with cognitive heuristic. To achieve scientific and practical adequacy, the aforementioned ideal-moral-reasoning-procedure should be determined by ideal-instrumental-moral-rationality (IIMR)—i.e., subjective morality (Ch. 4) (I will often use hyphens in multiple-word term to facilitate readability). IIMR/subjective-morality evaluates one's actions, etc. (i.e., means) in terms of their contribution to the ideal (vs. pragmatic) satisfaction of one's genuine moral ends (Ch. 4, §2.2). One's moral ends -or more generally, morality- is *genuine* by conforming with counterfactual idealization—more specifically, two-tiered idealization (i.e., an adaption of two-tiered internalism—Rosati, 1996), in which the agent is idealized per those conditions/endowments that she would consider authoritative under ordinary optimal conditions. Actions/decisions/judgments that deviate from that dictated by genuine IIMR/subjective-morality constitute *subjective moral errors* and when systematic, instantiate *subjective moral biases* (Ch. 4, §3). Since a moral-reasoning-procedure is ideal per only yielding morally correct (i.e., not morally erroneous) acts/decisions/judgments, and since the moral errors in-question are subjective-moral-errors, this yields *moral fallacy* as *subjective moral fallacy*.

As will be seen, various scientific, metaethical (loosely speaking), and practical hurdles and objections arise. For instance, in the opening of this chapter, I temporarily stipulated that racial discrimination is morally wrong and set aside issues regarding moral truth, etc. The use of this temporary stipulation reflects the challenge of grounding moral error within the context of a scientific research program (Ch. 4). Scrutinizing the *prima facie* MHB concepts, addressing challenges and objections that arise, and responding with revisions (both to the concepts themselves and to pertinent aspects of the accompanying paradigm) will comprise the bulk of this dissertation.

2. Potential Examples

Definitive examples are not available at this time, as there is not empirical inquiry regarding criteria that will be introduced later. Nevertheless, the following potential examples provide a general idea of what moral analogues of cognitive heuristics and biases might look like (more refined examples are provided in chapter 5). In other words, the general idea communicated below will be sufficient to get the inquiry off-the-ground, and I will be able to get more precise later.[9]

First, let's consider an exemplar instantiation of the *prima facie* concept of 'moral heuristic' (though this exemplar may not actually exist). A straightforward and if you will, "conceptually clean" hypothesized example is the *cuteness heuristic*: a cognitive process that infers an entity's moral status by "taking the shortcut" of assessing the proxy, cuteness (e.g., per neotenous characteristics or baby schema—Jia, Park & Pol,

[9] Regarding empirical inquiry, as far as I am aware, there has not been empirical inquiry regarding the later-introduced criteria—namely, criteria stemming from the standard, genuine ideal-instrumental-moral-rationality (IIMR) (see Ch. 4). The following potential examples are compatible with the *prima facie* concepts of 'moral heuristic' and 'moral bias.' With respect to the analogues that will ultimately emerge, while these potential examples do not address certain complexities, they still provide a general idea of what those analogues might look like.

2015, p. 169). Such a process would (be conducive to) yielding moral errors when moral status and cuteness diverge. For instance, consider Roomba vacuum cleaners, which are little, reminiscent of a face, and scoot around the floor, self-propelled. An extreme potential instantiation of errors due to the cuteness-heuristic is people taking their Roombas with them on vacation to avoid leaving the appliance home "alone" (Associated Press, 2007). Potential errors that could be more widespread are inconsistencies in the treatment of non-human animals that are attributable to cuteness, or a lack thereof (e.g., dogs vis-à-vis pigs, or squirrels vis-à-vis rats).[10] This heuristic would almost certainly yield some systematic moral errors—i.e., a moral bias.[11]

A further sense of what the moral analogies might involve can be gleaned from the identifiable victim effect. An exemplar of this effect is aid to victims more strongly correlating with *identifiability* than more plausibly justifiable factors, such as need (Small & Loewenstein, 2003). A victim is identifiable insofar as there is a definite referent (e.g., a particular person that is the victim). An exemplar of an identifiable victim is "Baby Jessica" McClure, who in 1987, at 18 months old, was trapped in a well; her plight and eventual rescue was a media sensation. In contrast to identifiable victims, statistical

[10] While this heuristic lacks empirical confirmation, nonetheless, it is plausible in light of: (a) cuteness perception is induced by baby schema instantiated in both humans (Kringelbach, Stark, Alexander, Bornstein, & Stein, 2016), other animals (Borgi, Cogliati-Dezza, Brelsford, Meints, & Cirulli, 2014) and even in non-living objects (Cho, Gonzalez & Yoon, 2011), such as Volkswagen Beetles (Jia, Park & Pol, 2015), and (b) cuteness perception motivates caretaking (Glocker, Langleben, Ruparel, Loughead, Gur, & Sachser, 2009). Especially with respect to cute, non-living objects, it is worth noting that while a process' being frequently overridden may frustrate its discovery and reduce its yielding of errors, that does not bear upon the process' existence. Furthermore, overridden processes are still detectable. For instance, if people more quickly toss Legos and toy helicopters into a trash compactor than they toss stuffed animals, such a delay would be some evidence of an overridden cuteness heuristic. Cuteness certainly plays some role in the moral valuation of non-human animals (Loughnan & Piazza, 2018; Piazza, McLatchie, Olesen & Tomaszczuk, 2016; Sherman, Haidt & Coan, 2009). While the plausibility of this heuristic renders it a more compelling potential example, the example's primary function is providing an exemplar instantiation of the concept, independent of its actuality.

[11] Such likely yielding of systematic errors is explicated in §3.

victims have indefinite referents. An exemplar statistical victim is *an* impoverished child that is a member of a group identified by sterile, bureaucratic language.

An example of the identifiable victim effect is that people donate more money to identifiable victims than statistical victims, even when donating to statistical victims would be more efficacious. The strength of this effect is reflected in its persistence despite seemingly inconsequential identifiability. For instance, Small and Lowenstein show that the effect persists even when the victim is only identifiable via a randomly assigned ID number that is drawn at random before a decision to donate is made, as contrasted with a statistical victim, whose ID number would be randomly drawn immediately after the decision. For example, a hat containing randomly assigned ID numbers is presented to experiment participants. The only difference in experimental conditions is whether: (a) the experimenter draws and reads a number aloud before the participant makes the decision (thus providing a definite referent and identifiable victim), or (b) the experimenter announces that a number will be drawn after the participant decides (thus rendering the referent indefinite, and the victim statistical at the time of the decision).[12]

Surely, the time at which the number is drawn is not morally relevant; surely, no plausible moral theory would say otherwise. As such—at least for the illuminative purposes of this introduction—experiment participants' decisions to make larger donations in the former case and smaller donations in the latter case potentially instantiates a moral error. The systematicity of such decisions renders them a potential instantiation of a moral bias. Such systematicity also suggests the existence of a distinct

[12] Confidence in this finding may require replication, though we can set that aside for our current illustrative purposes.

psychological process that explains the bias.[13] That process could involve assessing (a trait that roughly correlates with) identifiability and "taking the shortcut" of using that assessment as a proxy for more plausibly justifiable assessments for determining aid. Such a process would potentially constitute a moral heuristic.[14]

There are additional examples of established phenomena that can provide a sense of the sorts of things that the moral analogues might involve.[15] These phenomena include: implicit racial bias (Brownstein, 2015), blaming the victim (Lerner & Simmons, 1966), prejudicial violence involving displaced aggression (Hovland & Sears, 1940), and the role of psychic numbing in tolerating genocide (i.e., indifference to the plight of individuals who are one-of-many in a much greater problem) (Slovic, 2007, p. 79).[16]

3. Plausibility

Adapting cognitive heuristics and biases to yield scientifically and practically useful moral analogues is an intriguing proposal. However, do we have reason to think such analogues exist?[17]

Numerous considerations bear upon this. Some are philosophical considerations

[13] This dynamic of systematic judgments suggesting a distinct process is explicated at the end of Ch.2, §1.

[14] The above will suffice for a potential moral error, bias, and heuristic for now, as the criteria ultimately adopted (n. 8) require extensive exposition (Ch. 3-5). More refined examples are provided in chapter 5.

[15] The following phenomena -as standardly conceived- often do not neatly fit into categories of a process (e.g., a heuristic) and generated judgments (e.g., a bias). In this respect, the phenomena point to promising places to look for such analogues, but conceptions of the phenomena may need to be tweaked to fit the analogues' paradigm.

[16] This notion of "psychic numbing" differs from similar notions of the same name, such as that in which decreased affect is not accompanied by a decreased motivation to act (e.g., Lifton, 1967).

[17] (1) Considering the aforementioned employment of "moral heuristic and bias" in empirical research (e.g., Lindstrom, Jangard, Selbing & Olsson, 2018; Petersen, 2009; Wilkinson-Ryan & Baron, 2009), establishing such existence may seem unnecessary. However, the relationship between these employments and the moral analogues I will be developing is unestablished—for example, they may be compatible conceptions, competing conceptions, or conceptually orthogonal (n. 6). Furthermore, I do not necessarily think those employments are correct, useful, etc. As such, I will build the case for the analogues I am developing independently of such employments. (2) The arguments that follow are compatible with both the *prima facie* moral analogues mentioned, and the moral analogues that will ultimately emerge (given plausible assumptions) (see n. 8).

(if you will). These include metaethical worries, loosely speaking (e.g., can *moral error* be grounded within the context of a scientific research program? Ch. 4), and conceptual considerations (e.g., revisions to 'cognitive heuristic'—Ch. 3). Such considerations will be addressed later and are bracketed from the following. A separate consideration is whether there exists the right sorts of processes and actions/decisions/judgments for constituting the moral analogues (e.g., moral heuristics and biases). While this is ultimately an empirical question, we can still assess the plausibility of their existence (i.e., whether that plausibility is sufficient to proceed). This is addressed in the following. In other words, in the following, I will just establish that the right sort of processes, etc. plausibly exist, and get into the deeper philosophical issues later (which would be rendered moot if those processes, etc. were not plausible).

3.1. Per Capacities & Processes

Consider human cognitive capacities concerning epistemic and practical matters (e.g., perception, belief-formation, and decision-making). These capacities are not infallible, flawless, and perfect; they are *non-ideal*. For instance, sometimes humans fall prey to optical illusions and misjudge their environment. Sometimes they form unjustified beliefs due to invalid inference-making (e.g., affirming the consequent), inapt reasoning (e.g., *ad hominem* arguments), or motivated irrationality (e.g., wishful thinking). Sometimes humans perform imprudent actions due to *akrasia* (i.e., weakness of will). Sometimes they act irrationally due to inconsistent valuation (e.g., sensitivity to framing effects) or incoherent desires (e.g., intransitive preferences).

Since human cognition is prone to non-ideal capacities when epistemic and practical matters are concerned, it certainly seems plausible that it would be similarly prone to non-ideal capacities when moral matters were concerned. In other words, given that this kluge-ridden, mush of wetware sloshing around in our skulls exhibits such non-ideality regarding epistemic and practical matters, it certainly seems plausible that it would exhibit similar non-ideality regarding moral matters. Finding that the mind's propensity to non-ideality persisted concerning moral matters would be less surprising than finding that this propensity suddenly disappeared when moral matters were raised. As such, it is plausible that non-ideal capacities concerning moral matters exist.

This establishes the plausibility that cognition concerning moral matters goes awry. Let's now consider the plausibility of it going awry in the particular ways that would yield moral analogues of cognitive heuristics and biases. This requires shifting from capacities to processes. A pertinent way in which such processes can go awry involves *imperfect proxies*. For an illustration of this, recall the aforementioned availability heuristic. It involves answering a question about the probability of an event by assessing the ease with which similar events can be recalled. In this respect, ease-of-recall constitutes a proxy for probability. Sometimes—e.g., when events are dramatic—ease-of-recall assessments deviate from assessments of probability. In other words, ease-of-recall does not perfectly track probability. In this respect, ease-of-recall is an imperfect proxy. In cases in which such assessments deviate, the proxy is *inapt.* When the proxy is inapt, its use yields a response—e.g., a judgment—that deviates from the question's answer. In this respect, such responses are *inapt*. In sum, the availability heuristic uses an imperfect proxy (i.e., ease-of-recall), which is susceptible (e.g., when events are dramatic) to being inapt and yielding *inapt judgments*.

As is typically the case with processes featuring imperfect proxies, the availability heuristic yields some systematic inapt judgments. This stems from the factor, whether-an-event-in-question-is-dramatic, bearing a different relationship to ease-of-recall than it bears to probability—namely, dramatic events' being easy to recall but not correspondingly probable. In other words, the curve for the pair of variables, event-drama and ease-of-recall, has a different shape from the curve for event-drama and probability. If there is at least one psychologically useful factor or condition (e.g., event-drama) that has a different relation/curve with the proxy (e.g., ease-of-recall) than it has with the target (e.g., probability), then the proxy will be inapt and yield inapt judgments (e.g., probability overestimations) when the factor is present, and thus, the process (e.g., the availability heuristic) will yield some (relevantly) systematic inapt judgments—that is, not all of the inapt judgments will be (relevantly) random.[18] When a process uses an imperfect proxy, there is usually at least one such factor/condition.

Our epistemic and practical capacities frequently rely upon imperfect proxies. Examples of such imperfect proxies include: the use of representativeness as an imperfect proxy for probability (Tversky & Kahneman, 1982b), and the use of negative affect as an imperfect proxy for riskiness (Slovic, Finucane, Peters & MacGregor, 2002). Given that human cognition is prone to relying upon imperfect proxies when epistemic and practical matters are concerned, it certainly seems plausible that it would also be similarly prone to relying upon imperfect proxies when moral matters were concerned. Recall that when a process uses an imperfect proxy, there is usually at least one psychologically useful factor

[18] Such "(relevant) systematicity" and "(relevant) randomness" are only with respect to useful factors/conditions/categories in psychology—i.e., the sorts of things relevant to psychological descriptions and topics. In other words, such systematicity/randomness concerns the presence/absence of patternly-ness between the sorts of things that might be captured as values or variables in a psychologist's regression equation. It is worth noting that as such, Laplacian-style, low-level determinism neither precludes (relevant) "randomness," nor is sufficient for (relevant) "systematicity" in the pertinent senses. This is elaborated further in chapter 2 (§2.1, n. 21).

that bears a different relation with the proxy than the target, which yields systematic inapt judgments/decisions/actions. As such, it is plausible that amongst the processes concerning moral matters there are some that utilize imperfect proxies and yield systematic inapt judgments/decisions/actions.

The preceding inductions established the plausible existence of *non-ideal* cognitive capacities concerning moral matters, processes concerning moral matters that feature *imperfect* proxies, and systematic *inapt* judgments/decisions/actions concerning moral matters. These inductions were undertaken using terms such as “non-ideal” (capacities), “imperfect” (processes and proxies), and “inapt” (proxies and judgments/decisions/actions). These terms serve as alternatives to “irrational,” “erroneous” and other loaded terms that are standardly invoked in explications of heuristics and biases. Such terms are loaded in a sense that involves the aforementioned bracketed philosophical considerations (e.g., the philosophical issues involved in attributing moral erroneousness to judgments/decisions/actions within the context of a scientific research program). The difference between the (a) non-loaded versus (b) loaded terms represents the difference between (a) establishing the existence of processes and judgments/decisions/actions that are of the right sort to constitute moral analogues (upon resolution of the philosophical considerations) versus (b) establishing the plausible existence of moral analogues that satisfy the philosophical considerations. In this respect, the established plausible existence of (1) processes concerning moral matters that feature imperfect proxies, and (2) systematic inapt judgments/decisions/actions concerning moral matters is sufficient for (a) establishing the plausible existence of those processes and actions/decisions/judgments that are of the right sort to constitute moral analogues of cognitive heuristics and biases upon resolution of the philosophical considerations.

3.2. Per Paradigm Applications

An additional line of inductive support comes from the continuous expansion in the range of processes and outputs to which (adaptions of) the heuristics and biases paradigm have been successfully applied.

Firstly, Kahneman and Tversky's development of the cognitive heuristics and biases paradigm was itself inspired by the application of a similar paradigm to perceptual judgments. Such application is exemplified by Egon Brunswik's (1943) use of the lens model to capture the haze illusion (Kahneman & Frederick, 2002, p. 52; Tversky & Kahneman, 2003, p. 35). The haze illusion is an optical illusion in which objects in hazy, visibility-hindering weather appear farther away than they actually are. It was captured as the *perceptual bias* of systematically overestimating the distance of objects in hazy weather. It was explained by the *perceptual heuristic* (if you will) of inferring the distance of an object by assessing the perceived clarity of its edges (and the fact that in such weather, perceived edge-clarity decreases). Kahneman and Tversky's program was spurred by their statistics students' continuously falling prey to seemingly intractable non-ideal reasoning—e.g., the facetious "law" of small numbers (Tversky & Kahneman, 1982c). To capture this, Kahneman and Tversky adapted the perceptual heuristics and biases paradigm to probabilistic judgment-making (this heritage is reflected in the initial reference to cognitive biases as "cognitive illusions").

Applications of the cognitive heuristics and biases paradigm progressed. This progression included both expanding the range of phenomena captured by the paradigm, and increasing the extent of capture—e.g., advancing from a vaguely recognized tendency, to identifying a systematic bias, to explaining the bias by a hypothetical heuristic, to providing the empirical corroboration and falsification of alternative explanations necessary to establish the heuristic's existence.

Just as Kahneman and Tversky adapted the perceptual "heuristics and biases" paradigm and applied it to probabilistic judgments, Nisbett and Ross (1980) in turn, adapted Kahneman and Tversky's paradigm and applied it to psychosocial judgments. For instance, Nisbett and Ross identified the *act-observer bias*, in which assessments of others' behavior overestimate the influence of character and underestimate the influence of circumstances. As more researchers contributed to the research program, the domains of judgment to which (adaptions of) the heuristics and biases paradigm were applied expanded to include syllogistic inferences, causal attribution, memory, and prediction. As the presumed bright line dividing reason from emotion was revealed to be more and more blurry, applications of the heuristics and biases paradigm expanded beyond solely *cold*—i.e., unemotional—reasoning and incorporated emotion-laden processing (e.g., the affect heuristic). Another area of progress is the application of the paradigm to assessments of value (e.g., sunk-cost bias, framing effects, and status-quo bias). (Baron, 2008; Connolly, Arkes, & Hammond, 2000; Gilovich, Griffin, & Kahneman, 2002; Koehler & Harvey, 2004; Schneider & Shanteau, 2003)

As such, the range of judgments captured under (adaptions of) the heuristics and biases paradigm include emotion-laden judgments, value assessments, and judgments regarding the social domain. These judgments include ingredients that in combination, seemingly get fairly close to a moral judgment. Applying (an adaption of) the heuristics and biases paradigm to the domain of moral judgments would be in keeping with the paradigm's history.

Of course, the above does not entail that the paradigm could be adequately applied to moral judgments, as hurdles to such application abound. The conclusions yielded by the preceding induction regard merely what is plausible (not to mention the limited scope of considerations in play). Nevertheless, the precedent of an expanding range of outputs and processes being captured under the paradigm at least supports the mere plausibility, bracketing philosophical considerations, that the paradigm could be applied to the moral domain, which implies the plausible existence of the right sorts of processes and judgments/decisions/actions to constitute moral analogues of cognitive heuristic and biases (bracketing philosophical considerations).

The preceding mutually-supporting inductions (§3.1 and §3.2) allow us to conclude that bracketing philosophical considerations, the existence of moral analogues of cognitive heuristics and biases is plausible, and sets the stage for a future, less cluttered addressing of the remaining philosophical considerations.

4. Potential Benefits

There are several potential benefits from developing a paradigm, and, ultimately, a psychological research program regarding moral analogues of cognitive heuristics and biases. One benefit is the acquisition of additional scientific knowledge. The interest in this knowledge is buttressed by our interest in the particular subject matter—that is, we care about morality. Furthermore, morality is a characteristic feature of our species' nature; as such, understanding human morality is an important part of understanding ourselves.

Another potential benefit concerns theoretical progress. The heuristics and biases paradigm has been successfully employed in various domains, including the perceptual, cognitive, and psychosocial domains. When viewed as an overarching research program, exploring the applicability of the heuristics and biases paradigm to the moral domain can be seen as a fitting step for a progressive scientific research program (*a la* Lakatos, 2000). Even if the application was unsuccessful, such a result would be useful for determining the boundaries of the program's domain of application. As such, investigating its application to the moral domain furthers the maturation of a fruitful scientific research program.

There are potential benefits to moral philosophy as well. For instance, if certain immoral acts are to be explained by bad reasoning—as opposed to for instance, indifference to moral considerations—this would seemingly have implications for assessments of moral responsibility.

A major potential practical benefit is the development of debiasing techniques to reduce the incidence of moral errors—i.e., immoral actions/decisions/judgments. This would be in keeping with other applications of the heuristics and biases paradigm, which have spawned successful debiasing techniques regarding perceptual judgment (e.g., pilots' susceptibility to autokenesis illusions—Civil Aviation Authority, 2002), cognitive judgment (e.g., confirmation bias amongst CIA analysts—Cooper, 2005), and psychosocial judgment (e.g., overly-attributing others' actions to their character vis-à-vis circumstances—Tetlock, 1997). It is worth noting that merely the identification of moral biases—i.e., independent of any knowledge of the generating processes—can yield useful techniques. For instance, knowledge of the relation between dramatic events and probability overestimation (i.e., the aforementioned availability bias) can yield the technique of investing in laborious calculation when making such assessments without any knowledge of the heuristic that explains the bias.

Another potential benefit stems from other explanations and accounts in a similar spirit to moral analogues of cognitive heuristics and biases—for instance, *moral blind spots* (e.g., per Banaji & Greenwald, 2013, or Bazerman & Tenbrunsel, 2011). An especially prominent example is *implicit biases* (e.g., Brownstein, 2015). Implicit biases, such as implicit racial bias, are unconscious biases that, notably, can persist despite conflicting with reflectively endorsed views. Implicit racial bias has recently received a lot of attention for the explanatory role it may play in the police treatment of African-Americans. Influential supporters of such explanations include former U.S. Attorney General, Eric Holder, and former FBI director, James Comey (Ehrenfreund, 2015; Kaplan, 2015). Implicit bias training has been included in prominent policy vehicles (e.g., the Justice Department's Ferguson report, and the consent decree with the Seattle Police Department), and such training has already been adopted by several police departments (e.g., the Seattle, New Orleans, and St. Louis police departments) (Green, 2015; Mullainathan, 2015). In addition, appeals to implicit biases have also been raised to account for phenomena in other social domains, including healthcare, education, and law (Brownstein, 2015).

Especially given the rising influence of implicit bias accounts, it is important that implicit biases have a firm conceptual and theoretical footing. Developing a paradigm for moral analogues of cognitive heuristics and biases could contribute to this footing. For instance, the meaning of 'implicit *bias*' is unclear with respect to which of the following potential constituents of 'bias' are necessary conditions: (1) merely a statistical pattern (e.g., kids are biased towards mint chocolate chip ice cream), (2) a thin sense of ‘discrimination’ as differential treatment that lacks any attribution of erroneousness (i.e., is non-normative), and (3) and erroneousness, including moral erroneousness. The success of the implicit bias research program to date may depend upon equivocations regarding these constituents and limiting research to coextensive cases. This opens the door to questions such as: Do implicit associations of groups with positive attributes constitute an implicit bias? Must the group be defined by an immutable trait (which is a classic legal standard for unjust differential treatment)? What grounds whether differential treatment constitutes an error, especially a moral error? The moral analogue of ‘cognitive bias’—or more precisely, ‘cognitive error’—could provide a standard of error and bias that resolves these questions. In short, the analyses and arguments in this dissertation have potential applications outside of research programs on cognitive heuristics and biases and moral analogues thereof.

Given these various potential benefits, developing moral analogues of cognitive heuristics and bias is a worthwhile endeavor.

CHAPTER 2: BACKGROUND

1. Cognitive Psychology's (Default) Explanatory Paradigm

The scientific discipline that studies heuristics and biases is cognitive psychology.[1] Per its default explanatory paradigm, cognitive psychology seeks to explain *cognitive* capacities—that is, *informational* capacities (e.g., perception, language use, and reasoning).[2] In some contexts, 'capacity' has a positive connotation (*a la* 'skill' or 'talent'). The sense meant here regards merely things one does, regardless of their value. For instance, perceiving an optical illusion (i.e., *misperceiving* the environment) constitutes a cognitive capacity.

Capacities are specified or analyzed in terms of dispositions (e.g., per Marr's task analysis—1983). Dispositions are conditional regularities between (a) conditions/factors affecting an object, and (b) that object's responses. For example, a disposition of salt is that it is water-soluble. This is formalized as follows. Salt is such that: *ceteris paribus*, were it (a) submerged in water, it would (b) dissolve.[3] *Cognitive psychological* dispositions are conditional regularities between (a) *information* affecting *a mind* (i.e., inputs), and (b) *that mind's informational* responses (i.e., outputs). For example, most people are such that: were they (a) asked, "What's two plus three?" (input), they would

[1] The following account relies heavily on Robert Cummins (1975; 1983, I-II; 2000). To some extent, it is a rational reconstruction that may deviate from current cognitive psychological practices. The terms used are meant per their sense in philosophy and may conflict with uses in cognitive psychology. This especially pertains to "disposition."

[2] The term, "cognitive," is somewhat tricky. In philosophy, "cognitive" sometimes means truth-aptness. This sense is *not* employed in this dissertation. In psychology, "cognitive" usually means information-processing, *simpliciter*—for instance, it encompasses visual and linguistic processing. This is the meaning of "cognitive" in "*cognitive* psychology" (Anderson, 2005; Balota & Marsh, 2004; Eysenck, 2001; Kellogg, 2002). However, in the forthcoming context of heuristics and biases, "cognitive" refers to a specific type of information-processing: that approximated by "thinking," "reasoning," or "inference-making." This meaning is exemplified by inferences regarding probability, and it excludes perceptual and linguistic processing. In this respect, perceptual biases and cognitive biases are distinct species of bias.

[3] Henceforth, *ceteris paribus* clauses will often be omitted and should be considered implicit.

(b) answer, “Five” (output). While such regularities are invaluable, they are (arguably) not explanatory; such regularities are not proper explanantia (i.e., that which explains), but explananda (i.e., that to be explained).[4]

Outputs are not necessarily behaviorally expressed; they can be internal mental states or “entities,” such as unspoken judgments. For example, consider the McGurk effect (McGurk & MacDonald, 1976). Humans are such that: were they (a) simultaneously exposed to both the sound, “ba,” and the sight of lip movements that produce the sound, “ga,” they would (b) perceive the syllable, “da” (i.e., judge the sound to be “da”). This perception of “da” constitutes an output, despite it occurring within the mind. Outputs are simply products of cognition (whether behaviorally expressed or not). Cognitive psychology is often interested in behaviors (e.g., *saying* “Five”), as operationalizations of mental outputs (e.g., the internal mental judgment that the answer is five). Likewise, observable inputs are often operationalizations of presumed perceptions. Inputs can also be internal (such as when the judgment that the answer is

[4] This position conflicts with the subsumptive or nomological approach (e.g., deductive-nomological explanation—Hempel & Oppenheim, 1948). That approach contends that the pertinent proper explanantia are laws or generalizations. Such generalizations explain by either subsuming phenomena or when conjoined with initial conditions, entailing outcomes. This approach was embraced by cognitive psychology’s principal competing paradigm, behaviorism (*a la* Skinner, 1953), and harmonized with behaviorism’s rejection of appealing to processes within the “black box” of the mind. While such laws or generalizations are a vital component of progress in cognitive psychology, they (arguably) are not (sufficiently) explanatory, at least in the special sciences (Cummins, 1983; 2000). Such generalizations can be used to justify the identification of a particular disturbance as the precipitating condition for a particular system’s transition from one state to another; however, *justifying* the identification of a precipitating condition is distinct from *constituting* an explanation (Cummins, 1983, p. 6). In other words, identifying the input that affected the output merely yields a disposition to be explained. The nomological approach can entail (or support the likelihood of) outcomes (i.e., outputs); however, such entailment constitutes prediction, which is insufficient for explanation. Generalizations can identify dispositions as a special case of a more general disposition; however, such identification does not *explain* the disposition, it merely contextualizes the disposition to be explained. At least in cognitive psychology, the proper explanans of such dispositions is process-models (Barsalou, 1992; Bechtel, Abrahamsen & Graham, 1999; Bechtel & Wright, n.d.; Bermudez, 2005; Botterill & Carruthers, 1999; Cummins, 1983; 2000; Reynolds, 2007). While I think this analysis is generally correct and generally yields the right prescriptions, perhaps its tone (maintained for illuminative purposes) is overly dismissive and its use of “proper” is too rigid.

five is an input that leads to a subsequent output, such as translating that answer into another language).

To introduce an important feature of cognitive psychological regularities, and, in turn, their explanations, consider the following. Some cognitive systems possess the disposition identified by the McGurk effect; some do not. Alternately framed, the McGurk effect is a *regularity* that applies to some cognitive systems, but not all. This is because the McGurk effect does not apply to cognitive systems, *simpliciter*. That is, something's constituting a cognitive system (e.g., computers, non-human animals) does not entail that it conforms to the McGurk effect. The McGurk effect is not a regularity between cognitive inputs and outputs, *simpliciter*; in other words, the McGurk effect is not a "universal law of cognition." Instead, it is a regularity between cognitive inputs to, and outputs from, a *particular* type of system. In this case, that type of system is the human cognitive system. In this respect, the McGurk effect is a *system-specific* regularity that describes a *system-specific* disposition.[5] Most, if not all, cognitive regularities and dispositions are system-specific in this sense. In turn, most cognitive psychological explanations are system-specific.[6]

To illustrate the way that cognitive psychology explains cognitive dispositions, let's start with a simple example and work our way up. To begin, *systems* are similar to *objects*, and *system*-specific dispositions are similar to *object*-specific dispositions. An example of an object-specific disposition is the aforementioned water-solubility of salt—

[5] The primary contrast to *system-specific* dispositions (and as mentioned soon, object-specific dispositions) is *universal* dispositions. Universal dispositions are dispositions possessed by *all* things. For example: Each thing is such that: were it (a) in motion, it would (*ceteris paribus*) (b) stay in motion. Arguably, such distinctions ground different explanatory paradigms (e.g., as reflected in differing uses of laws).

[6] Such system-specificity (or perhaps more aptly, system-*relativity*) will later bear upon constituting an objective (or objective-ish) cognitive psychological kind (Ch. 3, §3).

that is, salt is such that: were it (a) submerged in water, it would (b) dissolve.

Why does salt do this? Why does salt have this disposition? Salt has this disposition by virtue of salt's makeup (i.e., parts) and structure. As such, salt's having this disposition is explained by providing a depiction of salt's makeup and structure that renders explicit how possessing that makeup and structure is sufficient for possessing the disposition.[7] For (a simplified) example:

Salt's makeup consists of Na^+ atoms and Cl^- atoms. They are organized in a crystalline structure of Na^+ atoms bound together by Cl^- atoms and visa-versa. That crystalline structure provides salt its granular form. The bond between Na^+ and Cl^- atoms is not as strong as the bond each can form with water molecules—i.e., H_2O. Upon submergence in water, H_2O molecules, while zipping past the crystal, bind to Na^+ or Cl^- atoms. Upon binding, the H_2O molecules (moving in various directions), pull the Na+ and Cl- atoms they "grab" in those various directions. The bond between the Na+ and Cl- atoms is not strong enough to bring the H_2O molecules to an abrupt stop; instead, the H_2O molecules keep going and rip the Na+ and Cl- atoms away with them. In this way, the H_2O molecules tear the crystal (i.e., grain of salt) apart atom by atom—that is, the salt dissolves.[8]

[7] The depiction explains by rendering explicit the sufficiency of the makeup and structure for possessing the disposition. Some readers may object that *rendering explicit* constitutes an inappropriately psychological criterion. To the extent possible, such readers are encouraged to interpret this criterion as a place-holder for an acceptable objective criterion of causal-mechanical explanation.

[8] A potential quibble with this explanation is that it is does not explain salt's water-solubility because it vacuously identifies the bond between Na+ and Cl- as weaker than the bond each has with sets of water molecules; this is (allegedly) vacuous because these bonding differences is already captured by the disposition's (i.e., explanandum's) stating that salt is water-soluble. The problem with this quibble is that it presupposes an established identification of dissolution (as referenced in the explanandum) with the depicted molecular pulling-apart process. Even if correct, the quibble's upshot would merely be rendering the explanation not an explanation of salt's water-solubility, but an explanation of generic solubility, that is illustrated with salt and water. This does not impugn the explanatory power of explaining dispositions with makeup and structure-oriented depictions.

Having given the salt example, we can start working our way up to cognitive explanations by upgrading one feature at a time. The first upgrade concerns replacing "passive" parts with "active" parts.

Consider a traditional mousetrap (pictured below). It has the disposition of: upon (a) a mouse disturbing the trip (i.e., trigger), the mousetrap (b) crushes the mouse under the hammer (i.e., swinging bar).

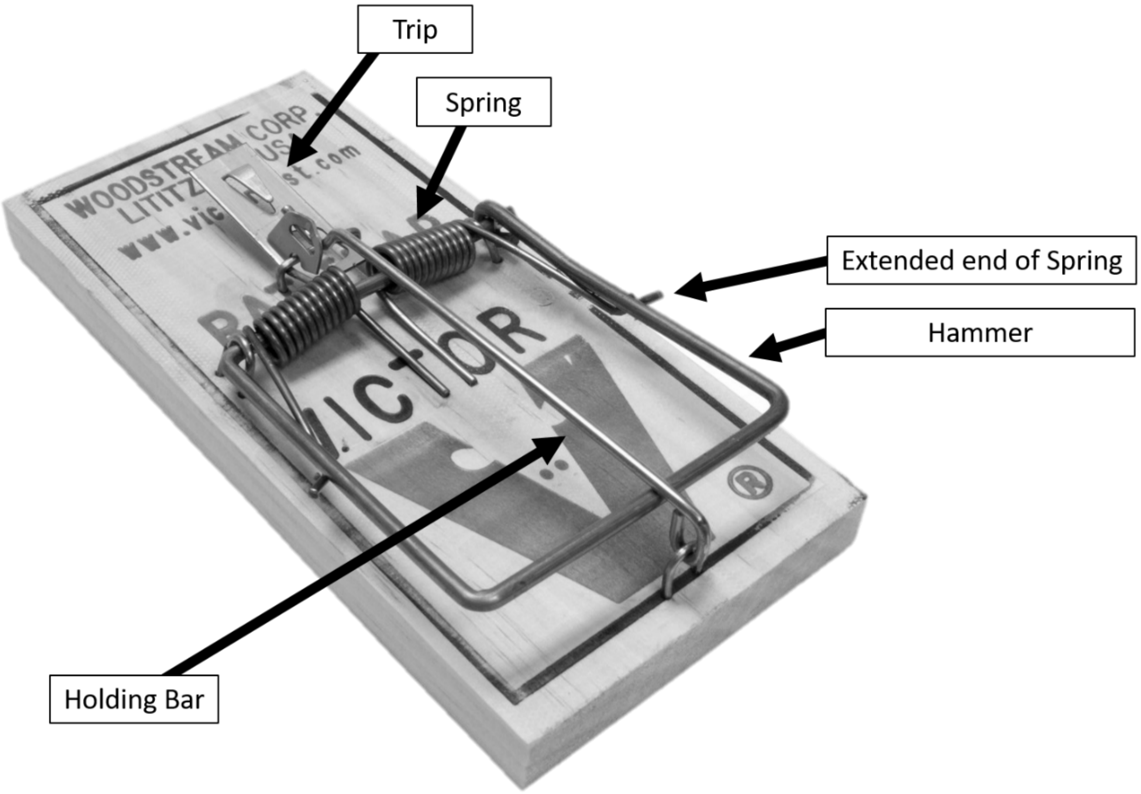

Figure 1. Victor metal pedal rat trap M200 (2019). This figure illustrates a traditional mousetrap and its parts.

The trip precariously restrains the holding bar, which holds back the hammer, which holds down the extended end of a tightly coiled spring. When the hammer is pulled back, it pushes down the extended end of the spring, which winds the coiled spring tighter (i.e., compresses the spring, which stores potential energy). Disturbing the trip releases the precariously restrained holding bar, which releases the hammer. This allows the hammer to be pushed by the extended end of the now unwinding (uncompressing) coiled spring, which swings the hammer down upon and mouse and crushes it. For simplicity's sake, we can use a bit of shorthand and just say that the spring "pushes" the hammer. In the mousetrap case, the disposition is manifested (in part) by virtue of parts "doing things," such as *pushing* the hammer.

In the salt case, there is a sense in which the parts of salt (i.e., the Na+ and Cl- atoms) were *passive*. They did not *do things*; instead, things were *done to* them (if you will). For instance, the Na+ and Cl- atoms did not *push* away from each other; instead, they were *pulled* apart. In light of this passivity, it was fitting to identify salt as an *object*. With more dynamic entities with more active and interactive parts (i.e., makeup) (e.g., the mousetrap), it is fitting to identify them as *systems* and identify their parts' activities as that system's *(inner) workings*.

When a system's parts "do things" that contribute to manifesting a disposition, the parts can be described and identified in terms of what function they perform in bringing about the disposition; in other words, the parts can be identified in terms of what contribution they make to (a) the antecedent affecting condition's yielding (b) the response.[9] Such identifications/descriptions (e.g., per *spring*) contrast with describing the constituents in terms of their composition (e.g., per *Na+* and *Cl-*). As such, we have moved from *compositional* descriptions to *functional* descriptions; in other words, a

[9] The functions in question here are *Cummins* functions (as opposed to proper or etiological functions).

functional perspective or level emerges. In the terminology of Daniel Dennett's framework (2002), we have moved from the "*physical* stance" to the "*design* stance."

With more active and interactive constituents, we begin to see more complex organization. For example, the mousetrap cannot manifest the aforementioned mouse-trapping disposition if the activities are performed simultaneously or at random. The activities must be performed in a specific order. For example, the spring cannot push the hammer before the holding bar releases the hammer. Such order in the performance of constituents' activities constitutes the organization. The organization is crucial to manifesting such dispositions, and its inclusion in the depiction is crucial to such explanations.

At this point, we can see that the explanatory paradigm in question is reflected in common explanations of an analogue clock's time-keeping capacity—or more precisely, its disposition to, upon (a) the passage of time, (b) track that passage. Such explanations explain the disposition by showing *how* the gears, etc. work. More precisely, the explanation explains a system's possession of a disposition by providing a depiction of its constitution and organization that renders explicit how possession of that constitution and organization is sufficient for possession of that disposition. This explanatory paradigm contrasts with appeals to laws and subsumptive generalizations (e.g., the deductive-nomological model of explanation—Hempel & Oppenheim, 1948; n. 4). In this respect, the explanations in question are causal mechanical explanations.

In some systems, such as mousetraps, watches, and vending machines, you can open up the system and see the inner workings. For example, you can see springs pushing hammers, gears turning dials, and spinning coils pushing bags of chips forward. In other words, the inner workings are tangible (if you will). The inner "workings" of a cognitive system are not tangible (in this sense); they are *informational* (though, assuming physicalism, ultimately token reducible to tangible physical constituents).

Things get very abstract on the informational level. Consider informational activities. For instance, in contrast to pushing levers and releasing latches, informational activities include adding numbers, alphabetizing names, and translating languages. Such informational activities are fittingly described as *operations*. One might think of the constituents of informational systems as that which performs such activities—for example, "adders" and "translators." However, these "entities" exist at very high levels of abstraction. Their abstraction is (presumably) similar to high-level computer software. Such abstraction is at a high level in the sense that there would be numerous levels to reduce through before one could connect activities such as translating a language to physical realizers. Indeed, this abstraction typically leaves the cognitive psychological level so far removed from the physical level that the entities performing the activities fall away. For instance, a particular addition operation might be performed by the "entity," "the calculator." We might be tempted to think of "the calculator" as a tangible thing that is distinct from the adding activity it performs (*a la* the spring that is distinct from its pushing-the-hammer activity). However, "the calculator" is not a thing in the way a spring, amplifier, or electrical distributor is. To illustrate this, consider a floppy disk for a chess video game. When used, what performs that activities that provide us a game to play? For instance, what makes the image of the chess board? We might think it is the floppy disk, itself. However, we can now download that same chess game. What provides the image of the chess board in that case? It cannot be a tangible entity, as no tangible entity has been added. Of course, if we drop down many levels of abstraction, it is all ultimately a matter of the states of transistors and other tiny tangible parts. Nonetheless, at the pertinent level of analysis, the activity of generating the chess board is performed by the chess *software*. There is not a thing of comparable tangibility to a spring.[10]

Ultimately, due to the cognitive level's extent of abstraction, our depiction of its "workings" need not feature parts that perform the activities. That is, we do not need to assign each activity to a performer. To do so can easily lead to problems. For one, it can lead to importing unjustified assumptions and entailments. For an extreme example, if the performer of the addition operation is dubbed, "the calculator," we might presume that the mind can perform any operation that we associate with the capacities of a typical digital calculator (e.g., identifying the square root of 11). Assigning performers to activities can also yield vacuous posits, such as saying the adding of 3 and 4 was performed by the "3 and 4 add-*er*." Such posits, though vacuous, can help us wrap our mind around abstract cognitive phenomena. For instance, as we subsume collections of dispositions under more general capacities (e.g., adding single-digit numbers), it is helpful to posit "a calculator" and think of it as that which performs the calculations. Nevertheless, it is important to remain aware of the nature of this crutch, and its limitations and pitfalls.[11]

Returning to the pervious narrative, the inner workings of a cognitive system are

[10] One upshot of the abstract nature of information is that it renders mental processing (at least in practice) highly unobservable (e.g., as compared with the circulatory system, which is observable upon dissection).

[11] Ultimately, the metaphysics of information, informational activities, "entities" (e.g., goals or desires), and programs (upcoming) is a very thorny matter. This is true even when dealing with artefactual informational systems, such as personal computers; however, with such systems, one can at least appeal to articulated designs and explicitly written lines of code. When dealing with evolved, biological information systems, the nature of these informational "entities" gets even trickier. This topic is beyond the scope of this dissertation.

cognitive processes (at least at the appropriate level of analysis). Cognitive processes consist of operations performed or executed in a particular order. That order and the more general organizational aspect of a cognitive process consists of the conditions governing the performance of operations. On the cognitive level, it is fitting to describe such organization as a *program*. The program "dictates" the execution of operations. More precisely, the program reflects the order and conditions under which particular operations are performed.

Without the benefit of tangible inner-workings, operations must often be discerned through functional analysis. Such analysis infers (or more precisely, generates hypotheses regarding) the operations and programs necessary to achieve the manifestation of the disposition. For instance, suppose a system manifests the disposition of calculating the cube of small numbers. For example, the input, 5, yields the response 125. A functional analysis of generating cubes reveals the intermediate step of generating the square of the number. That is, one "goes from" 5 "to" 125—i.e., "transforms" 5 into 125—by first, "transforming" 5 into 25. Given this method, the manifested disposition can be referred to as the *analyzed disposition*, and the operations inferred can be referred to as the *analyzing dispositions* (or *analyzing functions*, *functional roles*, or *activities*).[12]

[12] Carl Craver (2006) challenges the explanatory adequacy of such functional analytic models. He distinguishes "how-actually models" (i.e., that model how the system *actually* works) from "how-possibly models" (i.e., that generate the correct input-output relations, and thus, show how the system *might possibly* work, but do not necessarily show how the system *actually* works). Craver correctly assesses that how-actually models are explanatorily adequate and how-possibly models are not. He raises important worries about solely functional accounts' greater susceptibility to constituting merely how-possibly models and the expositions of these accounts' greater susceptibility to inadequately addressing "how-actuality." While he identifies various ways in which dual accounts (which include a componential analysis—i.e., a specification of the realizers of functional roles) allow for a variety of tests of how-actuality that are not available to solely functional accounts, nevertheless, the susceptibility of componential-inclusive accounts to those tests merely reflects such accounts' greater in-practice accessibility to how-actuality testing—i.e., it does not negate the legitimacy of methods of how-actuality testing available to solely functional accounts, let alone challenge solely functional accounts' in-principle susceptibility to how-actuality testing. While componential supplements to functional

Classical cognitive psychology depicts such processing in terms of symbolic computational transformations (*a la* digital computing).[13] The informational input triggers computational processing that, through the execution of operations in accordance with a program, transforms the informational input into an informational output.

One possesses a cognitive disposition *by virtue of* possessing (1) the analyzing dispositions—i.e., operations—and (2) their conforming to the organizing program; together, these comprise a *cognitive process*. It will sometimes be tempting to identify the cognitive process or the program as the performer of operations. This metaphorical language is unproblematic so long as one keeps in mind that it is merely metaphorical.

As with the salt, mousetrap, and other cases, cognitive psychological explanation capitalizes upon this *by-virtue-of* relation. In this respect, cognitive dispositions are explained by providing a depiction of the cognitive processes' operations and program that renders explicit how a system's possession of that process is sufficient for its possession of the disposition.

With respect to typical cognitive input-output dispositions, the cognitive process is that *by virtue of* which the reception of the input is followed by the production of the output. This yields a linear dynamic of: input-process-output. In a sense, such outputs are

accounts are very valuable for efficient how-actuality testing, nevertheless, their inclusion is neither a necessary condition of demonstrating resilience to how-actuality falsification nor a necessary condition of explanatory adequacy. Furthermore, such componential supplements are often unrealistic at the current level of development of cognitive psychology because of the number of layers between high-level operations and their physical realizers. In other words, componential supplements to a functional analysis of language translation is currently just an ideal, and research into such capacities should not be foresworn until such supplements are available. Furthermore, the wedding of high-level functional accounts and componential realizers is most likely to come from their meeting in the middle (if you will), as opposed to merely slogging uphill from the lower levels. In other words, a more effective strategy is simultaneous research in cognitive psychology, cognitive neuroscience, and neuroscience.

[13] *Classical* cognitive psychology contrasts with *connectionist* cognitive psychology, which models processes in terms of non-symbolic information propagating through artificial neural networks. Kahneman and Tversky's program utilizes the classical paradigm.

a function of the input, and the process' program specifies that function. Loosely speaking, the process "transforms" the input into the output.[14]

Such input-output dispositions are explained by providing a depiction of the mediating process that renders explicit why possession of that process is sufficient for possession of the disposition. Such depictions are contained in input-process-output models. Such process-models are (at least, akin to) causal-mechanical explanations on a cognitive level—that is, an informational level.[15]

The following is an example of a process model that explains a cognitive capacity. The capacity explained is determining whether a positive, whole number is prime. This capacity is formulated in terms of the following disposition: a regularity between (a) inputs of positive, whole numbers, and (b) outputs of correct displays of "Prime" or "Not Prime."

Explanandum: Disposition/Regularity:

Input: Positive, whole number

Output: Accurate display of "Prime" or "Not Prime"

Explanans:

Operations:

(a) Assess whether input is 1

(b) Assign input to dividend

(c) Assign 2 to divisor

(d) Assess whether the divisor is greater than one-half the dividend

(e) Assess whether the quotient is a whole number

(f) Add 1 to the divisor

(g) Display "Prime"

[14] Henceforth, scare-quotes for "transform" should be considered implicit.

[15] There is some room for quibbling over the causal-mechanical nature of such process models. Such quibbles can stem from disagreements over views of causation and problems concerning mental causation. Nonetheless, identifying such models as causal-mechanical (or something in that ballpark) is informative in a context in which such models are contrasted with the principal competing explanatory paradigm—namely, the subsumptive or nomological approach (e.g., deductive-nomological explanation) (see n. 4).

```
        (h) Display "Not Prime"
Organizing Program:
        (1) Perform (a)
        (2) If yes, perform (h) and stop
        (3) If no, perform (b)
        (4) Perform (c)
        (5) Perform (d)
        (6) If yes, perform (g) and stop
        (7) If no, perform (e)
        (8) If yes, perform (h) and stop
        (9) If no, perform (f)
        (10) Return to (5)
```

Cognitive processes are unobservable theoretical posits (*a la* subatomic particles). As such, claims about them must be inferred from observable phenomena. Such claims are inferred from observable input-output relations. (More precisely, such claims are inferred from relations between observable operationalizations of mental inputs and outputs.[16]) Such input-output relations are abstracted from sets of input-output pairs. An exemplar input-output pair is a judgment (a concept that will figure centrally in the chapters to follow). "Judgment" is sometimes construed as referring solely to an output; however, (at least in cognitive psychology) "judgment" implicitly refers to an input-output pair—that is, an output per an input.[17] Recall that per the cognitive psychological paradigm, inputs trigger cognitive processes, which transform those inputs into outputs.

[16] Regarding distinguishing operationalizations of inputs and outputs from their target constructs, sometimes, such distinctions can be ignored. For instance, regarding multiplying 8 and 5 in one's head, a person's *speaking*, "40," is, to be precise, an operationalization of the output of *concluding* that the answer is 40. In other cases, such as when a speaker's sincerity may be doubted, the distinction between an operationalization and the construct in question can be important (e.g., regarding answers to racially charged questions).

[17] This implicit reference to an input contained in 'judgment' (in the cognitive psychological sense, at least) is illustrated by the attributing of properties such as fallaciousness to judgments. Such properties can only be applied to an output *per an input*. This is akin to how an answer can only be erroneous *per a question*. For instance, even the answer, "P & ~P," can be correct per a certain question (e.g., what does applying *modus ponens* to Q and (Q > (P & ~P)) entail?).

Given the reception of a particular input, the particular output yielded is a function of the transformations executed. In addition, properties of the mediating processes can be reflected in properties of the mediated outputs-per-inputs. For instance, usually, the greater the number of transformations executed, the longer the time that elapses between the input and output. In this respect, judgments are a function and reflection of mediating processes. This is the basis of inferences from input-output relations to processes.

A single judgment rarely provides much information about the mediating process. This is because there are usually too many plausible processes that could equally well yield that particular input-output pair. For instance, an input-output pair of _2_ (input) - _4_ (output) does not tell you whether the mediating process is _+2_, or _*2_, or _^2_, or _(^3)-6_, etc. Unsystematic judgments are also not very informative. Such judgment sets suggest that different inputs in the set are triggering different processes. The plausible sets of processes that can account for such judgment sets are practically limitless. For instance, the set of input-output pairs, {5-8, 4-14, 2-11, 3-3, 9-1}, is not very informative. However, systematic judgments are informative, as they suggest mediation by a coherent process—that is, a process with a coherent program. For instance, the set of input-output pairs, {2-4, 3-6, 4-8, 5-10…}, suggests that the mediating process is: _*2_. Initially, an account of the process is put forward as a hypothesis. If that hypothesis is resilient to falsification, alternative explanations, etc., the process' existence is accepted.

With this background on cognitive psychology's explanatory paradigm, we can now turn to one of cognitive psychology's most influential discoveries: heuristics and biases.

2. Cognitive Heuristics & Biases

Heuristics and biases are notable in the social sciences for providing an alternative to descriptive rational choice theory (DRCT). DRCT was the dominant descriptive model of human judgment. In fields such as neoclassical economics, it still is. In many ways, the heart of heuristics and biases is its contrast to DRCT.

2.1. Strong-DRCT

Strong versions of DRCT (*strong-DRCT*) contend that humans are ideally rational.[18] Exemplars of ideal rationality include adherence to the rules and principles of logic, statistics, and probability theory (e.g., *modus tollens*, the relevance of sample size to inductive credence, and Bayes' theorem).[19] Strong-DRCT depicts the typical person as "homo-economicus"—a forward looking, self-interested, expected-utility maximizer who possesses consistent and stable preferences. Strong-DRCT entails that the cognitive processes underlying human judgment instantiate algorithms that are highly complex, computationally demanding, and consistent with sophisticated principles of ideal rationality (Simon, 1990, p. 194-198).[20] (Doherty, 2003, p. 658-659; Gilovich & Griffen, 2002, p. 1-6; Goldstein & Hogarth, 1997, p. 11; Little, 1991, p. 66) Henceforth, *ideal* rationality will often be referred to as simply, "rationality."

While strong-DRCT contends that "humans are rational," it does *not* contend that

[18] "Strong-DRCT" (and forthcoming, "weak-DRCT") are reconstructions of pivotal poles in DRCT theory. There are versions of DRCT that do not neatly fit this dichotomy.

[19] '*Ideal* rationality' is loosely synonymous with 'expected utility theory' (von Neumann & Morgenstern, 1947), 'subjective expected utility theory' (Savage, 1954), 'full rationality' (Selton, 2001), the 'standard idealization' (Cherniak, 1994), the 'normative standard' (Samuels, Stitch, & Faucher, 1999), and the 'standard picture' of rationality (Stein, 1996). Its most pertinent contrast is with models whose norms are tailored to accommodate human limitations (e.g., processing and memory limits); such models include bounded rationality (Simon, 1990) and prescriptive (i.e., pragmatic) models of decision-making (Baron, 1994).

[20] Whether strong-DRCT *entails* or *implies* ideally-rational processing—or does neither—is contestable. On the one hand, "homo-economicus" could be merely an *as if* model; on the other hand, inference-to-the-best-explanation is seemingly sufficient for at least, *implying* ideally-rational processing. I suspect that few DRCT theorists would have defended "homo-economicus" as a psychological theory; however, that is a separate question from whether such was entailed or implied.

every human judgment is rational. That is, strong-DRCT allows for some irrational judgments. Per terminological conventions, such irrational judgments are often referred to as “errors” or being “erroneous”; such “errors” can be presumed to be irrational and have the theoretical implications of irrational judgments.[21] The irrational judgments or “errors” that strong-DRCT allows for are those that result from breakdowns and mistakes in carrying out otherwise rational judgment formation. That is, strong-DRCT allows for breakdowns and mistakes in the *execution* of rational programs. That is, strong-DRCT merely contends that the *programs* of judgment formation are rational.

To elaborate (though still a simplification): an input triggers a program, and that program may or may not be successfully executed. Strong-DRCT contends that the programs are such that: *were* they successfully executed, they would yield rational judgments. The important feature of the judgment errors allowed for by strong-DRCT is

[21] There is some terminological tension here regarding rational/irrational and correct/erroneous. DRCT and the other theories of reasoning mentioned in this dissertation make claims regarding the rationality of human reasoning. Judgments whose outputs are the products of irrational reasoning can be referred to as irrational judgments. Psychological research tends to assess judgments not in terms of rational/irrational, but in terms of correctness and erroneousness. This is a sensible practice on epistemic-conservativism grounds. For instance, an experimenter’s knowledge of a question’s correct answer is sufficient for attributing erroneousness to an experiment participant’s divergent answers; attributing irrationality is a much trickier matter. Rationality and correctness can diverge. For instance, rational reasoning from information that is generally reliable, but in a particular case, turns out to be erroneous, can yield erroneous judgments. As such, precisely speaking, erroneousness is not sufficient for irrationality. However, cognitive psychological theoreticians and researchers speak as if it were sufficient. For instance, it is standard practice to speak about the implications of errors for DRCT despite the fact that such errors bear upon DRCT only if the errors are also irrational. This practice ends up being logically unproblematic because the practitioners exclude rational, erroneous judgments from the domain of pertinent judgments. Loosely speaking, an experiment that yields rational, erroneous judgments is considered a badly designed experiment whose findings are not theoretically relevant. In other words, erroneous judgments can be presumed to also be irrational because rational errors are tossed out as “not really errors.” Alternatively framed, in the pertinent contexts, “erroneousness” consistently means erroneous *and irrational*. While this use of “erroneousness” is imprecise, it is so entrenched that breaking from this convention (e.g., replacing such uses of “erroneous” with “irrational”) would probably cause more confusion than it would avoid. As such, I will follow convention and use “error” in the error-and-irrational sense. That is, in this dissertation, judgments identified as erroneous can be presumed to also be irrational (unless explicitly stated otherwise). This convention is asymmetrical in that correct judgments cannot be presumed to be rational; for instance, a prominent feature of heuristics is that they often generate judgments that are correct, but not (ideally) rational.

that the judgments' erroneousness is not attributable to the program. For a useful analogy, consider a perfect airplane design. Airplanes built from a perfect design can still be susceptible to mechanical failures. For instance, builders can fail to follow the design. Parts can wear out. Birds can be sucked into the engines. Various factors can happen to align in a way that causes a failure. Such failures are not the designers' fault. The fault may lie with the builders, the maintenance mechanics, or no one at all, such as when the failure is only attributable to bad luck. While the design is an important factor in the determination of the outcome, the design is not the only factor. Failures that occur are not necessarily attributable to design problems.

A cognitive program is like an airplane design in that both provide instructions (loosely speaking). In both cases, a perfect set of instructions does not necessarily entail a perfect execution of the instructions. Just as a perfect airplane design does not preclude mechanical failures, similarly, a perfectly rational program does not preclude breakdowns and mistakes that yield errors. In this respect, while strong-DRCT contends that the programs of judgment formation are rational, this contention does not preclude erroneous judgments. In sum, strong-DRCT entails that while the programs that generate judgments are ideally rational, they are not necessarily successfully executed; strong-DRCT allows for errors that are *not* attributable to programs.

This next part is a bit tricky. For the time being, breakdowns and mistakes that are not attributable to a flaw in a process' program can be assumed to be (relevantly) "random" breakdowns and mistakes (or more precisely, *relevantly unsystematic* breakdowns and mistakes).[22] Recall that features of an input-output pair reflect features

[22] The terms, "random"/"unsystematic" and "systematic" are potentially misleading (especially for philosophers). Recall from chapter 1 (n. 18) that such "systematicity" and "randomness" are *only* with

of the mediating process. Judgment formation that is susceptible to (relevantly) random breakdowns and mistakes is susceptible to correspondingly (relevantly) random judgment errors. Given a presumption in favor of processes' rationality, the extent of irrationality in a process suggested by (relevantly) random judgment errors is merely, susceptibility to (relevantly) random breakdowns and mistakes. As such, (relevantly) random judgment errors are consistent with strong-DRCT. (Relevantly) systematic errors, however, are not. The coherence of (relevantly) systematic errors suggests that such errors are attributable not to (relevantly) random breakdowns and mistakes, but from a (relevantly) coherent program that itself deviates from rationality. As such, (relevantly) systematic errors are *inconsistent* with strong-DRCT.

In sum, strong-DRCT contends that programs are ideally rational, but not necessarily successfully executed. In terms of competence vis-à-vis performance (as in linguistic competence vis-à-vis performance—Chomsky, 1980), programs that generate judgments can be construed as reasoning competences; the transformations executed (whether in accordance with such programs or not) can be construed as reasoning performances (Samuels, Stich, & Faucher, 1999). In these terms, strong-DRCT contends that humans possess ideally rational reasoning competences, are susceptible to

respect to useful factors/conditions/categories in psychology—i.e., the sorts of things relevant to psychological descriptions and topics. In other words, such systematicity/randomness concerns the presence/absence of patternly-ness between the sorts of things that would be captured as values or variables in a psychologist's regression equation. Attachment to metaphysical notions of systematicity/randomness can cause tremendous confusion here; to the extent possible, it is best to set those notions aside. For an extreme example, if you are disposed to fail to make correct judgments when a tank is running you over, this technically constitutes a systematic conditional regularity, but not one relevant to cognitive psychological inquiry. The presented senses of (relevantly) "random"/"unsystematic" and "systematic" are canonical in cognitive psychology. I cannot attest to this approach being perfect, but it certainly is useful. Perhaps a critique of it can be marshalled, but that is beyond the scope of this dissertation. In this respect, the adequacy of this approach is an assumed premise of this dissertation.

(relevantly) random performance errors, but are not susceptible to (relevantly) systematic performance errors.

2.2. Weak-DRCT

Contrary to strong-DRCT, it was discovered that humans indeed make (relevantly) systematic errors (e.g., Kahneman, Slovic & Tversky, 1982a; Gilovich, Griffen & Kahneman, 2002).[23] As such, strong-DRCT does not accurately describe human judgment. However, weaker versions of DRCT (*weak*-DRCT) can explain certain systematic errors by replacing the strong contention—that human judgment, *simpliciter*, is ideally rational—with a weaker contention, such as: merely the default processes, basic orientation, or "core" of human judgment is ideally rational.[24]

The important move here is taking the processes and mental phenomena that bring about judgments and dividing them into two groups. The first group is processes that belong to a core judgment system (or something of that sort). Weak-DRCT contends that these processes are rational. The second group is processes and "entities" (e.g., urges) that are not part of that core judgment system, but nonetheless, can affect judgment formation and determine what judgment is ultimately made. These processes and phenomena can be irrational.

Weak-DRCT identifies the core judgment system as (something like) the default judgment system. The core system includes a repertoire of rational processes. When processes or entities from outside the core judgment system affect judgment formation, such affecting constitutes *interfering with*, *disrupting*, or *overriding* the core judgment

[23] Henceforth, regarding "random"/"unsystematic" and "systematic," the qualifier, "(relevantly)," will often be omitted and should be considered implied.

[24] I do not necessarily contend that this is a viable distinction.

system's processes.

Weak-DRCT is particularly well suited to accounting for motivated irrationality, such as wishful thinking. The following illustrates a simplified, stereotypical weak-DRCT account of wishful thinking. A plane carrying a young woman crashes into the ocean. Her father initiates a rational process of assessing whether his daughter survived. His intense desire for her survival interferes with that process and prevents him from reaching the conclusion that she is dead. The result is an erroneous judgment that she must have swum away from the crash and is safe and happy on a deserted island.

In this case, the core judgment system, in accordance with a rational reasoning program, initiates a rational reasoning process. That process' formation of a judgment is affected by the intense desire for a given conclusion. The ultimate result is the yielding of an irrational judgment.

This account is consistent with weak-DRCT insofar as it does not posit irrational processes that belong to the core judgment system. Roughly speaking, weak-DRCT can account for irrationality and systematic errors that can be "blamed" upon a process or mental phenomenon from outside the core judgment system. For instance, suppose it were the case that everyone was such that: (1) were they assessing whether a loved one had died, they would draw a similarly irrational conclusion, and (2) were they in any other situation, they would only draw rational conclusions. These dispositions would yield sets of judgments with systematic errors. Namely, judgments' erroneousness would correlate with whether they were assessing whether a loved one had died. As such, mere systematic error is insufficient to falsify weak-DRCT.

In competence/performance terms, weak-DRCT allows for systematic performance errors, but only those that are attributable to processes and entities from outside the core judgment system. Weak-DRCT contends that the reasoning competences of the core judgment system are rational.

2.3. Heuristics & Biases

Daniel Kahneman and Amos Tversky provided judgment models that fundamentally departed from both strong-DRCT and weak-DRCT (e.g., prospect theory—Kahneman & Tversky, 1979). They contend that many of the *core* judgment system's processes were indeed irrational—that is, the processes deviated from ideally rational algorithms. This deviation did not consist of merely breakdowns and mistakes in executing algorithms; it did not consist of merely algorithms being disrupted, interfered with, or overridden. Instead, Kahneman and Tversky contended that many of the programs were adhering to wholly different inferential principles and that these principles were not ideally rational. Kahneman and Tversky were right.

A classic illustration of this involves the following experiment (Tversky & Kahneman, 2003). Experiment participants (i.e. subjects) were given personality descriptions, such as:

> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well-liked by his colleagues. (p. 5)

The experiment participants were told that the descriptions were of individuals, randomly sampled from a population of 100 (in total) engineers and lawyers. Participants were instructed to assess the probability that a given description was of an engineer or a lawyer. One group of participants were told that the population consisted of 70 engineers and 30 lawyers. The other group was told the population had the exact opposite composition—30 engineers and 70 lawyers. In the terminology of probability theory, the composition of the population constitutes a *base-rate*. Probability theory dictates that base-rates matter. This is especially the case with Dick, since his description provides little information regarding whether he is an engineer or a lawyer. That is, probability theory dictates that the probability that Dick is an engineer is much higher when the population base-rate is 70/30 engineers/lawyers, than when it is 30/70 engineers/lawyers. Since DRCT contends that humans are ideally rational -which includes adhering to the principles of probability theory- DRCT entails that the probability estimates would vary greatly between the participant groups. They did not. Both groups estimated the probability at approximately 0.5, regardless of population base-rate.

(Upon resilience to falsification and alternative explanations) this showed that people were often entirely insensitive to base-rates.[25] This is an important finding because it showed that participants did not merely execute the algorithms of probability theory *poorly*; for instance, they did not merely under-weight base-rates. Instead, they *did not execute those algorithms at all*; they employed an entirely different inferential procedure (in this case, the representativeness heuristic—discussed shortly). If people were merely executing the algorithms poorly, this would be relatively harmonious with

[25] Readers should note that these findings are well established. In this dissertation, unless indicated otherwise, mentioned experiments are intended to illustrate the nature of the evidential support for findings and generally illuminate the subject matter. As is often the case with any isolated experiment, the resultant data will be consistent with numerous explanations. Alternative explanations to those presented in the dissertation may occur to the reader and appear to undermine the findings; however, readers should keep in mind that Kahneman and Tversky's program has been well aware of a variety of alternative explanations and has taken measures, including subsequent experiments, to rule out many of them. This is not to say that an alternative explanation that might occur to the reader has necessarily been ruled out; it is only to say that one should not assess the feasibility of such alternative explanations merely by their consistency with the evidence that happens to be mentioned in this dissertation.

DRCT. More specifically, it would be consistent with bounded rationality variants of DRCT (e.g., Simon, 1990). Such variants conceded that people are not ideally rational, but only conceded it insofar as people can fail to execute ideally rational algorithms correctly. Such execution failures are still consistent with the basic orientation of people's cognitive systems being ideally rational. Loosely speaking, such errors would be mere performance errors. Kahneman and Tversky showed that many irrational judgments were not merely performance errors; they showed that the core judgment processes—i.e., the reasoning competences—were not ideally rational. (As with "capacity," "competence" does not necessary connote positive value). (Goldstein & Hogarth, 1997, p. 24-25; Kahneman, Slovic, & Tversky, 1982a; Kahneman & Tversky, 1996; 2000; Tversky & Kahneman, 2003).

Kahneman and Tversky primarily focused on intuitive judgment.[26] They contended that several intuitive processes constituted *cognitive heuristics*.[27] Cognitive heuristics are distinct processes of judgment formation that constitute reasoning shortcuts

[26] Tversky and Kahneman's conception of "intuitive judgment" changed over time. Initially, they defined it per Braine's (1978) formulation in which "A statement is intuitive only if its truth is immediately compelling and if it is defended in a single step" (Kahneman & Tversky, 1982b, p. 499). They also employed senses in which "intuitive judgment" was characterized as "an informal and unstructured mode of reasoning" (vis-à-vis "analytic methods or deliberate calculation") (p. 494), and "natural assessments that are routinely carried out as part of the perception of events and the comprehension of messages" (Tversky & Kahneman, 2002, p. 20). They situated intuitive judgment on a middle ground between perception and deliberation (Keren & Teigen, 2004, p. 93). Their later conceptions of "intuitive judgment" (vis-à-vis deliberative judgment) were more closely aligned with the system 1 (vis-à-vis system 2) judgments featured in early dual-processing theories of judgment (Evans, 2008) (*system* 1 and 2 were later replaced with *type* 1 and 2—e.g., Evans & Stanovich, 2013).

[27] The meaning of "heuristic" varies between disciplines and even within psychology (Hahn, Frost, & Maio, 2005). In this dissertation, "heuristic" connotes the classical conception from Kahneman and Tversky's program; it is distinct from the conceptions of the Gigerenzer school (Gigerenzer, Todd, & the ABC Research Group, 1999), the *cognitive miser* model (Fiske & Taylor, 1991), the *errors and biases* program (Nisbett & Ross, 1980), and the attribute-substitution model (see Ch. 3, n. 27) (Kahneman & Frederick, 2002). Readers should note that reliance upon familiarity with the term, "heuristic," from other contexts may be misleading as to its meaning within Kahneman and Tversky's program.

(vis-à-vis ideally rational algorithms); they are efficient and useful, but flawed.[28] As such, the judgments they produce are prone to characteristic errors (as assessed by inconsistency with the judgments entailed by ideal rationality).[29] Such characteristic errors instantiate *cognitive biases*, which are systematic errors (within a set of judgments).[30] Per this dynamic of being practical, yet flawed, heuristics are characterized as "quick and dirty" and "rough and ready" means of problem-solving (respectively: Gilovich & Griffen, 2002, p. 3; Over, 2004, p. 10). (Kahneman & Tversky, 1982b, p. 494, 499; 1996, p. 582; Tversky & Kahneman, 2002, p. 20; 2003, p. 35, 52)[31]

2.4. Representativeness Heuristic

An exemplar heuristic is the *representativeness heuristic* (Kahneman & Tversky, 1982a; 1982c; Tversky & Kahneman, 1982b; 1982c; 2002; 2003). "Representativeness"

[28] This dissertation focuses on the classical conception of cognitive heuristics (Ch. 1, n. 6) (e.g., Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1996; 2000). There is a subsequent conception per attribute-substitution (Kahneman & Frederick, 2002, p. 53-60); however, the classical conception is more useful for the purposes of adaption—this is explicated in chapter 3, n. 33 (after the main text introduces the analysis needed for the explication).

[29] The standard by which the program attributes erroneousness will be important later. That the CHB program's standard of error-attribution is per inconsistency with ideally-rational judgments) is illustrated in: e.g., Kahneman, Slovic, & Tversky, 1982b, p. xi-xii; Kahneman & Tversky, 1982a, p. 48; 1982b, p. 494; 1982c, p. 46; Tversky & Kahneman, 1982a, p. 24; 1982c, p. 91; 2002, p. 20; 2003, p. 38. It is also supported by: e.g., Gilovich & Griffen, 2002, p. 1; Goldstein & Hogarth, 1997, p. 5-6; Keren & Teigen, 2004, p. 93.

[30] The term, "cognitive bias," is sometimes used to refer to the disposition to generate judgments that systematically deviate from correctness (Keren & Teigen, 2004). This dissertation adopts the usage that refers to systematic deviation within a set of judgments or outputs. There is some ambiguity regarding the extension of cognitive biases. This stems from the question of how to classify instances in which a heuristic yields a judgment that *conforms* with ideal-rationality because the property that systematically correlates with judgments' deviations from ideal-rationality is absent (e.g., regarding the aforementioned availability heuristic, probability estimates of *non-dramatic* events, *ceteris paribus*). Cognitive biases are systematic errors—i.e., deviations from ideal-rationality *per* the presence of a property in inputs. One interpretation of this is that the extension of a cognitive bias is the judgments that deviate from ideal-rationality. Another interpretation is that the extension includes the judgments in which the systematic pattern manifests. The manifestation of the systematic pattern requires not merely instances of erroneousness being *present* when the property (e.g., dramatic-ness) is *present* in inputs (e.g., events whose probability is assessed), but also instances of erroneousness being *absent* when the property is *absent*. In order for a heuristic to explain a cognitive bias, it has to explain both types of instances.

[31] This extensive citation partly reflects the Kahneman and Tversky program's less than perfect fidelity to authoritative definitions.

(as in, "A is highly representative of B") is a relation akin to prototypicality, stereotypicality, or exemplariness. Tversky and Kahneman characterize it as "an assessment of the degree of correspondence between a sample and a population, an instance and a category, an act and an actor, or more generally, between an outcome and a model" (Kahneman & Tversky, 1996, p. 584).[32] For instance, a personality sketch can vary in the extent to which it is representative of one group versus another. For example: "Pat played with trucks as a child, played football in high school, and worked in construction as an adult." This personality sketch can vary in the extent to which it is representative of American males versus American females. Other examples of the representativeness relation include: a comment being representative of a type of person (e.g., a rigid ultimatum being representative of a military hawk), a sequence of heads and tails being representative of a fair coin (e.g., H-T-T-H-T-H is perceived as more representative of tosses of a fair coin than is H-H-H-T-T-T), and a (e.g., high) price of gold being representative of a (e.g., declining) world economy (Tversky & Kahneman, 1982b, p. 85).

The "logic" or principles of representativeness differ from the principles of probability theory (and thus, ideal rationality). Probability theory includes the *conjunction rule*, according to which, the probability of a conjunction (A&B) cannot be greater than the probability of one of its conjuncts (e.g., A): P(A&B) ≤ P(A). For example, the probability of Pierre being an accountant who speaks French (A&F) cannot be greater than the probability of Pierre being an accountant (A). This is because

[32] Kahneman and Tversky do not provide an "*a priori*" definition of *representativeness*; they define it "empirically" based upon usage (Tversky & Kahneman, 1996, p. 585). Criticism of this approach can be found in Gigerenzer (1996, p. 594; 1998, p. 2-3).

everyone who is an accountant who speaks French (A&F) is necessarily an accountant (A). Otherwise stated, the probability that an object is a member of a subset cannot be greater than the probability that it is a member of an encompassing set. For example, since an apple (A) is a fruit (F), the probability that the thing rattling around in a lunchbox is an apple (A&F) cannot be greater than the probability that it is a fruit (F).

Violations of the conjunction rule constitute commissions of the *conjunction fallacy*. While the conjunction rule applies to probability, it is not part of the *logic of representativeness* (Tversky & Kahneman, 1982b, p. 90). For example, someone who rides the subway, talks fast, and hates the Red Sox is more representative of the group, New Yorkers, than the group, Americans; this is the case regardless of the fact that New Yorkers are a subset of Americans.

2.5. The Linda Problem

Kahneman and Tversky discovered that when people are asked to assess probability, they often assess representativeness instead. One prominent demonstration of this involved the following question, which we can refer to as the "Linda problem" (Tversky & Kahneman, 1982b, p. 92):

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable:
>
> (a) Linda is a teacher in elementary school.
> (b) Linda works in a bookstore and takes Yoga classes.
> (c) Linda is active in the feminist movement
> (d) Linda is a psychiatric social worker.
> (e) Linda is a member of the League of Women Voters.

> (f) Linda is a bank teller.
> (g) Linda is an insurance sales person.
> (h) Linda is a bank teller and is active in the feminist movement.

The pertinent statements are (c), (f), and (h). Experiment participants perceived Linda's personality sketch to be highly representative of feminists, not representative of bank tellers, and somewhat representative of feminist bank tellers. When assessing the extent to which a profile (Linda) is representative of a group (feminist bank tellers) that is defined by membership in both a highly representative group (feminists) and an unrepresentative group (bank tellers), the high and low extents of representativeness are balanced out to produce a medium representativeness assessment. As such ("R" for representativeness): R(Feminist) > R(Feminist & Bank teller) > R(Bank teller). Because the logic of representativeness does not include the conjunction rule, this ranking is unproblematic.

The conjunction rule prohibits ranking "Feminist & Bank teller" as more probable than "Bank teller" (as the former is a subset of the latter). Nevertheless, when asked to assess the probability of Linda's membership in the aforementioned groups, subjects committed the conjunction fallacy. We can refer to such commissions as the *Linda-conjunction-fallacy*.[33] They also provided the same rankings given by others asked to assess the extent to which Linda was *representative* of those groups (Tversky & Kahneman, 1982b, p. 94). Tversky and Kahneman dubbed this (type of) phenomenon, the *conjunction effect*.

Responding to the Linda Problem with a ranking in which "Feminist & Bank

[33] I will often use hyphens with multiple-word terms to facilitate readability.

teller" is more probable than "Bank teller" constitutes an erroneous judgment. The systematic commission of such errors renders them instantiations of a cognitive bias. The bias's systematicity suggests that its constitutive judgments are mediated by a coherent process—that is, processing in accord with a coherent program. Kahneman and Tversky put forward the representativeness heuristic as an account of that program. In other words, they explained the Linda-conjunction-fallacy (and the conjunction effect) as the result of subjects employing the representativeness heuristic (Tversky & Kahneman, 1982b).[34]

More specifically, Kahneman and Tversky explained the Linda-conjunction-fallacy with process models featuring the representativeness heuristic (which we can call, the *Linda model*):

[34] Readers should note that the portrayals of research findings in this dissertation are often simplifications (in respects not relevant to the dissertation's arguments). For example, portrayals may imply that a cognitive heuristic is the complete determiner of a judgment. The more complicated reality is that often, the heuristic is not the sole process mediating a judgment, but an influential factor (*a la* one of several independent variables).

Linda Model

Explanandum: Linda-Conjunction-Fallacy (disposition)

- Input: The Linda Problem
- Output: Ranking (i.e., ranking *Linda is a feminist bank teller* as more probable than *Linda is a bank teller*)

Explanans: Linda Model

- Operations:
  - (a) *[Linda-Bank-teller] Representativeness:*
    - (a.1) Activate the prototype that best represents [the biographical description of Linda (which we can refer to as the "Linda prototype")]
    - (a.2) Activate the [Bank-teller] prototype
    - (a.3) Measure the extent to which the [Linda] prototype is representative of the [Bank-teller] prototype
    - (a.4) Assign the value of that measurement [low] to [Linda-Bank-teller]
  - *(b) [Linda-Feminist] Representativeness:*
    - (b.1) Activate the [Linda] prototype
    - (b.2) Activate the [Feminist] prototype
    - (b.3) Measure the extent to which the [Linda] prototype is representative of the [Feminist] prototype
    - (b.4) Assign the value of that measurement [high] to [Linda-Feminist]
  - *(c) [Linda-Feminist-Bank-teller] Representativeness:*
    - (c.1) Identify a compromise value [medium] that is approximately halfway between the value [low] assigned to [Linda-Bank-teller] and the value [high] assigned to [Linda-Feminist].
    - (c.2) Assign the value of that measurement [medium] to [Linda-Feminist-Bank teller]
  - *(d) Scale Conversion:*
    - (d.1) Convert the value [low] assigned to [Linda-Bank-teller] and the value [medium] assigned to [Linda-Feminist-Bank-teller] to the [Ranking] scale used in [the Linda Problem]
- Organizing Program:
  - (1) Perform (a) and (b)
  - (2) Perform (c)
  - (3) Perform (d)

As evidenced above, the Linda model contains a depiction of the cognitive system's mediating process (in terms of operations according to a program) that renders explicit how possession of the mediating process is sufficient for possession of the disposition, the Linda-conjunction-fallacy.

Recall that weak-DRCT contended that merely the core judgment system was ideally rational. As such, weak-DRCT was consistent with systematic biases that were attributable to processes and entities outside the core judgment system. In this respect, such attributions could explain away biases that, *prima facie*, might appear to falsify weak-DRCT. However, as the evidence of heuristics continued to pour in, such explaining away becomes increasingly ad hoc and strained. What might be considered the straw that finally broke weak-DRCT's back was the fall of the *cognitive miser* theory of heuristics. This last gasp of weak-DRCT conceded the prevalence of heuristics and biases but contended that humans intentionally execute heuristics due to "miserliness" about expending cognitive resources (Gilovich & Griffen, 2002, p. 4). If people employ heuristics for efficiency's sake, this is consistent with their nonetheless, possessing rational reasoning competencies. In other words, just because people often use quick and dirty methods, does not mean that they lack rational reasoning programs—they may just not bother to use them when stakes are low.

The principle rebuttal to the cognitive miser theory of heuristics was the persistence of heuristics despite the provision of financial incentives (Camerer & Hogarth, 1999; Gilovich & Griffen, 2002; Grether & Plott, 1979; Kahneman & Tversky, 1982b; Larrick, 2004; Tversky & Kahneman, 1981; Wilson, Houston, Etling & Brekke, 1996). In other words, if people still use heuristics when there is money on the line, it shows that heuristics is simply what the judgment system has at its disposal (or at least what the intuitive judgment system has at its disposal).

These findings supported Tversky and Kahneman's *natural assessments* model (Tversky & Kahneman, 2002), in which heuristics are executed automatically (*a la* the processes underlying perceptual illusions).[35] That is, heuristics are default processes that occupy the core of the judgment system (assuming this notion of a core remains viable). At least with respect to intuitive judgment, these findings arguably put the final nails in the coffin of weak-DRCT and DRCT, *simpliciter*—at least in cognitive psychology.

3. Judgments/Decisions/Actions

The following regards an important taxonomical clarification that may avert confusion. In the cognitive psychological paradigm, the primary ontological categories are informational inputs (e.g., questions like: What's 15 times 4?), cognitive processes (i.e., informational transformations, such as doubling, taking that product, and doubling again), and informational outputs (e.g., respondent answers like: 60). There are many kinds of outputs. Three such kinds are judgments, decisions, and actions. Initially, instantiations of immorality—i.e., moral errors—were identified as simply, *acts*. A more precise identification is: *acts, decisions, and/or judgments*. Some tension over how I should identify moral errors—and thus, moral biases—stems from this dissertation's incorporation of diverse subject matters and disciplines. For instance, on the one hand, the dissertation's focus on morality favors identifying moral errors as *acts*; this is because acts are often considered the paradigmatic objects of moral evaluation. On the other hand, the focus on cognitive processing favors emphasizing *judgments* and *decisions*, as they

[35] Per contemporary views, such judgments of "natural assessment" are one of the variants of "intuition" that were the theoretical predecessors of *type 1* judgment-making.

are the paradigmatic outputs in cognitive psychology. For some readers, the distinctions between acts, decisions, and judgments are very important. Nonetheless, perhaps counterintuitively, such distinctions are often beside the point for the project at hand.

To illustrate this, suppose that in a school lunchroom Owen offers Valerie a trade of his veal parmigiana for her spaghetti. A cognitive psychological theory might predict that she will form the *judgement* that her spaghetti is more valuable. An economic theory might predict that she will perform the *act* of trading away her spaghetti. Are these theories consistent? According to the theories' surface meaning, they are. This is because judgments and acts are simply different phenomena. The standard (often implicit) view regarding the relation between such outputs is a linear causal chain, wherein acts are precipitated by decisions (e.g., intentions to act), which are precipitated by judgments (which are something like beliefs, but with less exclusive conditions). Given these distinctions, claims about judgments never contradict claims about actions. Nevertheless, despite the above theories' surface consistency, such theories' implicit assumptions render the theories inconsistent. This is because such theories assume that *ceteris paribus*, judgments, decisions, and actions align—that is, their contents are in agreement. For example, *ceteris paribus*, Valarie's *judgment* that her spaghetti is more valuable aligns with a *decision* to decline the trade, which aligns with the *act* of declining the trade. There are exceptions, such as akrasia (wherein an action conflicts with something like a decision); however, these exceptions are accommodated by the inclusion of *ceteris paribus* clauses (henceforth, often left implicit). This assumed alignment is reflected in such theories' operationalizations—for instance, the cognitive psychological theory might operationalize its predicted judgment in terms of Valerie's act of declining the trade. Through the *alignment assumption*, a theory that predicts a particular judgment can be transformed into a theory that predicts the action that aligns with that judgment. The same transformability applies across acts, decisions, and judgments. As such, theories that vary in whether they refer to acts, decisions, or judgments are nonetheless, indirectly commensurable. For example, cognitive heuristics and biases (qua theory) usually refers to *judgments*; descriptive rational choice theory usually refers to choices -i.e., *decisions*- or when employed in economics, *actions*. Nevertheless, it is canonical to treat them as theoretical competitors.

The alignment assumption accommodates incorporating moral evaluation, which emphasizes acts. For example, suppose a decision theory predicts that Valarie will make the *decision* to trade for the veal parmigiana. A moral theory might entail that given the treatment of veal calves, Valerie's *act* of trading for the veal parmigiana would be morally wrong. The alignment assumption between decisions and actions allows the moral theory's entailments to be commensurable with the decision theory's prediction.

The alignment assumption also allows for explanations regarding judgments, decisions, or actions to (*ceteris paribus*) be applied to their counterparts. For example, an explanation of an action can be reformulated as an explanation of an aligned decision. It also licenses shorthand statements such as, "Some immoral acts are *explained by* moral heuristics," even if a more precise formulation might be "Some immoral acts *are explained by decisions, that are explained by judgments that are explained by* moral heuristics." Alignment allows the (*de dicto*) explanandum (i.e., that which is explained) to be slid up and down the causal chain of aligned phenomena that constitute outputs (i.e., judgments, decisions, and actions).

In all, given the alignment assumption, acts, decisions, and judgments are not only rendered *ceteris paribus* indirectly commensurable, but also *ceteris paribus* practically interchangeable (from the perspective of researchers, for instance). For example, it usually makes little difference whether a theory predicts a *judgment* that aligns with making a trade or predicts the *act* of making the trade (as operationalizations reveal).

The alignment assumption and its products, indirect commensurability and practical interchangeability, renders the distinction between acts, decisions, and judgments often beside the point for the project at hand. Further, thus far, appeals to judgments, decisions, and actions have been construed as referring to distinct phenomena. An alternative interpretation is that while such theories differ in what phenomena they mention, they are really all referring to a singular thing—namely, a broader, undifferentiated *output* that encompasses more thinly sliced phenomena, such as judgments, decisions, and actions. Carving up such undifferentiated outputs into acts, decisions, judgments, etc. would often involve a more finely grained perspective than is relevant. For example, suppose we are investigating racial disparities in sentencing. An explanandum-of-interest may be a juror's voting guilty given the trial observed (*qua* output-per-input). More fine-grained distinctions could differentiate between (a) the judgment (or assessment) of guilt (e.g., the thought, "He did it."), (b) the decision to vote guilty (e.g., the thought, "When I get my juror-ballot, I'm going to write 'guilty'"), and (c) the act of registering the guilty vote (e.g., writing the word, "guilty," on the ballot and dropping it in the hat at the center of the table). However, making such distinctions would involve more specificity than the investigation calls for; such distinctions would be beside the point. For (an extreme) example, one could separate the registering of the guilty vote into the bodily movements for each pen stroke, grasping the ballot with one's hand, locating the hat, reaching out to hold the ballot above the hat, etc. However, if the investigation concerns racial disparities in sentencing, such differentiation would be absurd. Just as one can identify *the registering of the guilty vote* as an explanandum and ignore each individual pen stroke subsumed under it, one can identify *voting guilty* as an explanandum and ignore differentiating the judgment, decision, and act subsumed.

The exception is when alignment is not the case, as with akrasia. An analogous illustration of the exception would be if such bodily movements contributed to an ultimate output that conflicted with the ultimate output *expected* (via the alignment assumption) from the preceding proximate links in the causal chain. For example, if following an affirmative decision to spank a child, raising one's hand in an aggressive pose triggered an aversion to proceeding (i.e., "caused one pause") and ultimately, yielded an output without spanking—that is, an ultimate output that conflicted with the output expected from the preceding decision to spank. It is only in such cases that we need to mention, let alone differentiate, bodily movements. In this respect, differentiating judgments, decisions, and actions is also usually unnecessary.

An additional advantage of construing the phenomena-of-interest as broader outputs is that it avoids commitments regarding how judgments, decisions, and actions relate. For instance, presuming a linear causal chain from judgment to decision to action sounds similar to presuming a linear causal chain from justificatory reasons for a moral judgment to the moral judgment. This presumption would be highly problematic given Jonathan Haidt's findings in "The Emotional Dog and its Rational Tail" (2001)—that is, the finding of justificatory reasons for moral judgments often being formulated *after* the formation of that moral judgment (i.e., rendering the reasons not explanatory reasons for the judgment, but post-hoc rationalizations). In short, construing the phenomena-of-interest as outputs side-steps this minefield.

Ideally, I could simply speak in terms of "outputs" (e.g., rephrasing the opening question as "Why do people generate immoral outputs?"). However, this would involve too large a break with conventional language. As such, within this dissertation, outputs of interest may be identified (explicitly or implicitly) as "acts," "judgments," etc. or in ambiguous terms (e.g., as with "blaming the victim"). Such identifications will often merely reflect conventional phrasing and should not be construed as implying substantive differentiation. In other words, the particular terms used are usually merely stand-ins for *output*. In this dissertation, the appropriate construal of such phenomena -including moral errors- is *as outputs* (unless explicitly stated otherwise). As such, in this dissertation, terms such as "act," "decision," "judgment," etc. can generally be treated as interchangeable.

Armed with this chapter's background information on the cognitive psychological paradigm, DRCT, and cognitive heuristics and biases, we are now in a position to develop moral analogues of cognitive heuristics and biases.

CHAPTER 3: ADAPTING ‘COGNITIVE HEURISTIC’

1. Worry: Is ‘Cognitive Heuristic’ an Adequate Basis for a Moral Analogue?

As stated (Ch. 1, §1), I am going to develop the moral analogue of cognitive heuristics by adapting ‘cognitive heuristic’ to the moral domain. This involves taking ‘cognitive heuristic’ and replacing its cognitive-domain-regarding features (or domain-general-regarding features) with moral-domain-regarding features to yield a *prima facie* concept of the moral analogue. However, as we shall see, a problem immediately arises (regarding the kind, *cognitive heuristic*).

The central meaning of ‘cognitive heuristic’ is reasoning shortcut. However, *shortcut* (*simpliciter*) is incomplete. This is because ‘shortcut’ is inherently relative (*a la* ‘taller *[than…]*’ or ‘faster *[than…]*’); it is only meaningful *vis-à-vis some relatum*. As addressed (Ch. 2, §2.3), that relatum is: the algorithm that the principles of ideal-rationality dictate for the given judgment task. Additionally, a process’ constituting a shortcut is understood in terms of the resources expended in executing that process. Such resources include time and cognitive resources, such as the number of computations required. In this respect, ‘shortcut-ness’ is understood as being “cognitively economical” (Hastie & Dawes, 2001, p. 95). Putting together the preceding, a more complete statement of *shortcut* is: requiring less resources to execute than would be spent by the algorithm that the principles of ideal rationality dictate for the given judgment task. We can summarize this as: shortcut vis-à-vis the ideally-rational algorithm (SCIRA).

SCIRA distinguishes cognitive heuristics from pertinent contrast class members (e.g., algorithms and the genus, reasoning processes, *simpliciter*). As such, SCIRA is the central property (*a la* essential property) of cognitive heuristics.[1] However, cognitive psychologists (at least those working within the Kahneman and Tversky tradition), do not investigate whether the processes that they call "cognitive heuristics" are indeed more economical than the corresponding algorithm—i.e., actually possess SCIRA. In this respect, they do not know whether those processes actually constitute cognitive heuristics. This apparent indifference to SCIRA-possession, and, in turn, constituting a cognitive heuristic, raises doubts about whether being a cognitive heuristic (or not) actually matters—i.e., it raises doubts about whether the kind, *cognitive heuristic*, matters. Especially considering the kind's centrality in the *cognitive heuristics* and biases (CHB) paradigm, it should matter. A desideratum of the pertinent mattering is being scientifically-useful.[2] If *cognitive heuristic* is not scientifically useful, then it warrants revision or elimination.[3] This threat of scientific inadequacy needs to be resolved before cognitive heuristics can serve as the basis for a moral analogue. As such, we need to

[1] There may be some fudging when jumping between the concept, 'cognitive heuristic,' and the kind, *cognitive heuristic*—i.e., jumping between linguistics and metaphysics. However, those jumps are unproblematic and rendering explicit the moves and architecture that makes this so yields an *extremely* lengthy and laborious read. Nonetheless, a worthwhile sense of such can be gleaned from the following alternative formulation. 'SCIRA' is the distinguishing feature of the meaning of 'cognitive heuristic' (more specifically, the distinguishing feature of that semantically-internal concept's intension). Cognitive psychologists' not testing for the corresponding/concordant property, SCIRA, reflects a divorce between (a) uses of the term, "cognitive heuristic," (to refer to those processes), and (b) the meaning of the concept, 'cognitive heuristic.' This divorce is not a clash between intension and extension, but a case of uses' unfaithfulness to meaning. (The possibility of such is reflected in the possibility of divorces regarding concepts that lack an intension—e.g., uses of "water" that are unfaithful to 'water.') This divorce raises doubts about conceiving of those processes per 'cognitive heuristic (per that meaning)' and raises the prospect of revising or eliminating that (descriptive) concept and in turn, its (dependent) corresponding kind.

[2] For simplicity's sake, I will often assess scientific usefulness in binary terms—i.e., a kind being scientifically useful *or not*. A more refined rendering accommodates a continuous scale of scientific usefulness (i.e., assessments of more/less). In this respect, "not scientifically useful" should be interpreted as implicitly, *insufficiently scientifically useful* (e.g., a kind being "*not* scientifically useful" by virtue of being much *less* scientifically useful than members of the kind's contrast class). In other words, while *jade* is not scientifically useless, we still might defensibly say, "*Nephrite* is scientifically useful, and *jade* is not." Such binary talk's implicitly accommodating a continuous scale also occurs with "making an explanatory contribution (*or not)*."

[3] Mentioned kinds are italicized (e.g., "Is *cognitive heuristic* a scientifically-useful kind?").

figure out whether or not *cognitive heuristic* is a scientifically-useful kind?

2. Does the Kind, *Cognitive Heuristic*, Make a Causal Explanatory Contribution?

Most objects instantiate numerous kinds. For example, a penny instantiates the following kinds: *currency*, *coin*, *copper*, *metal*, etc. Kinds vary in their scientific usefulness. For instance, in many contexts, it is scientifically useful to classify and theorize about objects in terms of their *natural kinds* (e.g., samples qua *nephrite* versus qua *jade*—Putnam, 1975).[4]

One basic way that a kind can be scientifically useful is entering into -or more precisely, *contributing* to- scientific explanations. Of particular interest in this context is the default form of explanation in cognitive psychology (Ch. 2, §1)—namely, process models, or more specifically, informational-level causal-mechanical explanations of cognitive dispositions.

Heuristics cause manifestations of cognitive dispositions (e.g., the representativeness-heuristic depicted in the Linda model causes the manifestation of the Linda-conjunction-fallacy—Ch. 2, §2.5).[5] As such, those heuristics contribute to causal-mechanical explanations of those dispositions. This may seem sufficient for the *kind*, *heuristic*, contributing to those explanations; however, it isn't.

[4] The colloquial kind, *jade*, encompasses two distinct mineralogical kinds, *jadeite* and *nephrite*. Jadeite and nephrite are superficially similar (e.g., in color and toughness); this accounts for their sharing the kind, *jade*. However, jadeite and nephrite differ in their microstructures (i.e., chemical compositions). Allowing for simplification, when all objects of a kind have the same microstructure (e.g., as all instances of *nephrite* do), this allows one to examine a single instance, discover certain properties (e.g., number of chains, or cleavage angle), and conclude that all other instances of the kind possess those properties. (Bird, 1998, p.69, n. 27) That is, the shared microstructure grounds inductive inferences that render those properties *projectable* to all instances of the kind. In this respect, *nephrite* constitutes a *natural kind*. *Jadeite* does as well. Since nephrite and jadeite have different microstructures, projectable properties of nephrite cannot be justifiably projected onto instances of jadeite, and visa-versa. In this respect, *jade* (which encompasses both *nephrite* and *jadeite*) is not a natural kind.

[5] For brevity, I'll refer to *cognitive* heuristics as simply, "heuristics" until §7.

To illustrate this, consider the following case (adapted from Dretske, 1988, p. 81).[6] Suppose you are at an opera. The plot is a murder mystery. At the end of the opera, the diva summons her highest tone and belts out the climactic line, “The butler did it!” This shatters a champagne glass. The opera goes on tour and the diva shatters numerous glasses in this way. Suppose you wanted to give a causal explanation of those glasses’ shattering. One way to identify the singings that caused the shatterings is as instantiations of the kind, *high-note*. Another way is as instantiations of *climactic-reveal*. *High-note* is an acoustic-oriented kind; *climactic-reveal* is a semantic-oriented kind. Of course, the singings caused the shatterings by virtue of their acoustic properties, not their semantic properties. As such, the singings shattered the glasses qua (instantiation of) *high-note*, and not qua *climactic-reveal*. This stems from the differing extent of correspondence between the kinds’ properties -especially their central properties- and the properties that caused the shatterings. In this respect, qua *high-note* has significant explanatory power that qua *climactic-reveal* lacks. That explanatory power is reflected in the greater understanding *qua-high-note* conveys. For instance, given knowledge that the singings caused the shatterings, there is a tremendous difference in understanding those events between (1) thinking of the singing in terms of the meaning of the words sung (as if it were a spell for shattering glasses), (2) total ignorance of the manner in which the singing bears upon the glasses’ shattering (e.g., a pre-scientific perspective), and (3) accurately (with respect to the shatterings) thinking of the singing in terms of its acoustic (i.e., physical) nature. This greater understanding is reflected in the differing conceptions’ different inductive utility (from those shatterings). In these respects, identifying the

[6] While Dretske’s original case regards properties’ causal-relevance, his analysis can be adapted for kinds’ explanatory-contributions.

singings as instantiations of *high-note* makes an explanatory contribution that the *climactic-reveal* identification does not. In this respect, the *kind*, *high-note*, makes an explanatory contribution that *climactic-reveal* does not.

Note that this does not undermine the singings' constituting climactic reveals (just as samples of nephrite indeed constitute jade). Furthermore, one can truthfully say that those climactic-reveals (i.e., those singings) caused and thus, explained those shatterings. Nevertheless, that is insufficient for the shatterings to be caused or explained by the singings *qua climactic-reveal*—i.e., it is insufficient for a (significant) explanatory contribution by the *kind*, *climactic-reveal*.[7]

Likewise, that (a) heuristics -i.e., those processes that constitute heuristics- cause and explain the manifestations of cognitive dispositions is not sufficient for (b) the *kind*, *heuristic*, making an explanatory contribution. For the contributions of those heuristics to yield a contribution by the *kind*, *heuristic*, the former contributions must be made by those processes *qua heuristics*. *A la* the contribution of qua-*high-note* (vs. qua-*climactic-reveal*), this comes down to properties. As stated (§1), the central property of *heuristic* is: shortcut vis-à-vis the ideally-rational algorithm (SCIRA). As such, a contribution by the kind, *heuristic*, requires a contribution by SCIRA.[8]

[7] While the referential utility of *climactic reveal* in those cases may constitute *some* explanatory contribution, it is no more than the explanatory contribution one's finger makes to a causal explanation when one identifies the cause by pointing at it. In this respect, that explanatory contribution is not significant and when using binary assessments for simplicity's sake, we can just say that the kind does not make an explanatory contribution.

[8] An alternative formulation of the preceding is that 'high-note' (as invoked by "high-note") makes an explanatory contribution that 'climactic-reveal' does not by virtue of the former's (greater) concordance between its intension and those properties of the referent that are causally-efficacious (or more generally, explanatorily-relevant) with respect to the explanandum. Likewise, for 'heuristic' to make an explanatory contribution, its intension -or more precisely, its intension's distinguishing feature, 'SCIRA'- must concord with explanatorily-relevant properties of the referent—i.e., the property, SCIRA, must be explanatorily-relevant to the manifestation of cognitive dispositions.

Since causal-mechanical explanation is a form of causal explanation, such an explanatory contribution requires a causal contribution to the manifestation of those dispositions. That requires that SCIRA is a causally-efficacious property with respect to those dispositions. As such, that those processes that constitute heuristics make causal and explanatory contributions to those dispositions yields an explanatory contribution by the *kind*, *heuristic*, only if SCIRA is a causally-efficacious property with respect to those dispositions. So, is it?[9]

2.1. Is SCIRA a Causally-Efficacious Property?

For a property to be causally-efficacious with respect to a system's manifestation of a disposition, that system must be sensitive to that property. System-sensitivity to a property does *not* require that the system (in some sense) appreciates that property—for instance, it does not require that the system has a concept of the property. It merely requires that the system possesses the capacity to differently interact with objects in accordance with whether those objects possess the property.

For example, consider a vending machine that is sensitive to the property "being-

[9] A couple of notes on *causal-efficaciousness*: (1) In philosophy, the most common appeals to the phrase, "causal-(in-)efficaciousness," concern the possession of (something like) any causal powers at all. For example, considerations of whether Platonic numbers, moral properties, or qualia are causally-(in-) efficacious, in this more general sense. In this dissertation, however, "causal-efficaciousness" is always appealed to *with respect to a particular (set of) explananda* (e.g., cognitive dispositions). For instance, despite *mass* being an exemplar causally-efficacious property in the general sense, a mirror's mass may be a causally-inefficacious property *with respect to the explanandum*, the mirror's capacity to reflect light. (2) Considering the weakness of theories of causal-efficaciousness (or more broadly, causal-*relevance*), I will rely upon intuition to assess properties' causal-efficaciousness. For an overview of leading theories, see McKitrick (2005) and Braun (1995). Most leading theories (e.g., counterfactual theories, exclusion theories) have significant problems (including strong counterexamples). Furthermore, (it is at least, my impression that) the theories are (rightly) given little weight versus intuitions. For instance, when the proffered theories conflict with an intuitive assessment of a property's causal-efficaciousness in a given case, few seem to side with the theory over the intuition. It is in this respect that these theories of causal-efficaciousness are weak. They can help illuminate the idea of causal-efficaciousness, but since they are seemingly rarely strong enough to overrule intuitive assessments, there is no need to specify causal-efficaciousness in terms of one of them. In other words, until such theories are strong enough to override some intuitions, you might as well just stick with the intuitions.

a-quarter." Suppose the vending machine will release both an item worth twenty-cents and a nickel if you insert a quarter; but will not do so if you insert a non-quarter (e.g., a dime). In this respect, the vending machine differentially interacts with coins in accordance with whether they possess the property, being-a-quarter. In this respect, the vending machine is sensitive to the property, being-a-quarter. As such, being-a-quarter can be a causally-efficacious-property with respect to dispositions of the vending machine. By contrast, the vending machine is insensitive to the property, minted-in-2018. This is despite the fact that coins that possess this property have "2018" embossed on them. For instance, the vending machine cannot differentially interact with coins in accordance with whether they possess the property, minted-in-2018, versus minted-in-2017. As such, minted-in-2018 *cannot* be a causally-efficacious-property with respect to dispositions of the vending machine.

Likewise, for SCIRA to be a causally-efficacious property with respect to cognitive dispositions, the cognitive system must be sensitive to SCIRA. An important contrast with heuristics is simple-yet-ideally-rational-algorithms, such as processes that instantiate disjunctive syllogism or &-elimination.[10] A system that is sensitive to SCIRA would be capable of differentially interacting with heuristics (which possess SCIRA) versus simple-yet-ideally-rational-algorithms (which do not). With this in mind, we can consider how the cognitive system might be sensitive to SCIRA.

2.2. Is the Cognitive System Directly-Sensitive to SCIRA?

The principal form of system-sensitivity to properties is (what we can refer to as) *direct*-sensitivity. Direct-sensitivity is simple, straightforward sensitivity, such as the

[10] I will often use hyphens in multiple-word terms for readability's sake.

vending machine's sensitivity to being-a-quarter. Is the cognitive system directly-sensitive to SCIRA?

Consider that a process cannot constitute a shortcut *vis-à-vis* an ideally-rational-algorithm if that process constitutes an ideally-rational-algorithm. As such, direct-sensitivity to whether a process possesses SCIRA requires ruling out the process' constituting an ideally-rational-algorithm. This requires satisfying the following disjunction: either (1) possession of the principles of ideal rationality (to evaluate the ideal rationality of a process that potentially possesses SCIRA) or (2) sensitivity to the pertinent ideally-rational-algorithm (to determine that a process-in-question is not that ideally-rational-algorithm).

Satisfaction of the first disjunct (i.e., possessing the principles of ideal-rationality) also requires distinguishing those principles from other possessed principles such that only the ideally-rational ones are employed for such evaluation. In other words, even if all of the principles are "in there," that is insufficient. Especially in light of the falsification of descriptive rational choice theory (DRCT) (Ch. 2, §2), there is not sufficient reason to believe that the cognitive system possesses and distinguishes the principles of ideal-rationality.

Let's turn to the second disjunct (i.e., sensitivity to the pertinent ideally-rational-algorithm). Considering that there are numerous heuristics, sensitivity to their SCIRA-possession via the second disjunct requires possession of the pertinent ideally-rational-algorithm for each one. This requires possessing a host of ideally-rational-algorithms (and distinguishing them from other processes, such that only the former is employed for comparison with a potential SCIRA-possessing process). In other words, this requires the cognitive system's possessing a distinguished list of ideally-rational-algorithms. Once again, especially in light of the falsification of DRCT, there is not sufficient reason to believe that the cognitive system possesses and distinguishes such a host of ideally-rational-algorithms.

As such, there is not sufficient reason to believe that the disjunction is true. Furthermore, both disjuncts are precisely the sorts of entailments of DRCT that brought down the theory (Ch. 2, §2). In a post-DRCT context, one cannot simply assume that the cognitive system satisfies this disjunction. There is not sufficient evidence to support such satisfaction. Furthermore, such satisfaction is discordant with the well-established findings of the cognitive heuristics and biases (CHB) research program (Ch. 2, §2.3-2.5). Since the cognitive system's direct-sensitivity to SCIRA requires the satisfaction of the disjunction, and we lack sufficient reason to believe the disjunction, we lack sufficient reason to believe that the cognitive system is directly-sensitive to SCIRA.

2.3. Is the Cognitive System Indirectly-Sensitive to SCIRA?

While *direct-sensitivity* is the principal form of sensitivity, it is possible for systems to be sensitive to properties *indirectly.* The simplest form of indirect-sensitivity to a property-of-interest is via direct-sensitivity to some other property that is coextensive with the property-of-interest—that is, a property that functions as a proxy for the property-of-interest. One way a proxy could be coextensive with SCIRA is by virtue of an analytical or conceptual relation between the proxy and SCIRA. At least as far as I am aware, there is no property that (1) is coextensive with SCIRA by virtue of analytical/conceptual relations, (2) is a property to which the cognitive system's sensitivity is not ruled out by CHB's refutation of DRCT, *and* (3) is a property to which the cognitive system is plausibly sensitive. As such, we can conclude that there is insufficient reason to believe that the cognitive system is indirectly-sensitive to SCIRA via direct-sensitivity to analytically/conceptually-related proxies.

Nonetheless, it is possible to have properties that are coextensive with SCIRA (other than via analytic/conceptual-relations). A major way in which this could occur involves (1) the isolation of processes that possess SCIRA in a distinct system (e.g., in an extreme case, a distinct module), and (2) the system's being *indirectly-sensitive* to SCIRA via direct-sensitivity to the proxy, emanating-from-that-system.

(2) is the easy part (if you will); (1) is the hard part. Recall that SCIRA (i.e., *shortcut-vis-à-vis*-the-ideally-rational-algorithm) entails *deviating-from*-the-ideally-rational-algorithm, which thus, excludes simple-yet-*ideally-rational*-algorithms (e.g., disjunctive syllogism). As such, the isolation of processes that possess SCIRA to a distinct system requires excluding simple-yet-ideally-rational-algorithms from that system. Considering the lack of direct-sensitivity to SCIRA, how would simple-yet-ideally-rational-algorithms be excluded?

The most plausible possibility is that while the system was insensitive to SCIRA-possession, *selection* (e.g., via evolution or reinforcement learning) was not—i.e., selection occurred per SCIRA-possession.[11] This would require something like the following scenario. At time 1 (t_1), conditions were such that evolution selected *only* heuristics and did not select *any* simple-yet-ideally-rational-algorithms. At t_2, conditions had changed such that a separate system developed that selected simple-yet-ideally-rational-algorithms and *only* processes that did not constitute heuristics. The conditions at

[11] The following is mostly presented in terms of evolution, as its easier to digest. Reinforcement learning is addressed at the end.

t_2, which yielded the non-selection of any heuristics, allowed the first system to be the *only* system containing heuristics. As such, the first system contained *only* heuristics and *all* the possessed heuristics—i.e., *only* processes that processed SCIRA and *all* the possessed processes that possessed SCIRA. As such, sensitivity to processes' emanation-from-that-system would be sufficient for *indirect*-sensitivity to SCIRA-possession (despite *direct*-insensitivity).

While this scenario is coherent, there is insufficient reason to believe it is true. This is because there is insufficient reason to believe that a suite of processes as complicated as heuristics (e.g., the representativeness heuristic, the availability heuristic, etc.) would evolve without the evolution of *any* simple-yet-ideally-rational-algorithms, such as &-elimination. For instance, there is insufficient reason to believe that a creature could execute the representativeness and availability heuristic and yet be unable to conclude for example, that since her tribemate has berries and meat, her tribemate has meat (&-elimination). In other words, there is no heuristic for &-elimination that could take the place of actually evolving &-elimination. Even if there were, there certainly is not a corresponding heuristic for *every* simple-yet-ideally-rational-algorithm whose failure to evolve was implausible given the evolution of various complicated heuristics. In addition, this scenario requires that no heuristics were acquired once t_2 was reached (or that such heuristics -and no non-heuristics- were somehow added to the first system). In sum, there is insufficient reason to believe this scenario is true.

The following is an alternative scenario that would also achieve the isolation of heuristics (despite direct-insensitivity to SCIRA). At t_1, conditions were such that evolution selected *all* now-possessed processes that do not constitute heuristics, including all now-possessed simple-yet-ideally-rational-algorithms. At t_2, conditions had changed such that a separate system developed that selected *all* now-possessed cognitive heuristics and *only* cognitive heuristics. The conditions at t_1, which yielded the selection of *all* now-possessed process that do not constitute cognitive heuristics, including *all* now-possessed simple-yet-ideally-rational-algorithms, (1) would allow the exclusion of those processes from the subsequently developed system and (2) would allow that all subsequently-acquired processes -which were confined to another system- consisted of *only* cognitive heuristics (i.e., processes that possessed SCIRA). As such, sensitivity to processes' emanating from the second system would be sufficient for *indirect*-sensitivity to SCIRA-possession (despite the lack of *direct*-sensitivity).

Again, this scenario is coherent, but there is insufficient reason to believe it is true. Ultimately, this is because the exclusion of all now-possessed heuristics from evolving prior to t_2 (and thus, being confined to the second system) is plausible only if evolution is sensitive to deviating-from-ideal-rationality—for instance, evolution rejecting heuristics prior to t_2 because at that time, evolution was *only* selecting for ideally-rational-algorithms. However, evolution selects per (things like) practical usefulness (with respect to things like survival and reproduction), not per conformity with ideal-normative standards, such as ideal-rationality. As such, *prima facie*, evolution is *insensitive* to ideal-rationality. Note that this conclusion is merely *prima facie* because despite evolution's *selecting for* practical-usefulness, the cognitive system could be *indirectly*-sensitive to ideal-rationality due to ideal-rationality being coextensive with practical-usefulness—i.e., processes might be practically useful only if they are ideally-rational.

One way in which this might occur is by definition—that is, if the ideal-rationality of a process is defined in terms of its practical usefulness. However, this would be a corruption of the notion of *ideal*-rationality, which is exemplified by conforming to the principles of logic, statistics, and probability theory. Furthermore, it would eviscerate the aforementioned distinction between *ideal*-rationality and *prescriptive*-rationality (i.e., practical or pragmatic rationality). As such, this "by definition" route to ideal-rationality and practical-usefulness being coextensive is blocked.

Could ideal-rationality and practical-usefulness be coextensive other than by-definition? No. The entire narrative surrounding cognitive heuristics is that they are practically useful *despite* their deviation from ideal-rationality. So long as the notion of *ideal*-rationality remains faithful to being exemplified by conforming to logic, statistics, and probability theory, ideal-rationality and practical usefulness are not coextensive. As such, evolution's selecting per (things like) practical usefulness and not per conformity with ideal-normative standards, such as ideal-rationality, yields sufficient reason to doubt evolution's sensitive to ideal-irrationality, and certainly, renders it sufficiently insensitive to rule out excluding all now-possessed cognitive heuristics from evolving prior to t_2, which sufficient reason to doubt that all now-possessed cognitive heuristics are confined to the second system, and in turn, that this scenario provides a plausible means by which the cognitive system could be indirectly-sensitive to SCIRA-possession.

These scenarios do not exhaust all the possibilities by which it is in-principle possible to confine all and only (possessed) heuristics to a distinct system. Nonetheless, they illustrate the inevitable pitfalls that any such scenario encounters. This includes scenarios in which process-selection is achieved entirely through reinforcement-learning (as opposed to evolution). This is because, like evolution, reinforcement-learning (e.g., per positive/negative feedback) selects per (things like) practical usefulness, not conformity to an ideal-normative standard, such as ideal-rationality. Furthermore, the scenarios in which solely reinforcement-learning would yield a distinct system to which all and only heuristics were confined (despite direct-insensitivity to SCIRA-possession) would be even more convoluted. As such, we can conclude that there is insufficient reason to believe in the confinement of all and only heuristics to a distinct system (i.e., isolation of heuristics to a distinct system). As such, there is insufficient reason to believe in the cognitive system's being indirectly-sensitive to SCIRA-possession via sensitivity to the proxy, system-of-emanation. Recall that indirect-sensitivity to SCIRA-possession via sensitivity to analytical/conceptually-related proxies is also blocked. There is insufficient reason to believe in any other means of indirect-sensitivity to SCIRA-possession. Recall that direct-sensitivity to SCIRA-possession is blocked. As such, we can conclude that there is insufficient reason to believe that the cognitive system is insensitive to SCIRA-possession. As such, SCIRA cannot be a causally-efficacious property with respect to cognitive dispositions.

Recall that the causal and explanatory contributions of (the processes that constitute) heuristics to cognitive dispositions yields an explanatory contribution by the kind, *heuristic*, *only if* SCIRA is a causally-efficacious property with respect to those dispositions (*a la* the causal and explanatory contributions of the singings that constitute climactic-reveals yields an explanatory contribution by *climactic reveal* only if semantic properties are causally-efficacious to the glasses' shattering). There is not sufficient reason to believe that SCIRA can be a causally-efficacious property with respect to such cognitive dispositions. As such, there is not sufficient reason to think that the contributions of heuristics to cognitive dispositions yields an explanatory contribution by the kind, *heuristic*. As such, we do not have sufficient reason to think that *heuristic* is scientifically-useful in this way. Recall that if *cognitive heuristic* is not scientifically-useful, then it warrants revision or elimination (which would wreak havoc upon the development of its moral analogue). As such, we have yet to fend off the threat of *heuristic* warranting revision or elimination.

3. Is *Cognitive Heuristic* an Objective Kind?

Another way a kind can be scientifically-useful is being an objective kind. An exemplar of kind objectivity is natural kinds carving nature at its objective joints. The objectivity at-hand in cognitive psychology is limited due to the domain-of-inquiry's being not nature, but a *specific-(type-of)-system* (i.e., not cognition, *simpliciter*, but the human cognitive system).[12] While the objectivity at-hand in cognitive psychology is limited, the general spirit of objectivity still applies, as we can sensibly talk about carving specific-(types-of)-systems at their objective joints.[13] Those objective joints are grounded in causality. In this respect, kinds are specific-system-objective insofar as they align with the causal operations of that system.[14] Regarding physiology, *lung* is such a kind; *white-organ* is not. In cognitive psychology, an extreme exemplar of a kind aligning with the causal operation of the system is a kind that maps onto a mental module (*a la* Fodor,

[12] This harmonizes with cognitive psychology's explaining not via "*universal* laws of cognition," but via causal-mechanical explanation (Ch. 2, §1).

[13] We might describe the limited objectivity of specific systems as "objective-ish-ness"; nonetheless, for readability's sake, I will refer to this notion as simply, "objectivity." Such objectivity harmonizes with - but does not require- a narrow view of causation.

[14] It is noteworthy that just as a given property can be causally-efficacious in one system and not another, likewise, a given kind can align with the causal operation of one system and not another. As such, the same kind can be a specific-system-objective in one system, but not another. For example, the kind, *bitter-tasting-food*, may be such with respect to the system, human-being, but not the system, catapult.

1983). More modest exemplars are kinds that map onto specific cognitive systems and/or are unified by identification with a distinct cognitive mechanism. Such kinds constitute (what we can call) *system-specific-objective-kinds* (SOK).[15] Though not presented in these terms, that constituting a SOK is way for *heuristic* to constitute a scientifically-useful kind is reflected in Gerd Gigerenzer's demand for -and Daniel Kahneman's (attempted—n. 33) acquiescence to- unifying heuristics through identification with a distinct cognitive mechanism (Gigerenzer, 1998; Kahneman & Frederick, 2002).

3.1. System-Intrinsic Objectivity

The pertinent objectivity of specific-systems -i.e., *system-intrinsic* objectivity- is vulnerable to conflation with *system-extrinsic* objectivity. To introduce this distinction, consider the following numbers: (a) 1163, (b) 1167, (c) 1171, (d) 1175, (e) 1181 and (f) 1187. A difference between (1) {a, c, f} and (2) {b, d, e} is that the numbers in the first set are prime, while the numbers in the latter set are not. Nonetheless, there is not sufficient reason to think that when you first read those numbers, your cognitive system represented or processed the numbers in the first set in a way that systematically differed from how it represented or processed the numbers in the latter set.

There are two types of objectivity at hand here. For the first type, consider that numbers that are prime (or not) are *objectively* prime (or not). As such, there is an objective distinction between the sets of numbers. This objectivity is independent of the human cognitive system. As such, it constitutes (what we can call) system-*extrinsic*

[15] In many ways, the spirit of SOK-hood resembles that of natural kind-hood; nonetheless, the former is the preferable category in this case. For one, it is more specific. What matters for our purposes is whether a kind is aligned with the causal operation of the human cognitive system; it does not matter whether that category maps onto natural kind-hood, which is sensitive to considerations that do not bear upon our concerns. SOK-hood is also preferable in its being more ecumenical, in that it avoids having to make contentious commitments regarding the criteria of natural kind-hood.

objectivity with respect to the human cognitive system. The system-extrinsic objective difference between the sets of numbers also applies to sets of *representations* of those numbers. That is, your representations of 1163, 1171, and 1187 were systematically system-extrinsic-objectively different from your representations of 1167, 1175, and 1181, in that the former are representations of prime numbers and latter are representations of non-prime numbers.

The other type of objectivity at-hand here is the objective "carving" sense realized by alignment with the causal operation of the human cognitive system. This sense regards whether your representations of 1163, 1171, and 1187's participation in the causal nexus of the human cognitive system was systematically different from that of your representations of 1167, 1175, and 1181—for instance, whether there was a systematic difference in how the numbers were processed, stored, retrieved, what mechanisms or systems they used, etc. This constitutes (what we can call) system-*intrinsic* objectivity with respect to the human cognitive system.[16] System-specific-objective-kinds are objective in this system-intrinsic sense.

To transition towards this distinction's bearing upon *heuristic*, first consider the following beliefs: (a) The capital of Peru is Lima, (b) The capital of Guinea is Bissau, (c) The capital of Cambodia is Phnom Penh, and (d) The capital of Croatia is Nicosia. Some of these beliefs are true; some are false. With respect to system-*extrinsic* objectivity, the set of beliefs that are true is objectively distinct from the set of beliefs that are false. However, with respect to the system-*intrinsic* objectivity of objective carving and causal-

[16] It might be tempting to consider such *system-intrinsic objectivity* (which could have been called *system-relative objectivity*) a form of subjectivity. In a sense, it is; however, we do not want to lose the reference to the sense of objectivity in "carving a system at its *objective* joints."

alignment, there is insufficient reason to believe that the true beliefs are systematically distinct from the false beliefs—e.g., stored in separate memory systems.[17]

Now, recall (the aforementioned species of) the availability heuristic (i.e., estimating the probability of an event by assessing the ease with which events of that type can be recalled) (Ch. 1, §1). Suppose someone provides probability estimates for numerous events, each time using the availability heuristic. Some of those events are dramatic; some are not. (*Ceteris paribus*) those estimates will instantiate (the aforementioned species of) the availability bias—i.e., systematically overestimating the probability of dramatic events (due to dramatic events' greater ease-of-recall). The overestimations are erroneous; suppose the non-dramatic events are estimated accurately. With respect to system-*extrinsic* objectivity, the overestimations are objectively distinct from the accurate estimations. Nonetheless, the distinction's system-*extrinsic* objectivity does not entail system-*intrinsic* objectivity with respect to the human cognitive system (e.g., the estimations of dramatic events being produced by a different system).

Heuristics (i.e., shortcuts vis-à-vis ideal rationality) deviate from ideal-rationality—e.g., the standards of logic, statistics, and probability theory. In contrast with the system-extrinsic (plausible) subjectivity of standards of beauty and taste, the standards of logic, statistics, and probability theory are system-extrinsic objective (or so we can stipulate). As such, heuristics can constitute an objectively distinct kind in the system-*extrinsic* sense of objectivity. Nonetheless, this does not entail that heuristics are objectively distinct in the system-*intrinsic* sense—i.e., the causal-alignment sense—i.e., per alignment with the causal operations of the human cognitive system—i.e., the

[17] If necessary, assume the beliefs have the same (immediate) causal origin (e.g., reading a young student's error-riddled homework).

"carving" sense (i.e., the objective distinction in the system-*extrinsic* sense does not entail that cognitive heuristics are for instance, housed in a distinct module or executed by a separate system from those executing otherwise similar processes). The system-*intrinsic* sense of objectivity is the sense of interest with specific-system-objective-kinds (SOK).

So, since constituting a SOK is a way for *heuristic* to constitute a scientifically-useful kind, is there sufficient reason to think that *heuristic* is a SOK? No. As established regarding the human cognitive system's *indirect* sensitivity to SCIRA (§2.3), there is insufficient reason to believe that all and only cognitive heuristics are isolated in a distinct system. As such, given insensitivity (including direct-insensitivity—§2.2) to SCIRA-possession (i.e., the central property of *heuristic*), there is insufficient reason to believe that the kind, *heuristic*, aligns with the causal operation of the cognitive system (e.g., that the distinction between cognitive heuristics and simple-yet-ideally-rational-algorithms, such as &-elimination, aligns with the causal operation of the cognitive system).[18] As such, this way of the kind, *heuristic*, being scientifically-useful also does not bear fruit and the threat of revising/eliminating *heuristic* remains.

4. Does the Kind, *Cognitive Heuristic*, Contribute to Why-Have Explaining?

An emerging theme here is that upon scrutiny, heuristics may be more problematic than they seem. Another episode of this regards the presentation of being-a-shortcut as the *function* or *purpose* of heuristics. This sounds right, but upon scrutiny -and

[18] *Heuristic*'s not constituting a SOK is rooted in the cognitive system's insensitivity to SCIRA, which in turn, is rooted in the causal-inefficacy of SCIRA. *Heuristic*'s not constituting a SOK is merely *rooted in* (vs. *entailed by*) SCIRA's causal-inefficacy due to the possibility of sensitivity despite causal-inefficacy via *indirect*-sensitivity. However, this possibility only yielded unpromising and convoluted paths to SOK. In sum, a lot ripples out from the causal-inefficacy of a kind's central and distinguishing property.

an uncharitable interpretation- this is *telos*-talk (if you will). A more charitable interpretation is that it is a nod towards the more refined versions of such talk, such as that being-a-shortcut explains the presence of *heuristics*—that is, explains *why* we *have* those processes that constitute heuristics—that is, *why-have explaining* of *heuristics.* More specifically, the thought might be that we have those processes that constitute heuristics because they were selected for because they constitute shortcuts—e.g., selected for via evolution (e.g., Gilovich & Griffen, 2002, p. 10-11) and/or reinforcement-learning (e.g., Rieskamp & Otto, 2006). This explanation (or at least, the proffering of its plausibility) was especially valuable in the context of cognitive heuristics and biases' (qua theory or model's) arrival as a theoretical competitor of descriptive rational choice theory (DRCT) (Ch. 2, §2.1-2.2). Especially given the presumption in favor of human rationality that surrounded DRCT's dominance, being-a-shortcut provided a plausible explanation for why we would have cognitive heuristics *instead of* ideally-rational algorithms. Recall that (a) the threat of revision/elimination looms for *heuristic* pending the establishment of its scientific-usefulness (§1), (b) one basic way that *heuristic* can be scientifically-useful is by contributing to an explanation, and (c) such a contribution by the kind, *heuristic*, requires a contribution by SCIRA (§2). So, does SCIRA make an explanatory contribution to why-have explaining of heuristics? More specifically, were heuristics selected for because they possess SCIRA?

4.1. Selection per SCIRA via the But-For Condition?

To illustrate what such selection might look like, let's examine a simple, fictional scenario in which heuristics would straightforwardly be selected for because they possess

SCIRA.[19] Suppose that at one point in our evolutionary history, the human cognitive system consisted of only ideally-rational-algorithms. While such processes are guaranteed to yield correct judgments (or at least, valid or justified conclusions), their execution can require extensive time and cognitive resources. Sometimes, fast reasoning that is *usually close* to correct can be more fitness-enhancing than slower reasoning that is correct. For example, it is better to quickly process the approximate trajectory of an approaching predator and react sooner than it is to assess the trajectory with perfect accuracy but do so slowly and react too late. Suppose that per such, we evolved cognitive heuristics that replaced ideally-rational-algorithms. Those heuristics would have been selected for because they constituted shortcuts *vis-à-vis the ideally-rational-algorithms* that they *replaced*. In this respect, we would have evolved cognitive heuristics *because* they possessed the property, SCIRA—that is, the presence of those processes in the human cognitive system today would be due to their possession of the property, SCIRA—that is, SCIRA would be *why* we *have* those cognitive heuristics. As such, ‘SCIRA’ and thus, *heuristic* would make an explanatory contribution to why-have explaining of the presence of heuristics.

Of course, the preceding scenario is fictional—there is not reason to think that the human cognitive system ever consisted of only ideally-rational-algorithms. Nevertheless, the scenario will help us wrap our heads around how SCIRA could so contribute (and provides an anchor case -if you will- from which we can devise more plausible variations). In particular, the scenario provides a clear illustration of an important aspect

[19] Discussion of the selection of heuristics will often be in terms of evolutionary selection (as opposed to for instance, selection via reinforcement-learning). This is because it is easier to explicate the upcoming topics in these terms. After drawing conclusions regarding evolutionary selection, I will then address whether those conclusions generalize to all selection (including via reinforcement-learning).

of SCIRA so contributing—namely, the role played by the appeal to ideally-rational-algorithms. Like with sensitivity to SCIRA (§2), for an explanatory contribution by SCIRA -i.e., shortcut-*vis-à-vis-the-ideally-rational-algorithm*- a contribution from merely the *shortcut*-part (if you will) is insufficient; a contribution from the *vis-à-vis-the-ideally-rational-algorithm*-part is required.

The key feature of the preceding scenario is that *but for* heuristics' possession of SCIRA, we would possess ideally-rational-algorithms. As such, our possession of cognitive heuristics would not merely be attributable to their being cognitively economical (i.e., the *shortcut*-part), it would be attributable to their being more cognitively economical *than* (i.e., vis-à-vis) *the ideally-rational-algorithms they replaced*. In this respect, the *vis-à-vis-the-ideally-rational-algorithm*-part would be contributing to explaining the presence of those cognitive heuristics. As such, so would SCIRA and in turn, *heuristic*.

In the above scenario, by previously possessing only ideally-rational-algorithms and then replacing them with heuristics, the *but-for-condition* clearly obtains (i.e., *but for* heuristics' possession of SCIRA, we would possess ideally-rational-algorithms). However, we do not need a scenario this outlandish to secure the but-for-condition. Even if we never possessed only ideally-rational-algorithms, if we merely *would have* possessed them *but for* heuristics' possession of SCIRA, that is sufficient for satisfying the but-for-condition (and an explanatory contribution being made from the *vis-à-vis-the-ideally-rational-algorithm*-part).

However, this is still a tall order. The realistic plausibility of our having otherwise evolved ideally-rational-algorithms is not very promising. In this respect, an explanatory contribution of *heuristic* via satisfaction of the but-for-condition is not very promising. Nonetheless, it segues into a something with greater potential.

4.2. Selection per SCIRA via SCIRA-Score?

Suppose ideally-rational-algorithms (or more generally, ideal-rationality) constituted a standard vis-à-vis which processes became more selectable as they approached. SCIRA could capture this by being construed in (something like) the following way: a process constitutes a shortcut-vis-à-vis-the-ideally-rational-algorithm to the extent that the process has a high ratio of (a) the outputs' closeness to the output that the ideally-rational-algorithm would produce, over (b) the resources (including time) expended to execute the process. We can call this ratio the *SCIRA-score*. The higher the SCIRA-score, the more a process constitutes a heuristic.[20] If processes are selected for per their SCIRA-score, this would be sufficient for explanatory work by the *vis-à-vis-the-ideally-rational-algorithm*-part of SCIRA.

So, are processes selected for per their SCIRA-score? Let's stick with the evolutionary context for now to make things easier. Firstly, a correlation between SCIRA-score and selection is *insufficient* for processes to be *selected for* per their SCIRA-score. In other words, correlating with something that matters for selection is not the same as mattering for selection. Adapting a canonical example (e.g., Wright, 1973, p. 141), the heart both pumps blood and makes "thump-thump" sounds. Making-"thump-thump"-sounds correlates with pumping-blood. We can even suppose that the things that pump blood are coextensive with the things that make "thump-thump" sounds.

[20] Constituting a heuristic may be an all-or-nothing condition -and not a continuous matter- so there might need to be threshold conditions, etc. Nonetheless, this complexity does not affect the argument; so, we can set it aside.

Nevertheless, the heart was not selected for *because* it makes “thump-thump” sounds. As such, making “thump-thump” sounds does not explain *why* we *have* hearts; making-“thump-thump”-sounds does not make a why-have explanatory contribution.

With this in mind, we can return to the question: Are processes selected for per their SCIRA-score? There is strong reason to doubt that they are. As broached in §2.3, processes are selected for because of (something like) practical usefulness. As such, *prima facie*, they are not selected for because of their (extent of) conformity to an ideal-normative-standard, such as ideal-rationality. However, this conclusion is merely *prima facie* because processes might be practically useful *due to* their being ideally-rational.

As addressed (§2.3), one way in which this might occur is by definition—that is, if the ideal-rationality of a process is defined in terms of its practical usefulness. However, recall that this would be a corruption of the notion of *ideal*-rationality, which is exemplified by conforming to the principles of logic, statistics, and probability theory.

Another way in which processes’ practical usefulness could be due to their SCIRA-score would be if SCIRA-score was constitutive of practical usefulness. There is strong reason to doubt this. To illustrate this, consider a scenario in which one hears a rustling in the bushes that could be from either a tiger or a rabbit. Ideal-rationality dictates that one’s (credence in) beliefs should be proportionate to the evidence. However, it is practically useful to almost always believe that it is a tiger. This is because believing a tiger is present when it is merely a rabbit has a trivial practical cost (e.g., running away for no reason). However, believing that a rabbit is present when it is a tiger has tremendous practical cost (e.g., your life). That is, regarding the presence of tigers, false-positives have trivial costs, while false negatives have tremendous costs.[21] Whenever this dynamic applies (e.g., as with the lethality of unfamiliar berries), overwhelmingly erring in favor of false-positives will be more practically useful than proportioning one's beliefs to the evidence. As such, SCIRA-score is not constitutive of practical usefulness; SCIRA-score presumably correlates with practical usefulness, but this is insufficient for constitutive-ness. In short, the link between practical usefulness and rationality regards *prescriptive*-rationality, not *ideal*-rationality.

Up until now, we have been discussing selecting-per-SCIRA in terms of evolutionary scenarios. However, another means of selecting for SCIRA (and thus, SCIRA making a why-have explanatory contribution) is selection through reinforcement-learning. While selection in evolution is governed by fitness (i.e., survival and reproduction), selection in reinforcement-learning is governed by (something like) positive/negative feedback. Nonetheless, the same dynamics of practical usefulness apply. In short, conformity with an ideal normative standard, such as ideal-rationality, just is not constitutive of practical usefulness—neither practical usefulness per survival and reproduction, nor practical usefulness per positive/negative feedback.

Even if one is not fully convinced by the preceding, recall that the question of this section regards whether an explanatory contribution by *heuristic* -i.e., SCIRA- can be *established*. The preceding is sufficient to block the *establishment* of *heuristic* -i.e., SCIRA- making an explanatory contribution to why-have explaining of the presence of (the processes that constitute) cognitive heuristics, and in turn, block the *establishment* of *heuristic* being scientifically-useful in this respect. As such, *heuristic* has still yet to fend

[21] My thanks to Sara Worley for this point.

off revision/elimination.

5. Does the Kind, *Cognitive Heuristic*, Contribute to Explaining Cognitive Biases?

Suppose that *heuristic* was deemed not scientifically-useful and warranted elimination. This would prompt the question: "What about cognitive biases? Surely, they are scientifically-useful, right?" Are they? Recall that the central meaning of *cognitive bias* is systematic error. Within the CHB paradigm, judgments are erroneous by virtue of possessing the property, deviating-from-the-ideally-rational-judgment. (Ch. 2, §2.3) Recall that there is insufficient reason to believe that the human cognitive system is sensitive to SCIRA (i.e., shortcut-vis-à-vis-the-*ideally-rational*-algorithm).[22] This reflects an insensitivity to ideal-rationality, in general. (§2) As such, the system is insensitive to whether a judgment it produces conforms to or deviates from ideal-rationality. As such, a kind that carves up judgments per their deviation from the ideally-rational judgment does not align with the causal operation of the system.[23] In this respect, *cognitive bias* is not a system-specific-objective-kind (SOK) (with respect to the human cognitive system).

Nonetheless, regardless of whether *cognitive bias* is a SOK, cognitive biases are still phenomena of interest—that is, we care about them. Recall that cognitive biases are systematic errors. We care about errors because we want to avoid them—we disvalue them. We care about *systematic* error because such systematicity (a) makes errors predictable, (b) renders them conducive to ameliorative measures (e.g., de-biasing), and (c) suggests the presence of a coherent problematic cognitive process that explains the

[22] In the interest of avoiding unwieldy repetition of "*lack of sufficient reason to believe that* [P is the case]," this phrase will be replaced with formulations of "P is not the case," though the former should be considered implied.

[23] *Alignment with the causal operation of the human cognitive system* is a descriptor that most aptly applies to cognitive processes. Applying it to judgments involves applying it to the *products* of such processes. This is slightly awkward, but ultimately, still apt. It could be argued that a different descriptor should be used, but I will stick with it for simplicity's sake.

errors. Just because a kind we care about is not objective (e.g., neither a natural kind nor a SOK), that does not mean we should refrain from utilizing science to investigate it. For instance, worries have been raised about whether some -or even all- medical diseases and psychiatric disorders are non-objective kinds (or something of that sort) (Murphy, 2015, §2; Perring, 2010, §3). Nonetheless, this does not (and should not) threaten the practice of scientific medicine and psychiatry. In other words, *whether* we use science to address clef pallets and anti-social-personality-disorder does not hinge upon their ontological status, though it does affect how we should undertake such science—for instance, in kind-selection, the weight we should assign (if any) to kind-objectivity (e.g., constituting a SOK) or more precisely, causation-oriented kind unification (e.g., unification via identification with a distinct cognitive mechanism). We often appropriately want scientific investigation to recognize distinctions, categories, and kinds that are grounded not in objectivity (e.g., causation), but in our cares and interests. Such kinds are (what we can call) *practical kinds.*

Per the CHB paradigm, judgments are erroneous by virtue of deviating from the-ideally-rational-judgment. Our caring about such errors reflects our caring about the *standard*, ideal-rationality. Caring about a standard is sufficient grounds for conceiving of objects/phenomena in terms of their relation to -or more pertinently, their *contrast* with- that standard, and appealing to or even creating categories and kinds based upon those construals. As such, caring about ideal-rationality is sufficient grounds for conceiving of judgments in terms of whether they are errors vis-à-vis ideally-rational judgments, and when such errors are systematic, conceiving of them in terms of the kind, *cognitive bias*. Such grounding of *cognitive bias* in our cares and interests renders it an adequate practical kind. *Heuristic* can thus be a scientifically-useful kind by contributing to explaining instantiations of this practical kind. So, does it?

Recall (§2) that while heuristics cause and thus, explain cognitive dispositions, this was insufficient to yield a contribution by the *kind*, *heuristic*, due to the causal-inefficaciousness of SCIRA (*a la* the non-contribution of *climactic-reveal* due to -with respect to the glasses' shatterings- the causal-inefficaciousness of its central properties). In this respect, SCIRA does not make a *causal*-explanatory contribution. So, how could SCIRA make a contribution to explaining cognitive biases? By making a *non-causal* explanatory contribution—more specifically, by making a *contrastive* explanatory contribution.[24]

The central property of *cognitive bias* is systematic error. Such erroneousness is per *deviating from the ideally-rational* judgment—or more generally, deviation from ideal rationality. In this respect, *cognitive bias* is defined *relative* to ideal-rationality—that is, in *contrast* with ideal rationality. This renders *cognitive bias* (what we can call) a *contrastive kind*. Contrastive kinds are defined by objects/phenomena's contrast with something else, such as an abstract standard (e.g., ideal-rationality) or a specified object (e.g., the ideally-rational-judgment).

Contrastive kinds are not novel (though they might not be conceived of in these terms). For instance, in clinical psychology, various disorders and pathologies are conceived of in terms of their contrast with (i.e., deviation from) the cared about standard (qua *contrastum*) of healthy mental and behavioral life—the standard often expressed as

[24] Such non-causal-ness of contrastive explaining may depend upon a narrow view of causation. Verbal disputes regarding 'causation' aside, there is a substantive difference between ("non-causal") contrastive and (narrow) causal explaining, as shall be shown.

well- or *normal-functioning*. And this is how it should be. We would be much worse off if we did away with scientific inquiry into contrastive kinds.

Often, the root of contrastive kinds is viewing objects and phenomena through the lens of what we care about. This allows for causally-unaligned contrasta to enter the picture. Note that once DRCT is refuted, ideal-rationality only enters the picture via its being something we care about. In this respect, there is a harmony between practical, contrastive, and non-SOK kinds. Often, a crucial part of yielding causally-unaligned contrastive kinds is not merely *selecting* objects in accordance with what we care about but *conceiving of them* in terms of what we care about.

To illustrate this distinction, suppose we care about things that are the tallest or biggest on Earth (perhaps we are a fan of world records). Also, suppose that thousands of years ago, the tallest mountain on Earth was vaporized by a functionary from the Alpha Centauri Zoning Commission. Before that vaporizing, Mount Everest did not possess the property, being-the-tallest-mountain-on-Earth. Upon such vaporizing, it did.

Consider the difference between the following explananda: (1a) How Mt. Everest came to be, versus (2a) How Mt. Everest came to be *the tallest mountain on Earth*. Explanandum 1a regards how an object came into existence. Explanandum 2a regards how an object came to possess a property. Explanandum 1a involves our cares *motivating the selection* of an object-of-explanation; explanandum 2a involves *conceiving of* that object *in terms of* our cares. Explanandum 1a can be sufficiently explained by appeals to plate-tectonics (or something of that sort). However, that would be insufficient to explain explanandum 2a, since you cannot explain Mt. Everest's being the tallest mountain on Earth without mentioning the Alpha Centauri bureaucrat vaporizing the once-taller mountain.

To see how this distinction can get progressively subtler, let's now replace the name, "Mt. Everest," with "the mountain that is the tallest mountain on Earth." As such, we will consider the difference between explaining (1b) How the mountain that is the tallest mountain on Earth *came to be*, versus (2b) How the mountain that is the tallest mountain on Earth *came to be the tallest mountain on Earth*. Regarding explanandum 1a, the property, being-the-tallest-mountain-on-Earth, merely identifies the object; this is reflected in the fact that the parallel explanandum that used the name, "Mt. Everest," did not mention "the tallest mountain on Earth." Regarding explanandum 2b, the property, being-the-tallest-mountain-on-Earth, both identifies the object and is that whose coming to be (i.e., whose coming to be possessed by the object) is to be explained.

Let's now consider cognitive psychologists explaining errors. Ideally, we might want the explanandum, "errors," more precisely articulated; however, scientific practice is not always so accommodating. For instance, a cognitive psychologist might look at an erroneous judgment and simply say, "Now what explains that?" without further clarifying the *that* to be explained. Here are two possible explananda: (1c) How the response that is erroneous *came to be* and (2c) How the response that is erroneous *came to be erroneous*.

Let's add a concrete example. Suppose Mathew took a math quiz. It included the following question: 2 + 1 * 4 = ? Matthew gave the answer, 12. The correct answer is 6. What happened? Mathew forgot to follow the order-of-operations rule (i.e., multiplication before addition). Mathew should have first, turned "1 * 4" into "4" and then added it to 2, to get 6—i.e.:

2 + 1 * 4 = ?

2 + ~~1 * 4~~ 4 = ?

2 + 4 = 6

Instead, he did the operations from left-to-right: first turning “2 + 1” into “3” and then multiplying that result by 4, to get 12—i.e.:

2 + 1 * 4 = ?

~~2 + 1~~ 3 * 4 = ?

3 * 4 = 12

Explanandum 1c is: How the response that is erroneous came to be—that is, the object-of-explanation is simply, “12” (and how it came to be). This can be answered by providing a solely causal-mechanical explanation that depicts the transformations that Mathew performed to turn the input, 2 + 1 * 4 = ?, into the output, 12. This would be a sufficient explanation without any appeal to the correct order-of-operations. In this scenario, the cognitive psychologists’ caring about errors motivated them to direct their attention towards not just any response, but upon a response that was erroneous. In this respect, the property, constituting-an-error, was used to identify the explanandum. This is one way in which what one cares about can influence research. If the cognitive psychologist goes on to explain all the answers that Mathew got wrong in this way, she is selecting the kind, *Mathew’s errors*. This is one way of viewing kinds through the lens of what she cares about. However, this way does not yield contrastive explaining.

To be conducive to contrastive explaining, she needs to not merely *select* a set of objects in accordance with her cares, but to *conceive of* them per her cares. She needs to explain not explanandum 1c (i.e. How the response that is erroneous came to be), but explanandum 2c: How the response that is erroneous *came to be erroneous*.[25] This yields the object-of-explanation' being: 12-*instead-of-6* (and how it came to be). That is, the explanandum includes an appeal to the correct answer, 6. The explanation that was sufficient for 1c (i.e., a solely causal-mechanical explanation that depicts the transformations that Mathew performed to turn the input, 2 + 1 * 4 = ?, into the output, 12) would be insufficient for 2c. A sufficient explanation for 2c requires conceiving of Mathew's performing the operations as: left-to-right-*instead-of-per-the-correct-order-of-operations*. Construing the object-of-explanation as 12-*instead-of-6* (and how it came to be) is conceiving of it in a contrastive way—i.e., a *contrastive construal*. Its sufficient explanation involves an appeal to a *contrastive construal* of the operations that Mathew performed—that is, an appeal to left-to-right-*instead-of-per-the-correct-order-of-operations*. That is, the erroneousness of Matthew's response is explained by the erroneousness of the process he executed. In this respect, the explanation is contrastive (or includes a contrastive component). In this respect, left-to-right-*instead-of-per-the-correct-order-of-operations* makes a contrastive explanatory contribution to the explanation of 12-*instead-of-6.*

Returning to cognitive heuristics and biases, conceiving of judgments as instantiations of *cognitive bias* is to contrastively construe them as judgments-made-

[25] This is somewhat ambiguous. The pertinent sense of "came to be erroneous" regards how Mathew got the answer wrong—i.e., the mistake Mathew made. It does not regard how mathematicians established the order-of-operations rule or anything like that.

instead-of-ideally-rational-judgments—i.e., as judgments *qua deviations-from-ideal-rationality*. The central property of *heuristic*, SCIRA (*shortcut-vis-à-vis*-the-ideally-rational-algorithm), entails *deviation-from*-ideal-rationality (this entailment will be important later). In this respect, conceiving of a process as an instantiation of *cognitive heuristic* (*insofar as that entails deviating-from-ideal-rationality*) is to contrastively construe it as a-process-executed-instead-of-the-ideally-rational-process or a process *qua deviation-from-ideal-rationality*. Appealing to this makes a contrastive explanatory contribution to the explanation of judgments *qua deviations-from-ideal-rationality*. In other words, such judgments' deviation from ideal-rationality is explained by their generating process' deviation from ideal-rationality. For example, regarding the aforementioned availability heuristic and bias, the bias' deviation-from-ideal-rationality of systematically overestimating the probability of dramatic events is explained by the heuristic's deviation-from-ideal-rationality of inferring probability by assessing ease-of-recall (coupled with the fact that dramatic events are disproportionately easy to recall).[26] In such respects, appealing to a process *qua cognitive heuristic* makes a contrastive explanatory contribution to the explanation of judgments *qua cognitive bias*. That is, appealing to the instantiating of *cognitive heuristic* makes a contrastive explanatory contribution to the explanation of cognitive biases. In this respect, the kind, *cognitive heuristic* (insofar as it entails deviating from ideal-rationality), makes a contrastive explanatory contribution to the explanation of cognitive biases (*a la* the kind, *high-note*, but not the kind, *climactic-reveal*, making a causal explanatory contribution to the

[26] Note that the heuristic's embedded contrastum, ideal-rationality, played an explanatory role without (needing to) play any causal role at all. In fact, the contrastum's entire absence from the system would not threaten its explanatory contribution. This reflects the explanatory contribution's contrastive dynamic, as opposed to a (narrow) causal dynamic.

explanation of the glasses' shatterings). As such, *cognitive heuristic* is a scientifically-useful kind, *but only insofar as it entails deviating-from-ideal-rationality*. Such scientific-usefulness manifests in the sense it makes to group processes that constitute cognitive heuristics together as explainers of cognitive biases.[27]

[27] Some big-picture points. Since *cognitive bias* is an adequate practical kind and since instances of *heuristic* (qua *heuristic*) contributes to explaining cognitive biases (qua *cognitive bias*), *heuristic* is a scientifically-useful kind on explanatory grounds. In other words, *heuristic* is thereby an adequate *explanatory* kind. In this respect, it is scientifically-useful to group instances of *heuristic* together—that is, it is scientifically-useful to employ the kind, *heuristic*. Constituting an *explanatory* kind with respect to a *practical* kind is insufficient to render *heuristic* an adequate *practical* kind, *per se*; however, it does ground *heuristic*'s adequacy in practicality. CHB's featuring practically-grounded kinds renders it an *applied* scientific research program, in the sense in which clinical psychology (vis-à-vis for instance, investigating linguistic processing) is an applied science (vis-à-vis a basic or "pure" science). As an adequate explanatory kind, *heuristic*, need not achieve adequacy through constituting a practical kind. Likewise, *heuristic* need not be an objective kind—e.g., a SOK or a kind unified via identification with a distinct cognitive mechanism. Furthermore, *heuristic* should not have been expected to be an objective kind. This stems from ideal-rationality's being a *normative* standard. Insofar as it is unsurprising that we (and our competences) do not conform to normative standards, it is unsurprising that a kind conceived of in terms of its contrast with a normative standard would have a causally-unaligned contrastum and would not be an objective kind. Constituting an objective kind -e.g., a kind unified through identification with a distinct causal mechanism- is not a (mandatory) desideratum for explanatory kinds in an applied scientific research program. As such, Gigerenzer's (1998) demand and Kahneman's (Kahneman & Frederick, 2002) acquiescence to unifying *heuristic* through identification with a distinct causal mechanism was misguided (or it was misguided to present cognitive heuristics in a manner that rendered such unification apparently apt). Kahneman's exiling anchoring-and-adjustment from the extension of *heuristic* to accommodate such unification was also misguided. Why was *cognitive heuristic* presumed to be a kind for which unification on a causal basis was mandatory? This goes back to DRCT. DRCT contended that (core) human processes were ideally-rational. Such processes befit constituting a SOK or otherwise causally-unified kinds. In this respect, they befit (narrow) causal explaining. So, DRCT befits causal explaining. Since heuristics deviated from ideal-rationality (per SCIRA), they falsified DRCT. In this respect, conceiving of such processes qua *heuristic* was useful -i.e., *heuristic* was a useful kind regarding causal explaining- *but only insofar as DRCT remained a live theory to be falsified.* Once DRCT was rejected, constituting a falsifier of DRCT is no longer a basis of causal-explaining usefulness. At that time, conceiving of those processes qua *heuristic* should have ended in causal-explaining contexts (regardless of whether those processes kept the name, "heuristic"—*a la* referring to the causes of the glasses' shatterings as "climactic-reveals" -or better yet "*the* climactic-reveal"- without presuming that semantic properties or *climactic-reveal* mattered regarding the shattering). However, this did not happen. Instead of just falsifying DRCT, CHB replaced it as the leading positive, descriptive theory—i.e., the leading causal-explaining-theory. The deviation from ideal-rationality in *heuristic* (which made it a falsifier of DRCT) constitutes an appeal to ideal-rationality (even though it is a negative appeal). This embeds the normative standard within *heuristic* (or more precisely, 'heuristic'). In other words, ideal-rationality remained a variable in *heuristic* (thus the exclusion of simple-yet-ideally-rational-algorithms like disjunctive syllogism). As such, CHB -with its embedded normative standard- was employed in causal-explaining. Ironically, CHB, whose thrust was demonstrating how we fundamentally deviated from ideal-rationality (*contra* DRCT), ended up carrying on DRCT's embedding of ideal-rationality in causal-explaining. That is, CHB carried on DRCT's original sin of embedding a normative standard in a causal theory. The way in which this might have happened is the following. Use of the term, "heuristic" became engrained. Its explanatory shortcomings went unnoticed because even direct consideration of whether heuristics are explanatory

6. Conjuncts of SCIRA

Recall that the explanatory-contribution of SCIRA (*shortcut vis-à-vis* the ideally rational algorithm) and thus, *heuristic*, to contrastively explaining biases was only insofar as SCIRA entails deviating-from-the-ideally-rational-algorithm (DEVIRA). While SCIRA entails DEVIRA, that is not all there is to SCIRA (and in turn, *heuristic*). This is reflected in SCIRA's being a *prima facie* decent candidate for why-have explaining of *heuristics* (i.e., the processes constituting heuristics being selected for because of their possession of SCIRA). Yet DEVIRA cannot account for this, as *deviating* from the ideally-rational algorithm is not something to be selected for—if anything it is a cost. As this reflects, SCIRA includes (being)-more-economical-than-the-ideally-rational-algorithm (MEIRA). This accounts for SCIRA being a good candidate for why-have explaining. It also captures what cognitive psychologists were neglecting when not testing for SCIRA (§1).

Another aspect of SCIRA (or at least, 'heuristic,' as employed in the CHB program) is *natural-assessment-process*. Such processes occupy a middle ground between perception and deliberation; the notion more-or-less amounts to *intuitive*-process (or upon the rise of dual-processing models, *system-1-process*[28]) (e.g., Kahneman & Tversky, 1982b, p. 494, p. 499; Keren & Teigen, 2004, p. 93; Tversky & Kahneman, 2002, p. 20). With the replacement of DRCT with CHB (qua theory), ideally-rational-

only reveals the problematic construal if one is sensitive to the subtle distinctions of (1) referential-adequacy versus sense-adequacy (e.g., heuristics explain, but qua-*heuristic* does not) and (2) contrastive versus (narrow) causal explaining. The shortcoming only concerns sense-adequacy regarding causal-explaining, which is why unifying *heuristic* (which regards sense-adequacy) via identification with a distinct cognitive mechanism (which is a causal-explaining matter) was such a sore spot (as *heuristic* is not a SOK).

[28] The *system* 1 (intuitive system) vs. *system* 2 (deliberative system) distinction of dual-process models has since been superseded by *type* 1 and *type* 2 *processes* (Evans & Stanovich, 2013). Kahneman (2013) now considers those systems to be useful fictions.

algorithms were associated with deliberative thinking. SCIRA (or being-a-shortcut) was meant not only with regard to each heuristic relative to its corresponding ideally-rational-algorithm; it also regarded heuristics as a class relative to the class with which ideally-rational-algorithms were associated—namely, deliberative processes. In this respect, SCIRA implied that heuristics were intuitive processes (or more precisely, were processes that originated as intuitive processes). In other words, granted, 'SCIRA' -or 'fast' and 'frugal'- are relative concepts—e.g., fast-and-frugal-*vis-à-vis-ideally-rational-algorithm*; nonetheless, there is also a non-relative aspect to "fast and frugal" insofar as they imply membership amongst *the* "fast and frugal processes" (i.e., intuitive processes) or being of the "fast and frugal system" (i.e., the intuitive system).[29]

As such, SCIRA (or at least, 'heuristic,' as employed in the CHB program) has three aspects (or constituents): DEVIRA, MEIRA, and intuitive-process. These 3 aspects befit 3 different types of explanation. DEVIRA befits contrastively explaining cognitive biases (§5), MEIRA befits why-have explaining (why we have those processes that constitute cognitive heuristics) (§4), and intuitive-process befits causal-mechanical explanation (§2-3). Especially if an intuitive system is presumed, the pertinent feature of intuitive-process is that as a kind, it is good candidate for a system-specific-objective-kind (SOK) and thus, befits participation in causal-mechanical explanations. This notion of *heuristic* as a prospective SOK is consistent with Gigerenzer's (1998) demand to identify cognitive heuristics with a distinct cognitive mechanism, and Kahneman's

[29] One could argue that being-an-intuitive-process was not embedded in (the employed meaning of) SCIRA but was a separate central property (or feature) of 'heuristic' or even, was a theoretical commitment entirely separate from the meaning of 'heuristic.' I think it was embedded in SCIRA, but even if it were not, it ultimately does not affect the conclusions that follow—though it makes getting there a lot more laborious. Ultimately, resolving this issue would take us far afield into the analytic/synthetic distinction, holism, and notions (e.g., essential properties) to which the CHB program was not sensitive.

acceptance of this demand as appropriate and (attempted) acquiescence to it (Kahneman & Frederick, 2002).[30]

When appealing to *heuristic*, all three aspects are present. In this respect, the three aspects are not *disjuncts* to select amongst depending upon the context; they are *conjuncts* that are always present. We can emphasize this dynamic by identifying the aspects as "conjuncts."[31]

Recall that the only explanatory contribution found from SCIRA was contributing to contrastively explaining cognitive biases insofar as SCIRA entailed DEVIRA. In this respect, DEVIRA contributes to explaining cognitive biases and thus, is scientifically-useful. As such, any revision to *heuristic* should retain DEVIRA.

Does DEVIRA contribute to the other discussed ways of being scientifically-useful? [32] Recall (§2) that the kind, *heuristic* -i.e., such processes qua *heuristic* (*a la* the singing qua *climactic-reveal*)- did not make a causal-explanatory contribution due to the central property, SCIRA, being causally-inefficacious. That was so because of the human cognitive system's insensitivity to SCIRA. That stemmed from such sensitivity requiring implausible DRCT-style capacities in order to differentially interact with processes depending upon their possession of SCIRA. The need for such capacities stems from

[30] This acceding yielded the aforementioned attribute-substitution model of heuristic (see n. 33). The seriousness and sincerity of this acceding is reflected in the fact that Kahneman and Frederick excluded one of the original heuristics, anchoring-and-adjustment, from the extension of 'cognitive heuristic' in order to comply with the demand.

[31] It is worth noting that since each befitted type of explanation has at least one befitting conjunct, vagueness in the meaning of 'heuristic' can yield the appearance of being unproblematically explanatory in each type of explanation by, when a specific type of explanation is under consideration, emphasizing the corresponding conjunct and ignoring those that might be problematic.

[32] For readability's sake, in the following examination of the conjuncts and means of scientific-usefulness, the unwieldy and repetitive qualifier, "there is not sufficient reason to believe [X]," is usually replaced with simply, "not [X]" (or transformations thereof). As such, the claims therein should be interpreted per the conservative assessment of warrant that the qualifier would have provided.

SCIRA excluding simple-yet-ideally-rational-algorithms. That exclusion stemmed from SCIRA encompassing the conjuncts, DEVIRA and MEIRA. In this respect, DEVIRA not only does *not contribute* to *heuristic* making a causal-explanatory contribution, it *precludes heuristic* from doing so and from being scientifically-useful in this way.

Similarly, DEVIRA (and MEIRA) were behind *heuristic*'s not constituting a system-specific-objective-kind (SOK). This is because DEVIRA and MEIRA, by virtue of their contrast with the-ideally-rational-algorithm, exclude processes that constitute simple-yet-ideally-rational-algorithms (e.g., disjunctive syllogism and &-elimination), and we lack sufficient reason to believe that processes would belong to different systems depending upon their consistency with ideal-rationality. As such, DEVIRA precludes *heuristic* from being scientifically-useful via being a SOK.

Recall that *heuristic* did not contribute to why-have explaining because SCIRA was not the right property for selection (not even per SCIRA-score) because the right standard for selection regarded (something like) resource-independent practical usefulness, not conformity with ideal-rationality. DEVIRA (and MEIRA) entail the appeal to ideal-rationality. As such, they preclude *heuristic* from being scientifically-useful via contributing to why-have explaining.

Since any revision to *heuristic* should retain DEVIRA, *heuristic* (either as is or upon revision) should not be expected to deliver either a SOK or contributions to either causal explanation or why-have explaining.

How do the other conjuncts fare with respect to explaining cognitive biases? Intuitive-process is neither (a) necessary, nor (b) sufficient for yielding and thus, explaining cognitive biases. This is because (a) deliberative processes can yield cognitive biases (e.g., deliberative processes that deviate from ideal-rationality)—thus, intuitive-process is not necessary, and (b) intuitive processes can fail to yield cognitive biases (e.g., if the intuitive process is disjunctive syllogism)—and thus, it is not sufficient. As such, intuitive-process does not contribute to explaining cognitive biases.

Regarding MEIRA-(process), MEIRA is not necessary for yielding and thus, explaining cognitive biases, as a process possessing DEVIRA can yield them without possessing MEIRA. For instance, a DEVIRA process that is *less* economical than the-ideally-rational-algorithm can yield cognitive biases (e.g., if it turns out that the representativeness heuristic is *less* economical than the corresponding ideally-rational-algorithm, this would not impede its yielding the Linda-Conjunction-Fallacy). Since MEIRA entails DEVIRA, MEIRA is indeed sufficient for yielding cognitive biases. However, MEIRA regards the process' economy, which does not make a unique contribution to yielding cognitive biases (beyond that made by DEVIRA). That is the explanatory contribution of MEIRA is exhausted by DEVIRA. In this respect, MEIRA is explanatorily superfluous—*a la* if objects of gold metal *qua metal* caused an effect and such objects were identified as *gold metal*, the inclusion of *gold* -vs. *metal (simpliciter)*- would be explanatorily superfluous. Once DEVIRA is invoked, including MEIRA (i.e., excluding processes that lack MEIRA) amounts to adding an arbitrary factor in extension designation or kind-scope. That is, it makes the kind too narrow—it includes too little. To illustrate this with (an extreme) example, one could add the conjunct, more-than-four-and-less-than-6-times-more-economical-than-the-ideally-rational-algorithm. Like MEIRA, this conjunct would be sufficient for yielding cognitive biases, as it would entail DEVIRA. However, once DEVIRA is being invoked, appealing to more-than-four-and-less-than-6-times-more-economical-than-the-ideally-rational-algorithm would obviously be explanatorily superfluous and an arbitrary factor in kind-scope. In this respect, it would not make an explanatory contribution to explaining cognitive biases. In a similar respect, neither does MEIRA.

Given the proceeding, we lack sufficient reason to believe that either intuitive-process or MEIRA contribute to explaining cognitive biases, and thus, lack sufficient reason to believe that either intuitive-process or MEIRA are scientifically-useful to explaining cognitive biases. As such, DEVIRA is the only conjunct that we have sufficient reason to believe is scientifically-useful with respect to explaining cognitive biases. As such, the only way in which (we have reason to believe that) the kind, *heuristic*, is scientifically-useful is with respect to contrastively explaining cognitive biases and only insofar as SCIRA entails DEVIRA.[33]

[33] Recall that this dissertation focuses upon *cognitive heuristic* per its classical conception within the Kahneman and Tversky research program (Ch. 1, n. 6; Ch. 2 n. 27) (e.g., Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1996; 2000). This notion is distinct from the subsequent notion of *cognitive heuristic* per *attribute-substitution* (Kahneman & Frederick, 2002, p. 53-60). Attribute-substitution involves (e.g., with respect to the aforementioned availability heuristic), substituting the *target attribute* of assessment (e.g., the probability of an event) with a *heuristic attribute* (e.g., the ease with which similar events can be recalled). The heuristic attribute can be thought of as the *proxy* (attribute) (*a la* Ch. 1, §3.1). For both the classical and attribute-substitution conceptions, an important dynamic is *cognitive heuristic* being (or attempting to be) (1) a descriptive-oriented category, or more precisely, a *causally*-significant category (e.g., either being a SOK/ kind-unified-by-identification-with-a-distinct-mechanism, or making a narrow-causal explanatory contribution), while also being (2) a normatively-significant category (e.g., heuristics being essentially non-ideal) that explains normatively-significant explananda (e.g., explaining aspects of human irrationality), which cashes out as making a contrastive explanatory contribution within a normatively-construed taxonomy. In other words, *cognitive heuristic* is a kind that (attempts to) reflect carving in accordance with both (1) causation, and (2) normative evaluation. These two orientations/roles are in tension. For instance, regarding simple-yet-ideally-rational-algorithms, causal-carving dictates including them in the extension of *cognitive heuristic*, but normative-carving precludes it. Blindly importing this tension into the moral analogue would be a mistake. This makes the classical conception of *cognitive heuristic* more useful for adapting because it is more conducive to being reduced into elements that allow us to separate-out these orientations/roles—namely, the reduction of *cognitive heuristic* (or ‘cognitive heuristic’ or more specifically, the central property/feature, SCIRA), into the “conjuncts,” DEVIRA, MEIRA, and intuitive-process. *Intuitive-process* isolates the kind’s potential as a causally-significant category. *DEVIRA* (and *MEIRA*, though that’s a bit more complicated) isolates the kind’s normative-orientation or more precisely, its contrastive-explanatory potential. The attribute-substitution conception of *cognitive heuristic* also encompasses these two orientations/roles. It encompasses (1) potential causal-

7. ‘Moral Fallacy’

Let’s return to the topic of developing the moral analogue of cognitive heuristic. Adapting ‘cognitive heuristic’ without revision would involve importing moral analogues of SCIRA and its encompassed conjuncts—that is, importing shortcut-vis-à-vis-*the-ideal-moral-reasoning-procedure* (SCIMP) and the conjuncts, deviating-from-*the-ideal-moral-reasoning-procedure* (DEVIMP), more-economical-than-*the-ideal-moral-reasoning-procedure* (MEIMP), and moral-domain-regarding-intuitive-process. However, the preceding examination showed that the only way in which (we have reason to believe that) the kind, *cognitive heuristic*, is scientifically-useful is with respect to contrastively explaining cognitive biases and only insofar as SCIRA entails DEVIRA. As such, that is the only part of *cognitive heuristic* that should be imported. This yields a moral analogue of cognitive heuristic with the central property DEVIMP and an expectation of scientific-usefulness only regarding contrastively explaining moral biases. Given that the central property is not SCIMP, calling it “moral heuristic” would be misleading. As such, the moral analogue should be called “moral fallacy” (it will be amended to “*subjective* moral fallacy” in chapter 5). Assuming that *moral bias* is an adequate practical kind, *moral fallacy* explaining *moral bias* would be sufficient scientific-usefulness for *moral fallacy* to not have to constitute an objective kind or SOK; as such, *moral fallacy* need not

significance by identifying the kind with the distinct (allegedly) causal mechanism of substitution—i.e., *replacing* the target-attribute with the heuristic-attribute. (I am skeptical of whether the distinction between a “replacement” occurring or not is a causal difference, as opposed to merely a difference in there being a good versus bad match between the given target attribute and the assessment executed; if there is not a causal difference, the kind loses its causal-significance.) It also encompasses (2) potential normative-significance in that the attribute that (ideally) *should* be assessed is the target attribute, not the heuristic-attribute. In other words, assessing the heuristic-attribute is a *deviation* from the normative procedure of assessing the target-attribute. However, reducing attribute-substitution into elements that separates these two orientations/significances/roles is much more difficult and its results much messier than doing so with the classical conception (i.e., shortcuted-ness—i.e., SCIRA).

require unification via identification with a distinct cognitive mechanism. In addition, *moral fallacy* would presumably also be an adequate practical kind.

CHAPTER 4: ADAPTING 'COGNITIVE BIAS'

1. Worry: Scientifically-Admissible Moral Standard?

The definition of 'cognitive bias' is systematic error. Initiating adaption by replacing its cognitive-domain-regarding (or domain-general-regarding) features with moral-domain-regarding features yields a *prima facie* concept of 'moral bias' with the definition: systematic *moral* error. Identifying moral errors constitutes attributing the property, *moral erroneousness*, to judgments/decisions/actions. Recall the objective of developing a paradigm that can undergird fruitful moral psychological research—that is, fruitful *scientific* research. This raises a worry: Are attributions of the property "moral erroneousness" compatible with science? Consider a judgment that abortion is morally (im-)permissible. Should a *scientific* research program be in the business of attributing moral erroneousness (or correctness) to such judgments? It seems like such attributions would be inadmissible (if you will) within science. To specify *erroneousness*, judgments are not erroneous, *simpliciter*, they are erroneous with respect to some standard of correctness. Thus, we can specify the worry as: What standard, if any, could ground moral erroneousness within a scientific context? That is, what moral standard might be scientifically acceptable—i.e., scientifically *admissible*? More precisely, within the context of a scientific research program, what standard, if any, would be admissible such that a judgment/decision/action's deviation from that standard would warrant the attribution of the property, moral erroneousness, or the description of that judgment/decision/action as a *moral error*? For instance, surely the standard "Mark's moral view" is not a scientifically-admissible standard—that is, surely, deviating from Mark's moral view is an unacceptable basis for attributing moral erroneousness within a scientific context. Consider Lawrence Kohlberg's problematic adapting of Jean Piaget's epistemic model to the moral domain, as revealed by Carol Gilligan's critique (Gilligan, 1982; Kohlberg, 1958; Piaget, 1970). In hindsight, this episode shows how easily psychological research programs can mishandle the selection and privileging of moral norms and standards. As such, great care needs to be taken when identifying the MBF paradigm's standard of moral error.

Before attempting to identify that standard, I should address the basis upon which I will assess *scientific-admissibility*. One option is to formulate and justify a theory of scientific-admissibility. This would yield assessing whether a given moral standard is one that (I contend) the scientific community *should* deem admissible. A second option is to accept scientific-admissibility as it currently stands. That is, assess whether a given moral standard is one that the scientific community currently *would* deem admissible. This can be operationalized as whether the standard would pass muster with editors of scientific journals and members of grant-awarding committees. I adopt this latter option.

One objection to that option is (something like): it implies a dubious cultural-relativism regarding scientific-admissibility. However, recall that the objective of this dissertation is to develop an MBF paradigm that can undergird a fruitful cognitive psychological research. In other words, the point of this dissertation is to develop a paradigm that would actually be employed. This requires compliance with current scientific-admissibility. We can consider this a practical constraint on the larger project in which this dissertation partakes, and an aspect of that project that renders the dissertation an instance of applied philosophy. Adhering to this practical constraint does not imply rejecting challenges to current scientific-admissibility; it merely considers such to be a project beyond the scope of this dissertation.

1.1. Adapt Ideal-Theoretical-Rationality?

To conduct the search for a scientifically admissible moral standard, it will be useful to: (1) identify standards that have unproblematically grounded erroneousness in other applications of the heuristics-and-biases paradigm—we can refer to these as *unproblematic standards*, and (2) consider whether they can be adapted to identify *moral* errors within the context of a scientific research program.

One unproblematic standard is ideal-theoretical-rationality.[1] Exemplar constituents of ideal-theoretical-rationality are the principles of logic, statistics, and probability theory (e.g., *modus tollens*, and the relevance of sample size). In the terminology of judgment-and-decision-making psychology, this standard is often referred to as "epistemic rationality." An example of this standard being employed within the heuristics-and-biases paradigm regards ranking "Linda is a feminist bank-teller" as more probable than "Linda is a bank-teller" and identifying the ranking as an error by virtue of its violating the conjunction rule (i.e., no conjunction can be more probable than one of its conjuncts) (Ch. 2, §2.5). The conjunction rule is a constitutive principle of probability

[1] It is worth noting that there have been prominent disputes regarding whether deviations from ideal-theoretical-rationality (e.g., despite ecological validity) are indeed "irrational" (e.g., Goldstein & Gigerenzer, 2002; Lopes, 1991). These disputes have been referred to as the "rationality debate" or "rationality wars" (Doherty, 2003; Samuels, Stich & Bishop, n.d.). Such disputes may appear to cast doubt upon whether ideal-theoretical-rationality is indeed an unproblematic standard. However, the dispute can largely be disarmed by disambiguating "rationality," and clarifying that the deviations-in-question are departures from *ideal* rationality and not necessarily departures from *pragmatic* rationality—or in Jonathan Baron's (2004) terminology, departures from *normative* rationality, but not necessarily departures from *prescriptive* rationality. Such disambiguation may leave open the question of which type of rationality constitutes "rationality, *simpliciter*" (if that question even has an answer); but settling this is unnecessary for proceeding with productive research. Both types of rationality -and people's conformity with or deviation from them- are of interest. In fact, judgments/decisions/actions and processes that conform to *pragmatic* rationality, but violate *ideal* rationality constitute an interesting class that can only be identified upon accepting the legitimacy of both standards. In this respect, ideal-theoretical-rationality provides an unproblematic grounding of error, so long as such error as clarified as *ideal* error, and not necessarily *pragmatic* error. This distinction is discussed further in §2.1.

theory and in turn, ideal-theoretical rationality.

Can this unproblematic standard be adapted to ground *moral* erroneousness within the context of a *scientific* research program? The pertinent feature of ideal-theoretical-rationality is its provision of (perhaps, loosely speaking) *a priori* knowledge (e.g., the knowledge that ranking “Linda is a feminist bank-teller” as more probable than “Linda is a bank-teller” is erroneous). Adapting ideal-theoretical-rationality to assessing moral erroneousness would yield a standard that purportedly provided *a priori moral* knowledge. Exemplars of moral theories in the ballpark of such adaptions are those that are purportedly derived from rationality, such as Kantian ethics. An exemplar of the purported *a priori moral* knowledge that such a standard might provide is that violating the categorical imperative is morally erroneous. Such purported *a priori* moral knowledge is scientifically-inadmissible. In other words, while a Kantian may claim to have made unassailable derivations from rationality, nonetheless, the notion that the scientific admissibility granted to the conjunction rule be bestowed upon the categorical imperative is simply a non-starter.

Recall that due to the interest in actually conducting fruitful MBF research in the near term, practical constraints are heeded, and the basis of scientific-admissibility is what the scientific community *would* deem admissible. While unnecessary for this assessment, I will add that I am highly sympathetic to such non-admission. This is because the case for moral uncertainty and skepticism is simply too strong to admit purported *a priori* moral knowledge into scientific contexts. Such non-admission may seem more exclusionary than it actually is. For instance, scientific-inadmissibility does not entail that purported *a priori* moral knowledge is inadmissible in all contexts or domains of inquiry. In other words, just as the standards of evidence differ between the criminal and civil justice systems, the (in-practice) standards of justification may differ between scientific and other contexts. Furthermore, scientific-inadmissibility does not entail that for instance, the categorical imperative is false. Ultimately, the scientific-inadmissibility of purported *a priori* moral knowledge merely reflects the fact that metaethics and moral philosophy have not progressed to the point of delivering sufficiently incontrovertible pertinent moral *knowledge* (assuming providing such is even a proper aim of metaethics and moral philosophy). Reaching such a point is an *incredibly* steep climb. It requires incontrovertibly establishing that (1) there are moral facts, (2) those facts are *knowable*, (3) such facts are known, and (4) such known facts are sufficiently specific and pertinent to ascribe erroneousness to particular moral judgments (such as those regarding abortion or the death penalty) with a comparable epistemic warrant to that held when ascribing erroneousness to ranking "Linda is a feminist bank-teller" as more probable than "Linda is a bank teller." The conclusion here is not that metaethics and moral philosophy will never reach (4); it is merely that they have not reached it *as of now*.[2]

The appeal to moral skepticism above raises a partners-in-guilt objection. An interlocuter might voice this objection as follows: "Granted, skepticism can be marshalled against *a priori* moral knowledge. Nevertheless, skepticism can also be

[2] Ultimately, I will conclude that (what I call) *ideal-instrumental-moral-rationality* (IIMR) is a viable standard of moral correctness and error and I will adopt it for the MBF framework. Given the scientific-inadmissibility of adaptions of ideal-theoretical-rationality (and later, adaptions of *a posteriori* knowledge), one is pushed towards an option like IIMR. However, so long as one grants the viability of IIMR, agreeing that one is pushed towards adopting it is not necessary for being on-board with this dissertation's project. For instance, ratcheted-down versions of the scientifically-inadmissible claims (such as merely acknowledging that adapting ideal-theoretical-rationality, etc. would raise problems whose avoidance is reasonable) can be *practically* consistent with accepting IIMR. Nevertheless, I think the stronger scientific-inadmissibility claims are true and buttress adopting IIMR.

marshalled against ideal-theoretical-rationality. Since we do not think such skepticism precludes grounding erroneousness in ideal-theoretical-rationality, why think moral skepticism precludes grounding moral erroneousness in *a priori* moral knowledge?"

The difference is that, unlike moral principles, many of the principles of ideal-theoretical-rationality that have undergirded error-ascriptions (e.g., *modus tollens* and the relevance of sample size) are indispensable presuppositions of psychological research, and science in general. Acquiescing to skepticism of ideal-theoretical-rationality would make conducting science impossible. Recall that the ultimate objective of this dissertation is undergirding scientific research. As such, a presupposition of this project is accepting the presuppositions of science. A presupposition of science is rejecting skepticism of ideal-theoretical-rationality. As such, a presupposition of this project is rejecting skepticism of ideal-theoretical-rationality. In other words, ideal-theoretical-rationality is indispensable to the practice at-hand. Rejecting ideal-theoretical-rationality would amount to a self-excluding stance (*a la* a self-defeating argument). By this I mean, such rejection would exclude the project from the practice it intends to undertake. Such a stance would be akin to objecting to a theory proffered at an epidemiology conference by appealing to solipsism. Unlike skepticism of ideal-theoretical-rationality (or jumping ahead, skepticism of *a posteriori* knowledge via skepticism of the external world), moral skepticism leaves science intact. In other words, appeals to *a priori* moral knowledge are unnecessary for conducting science. As such, this partners-in-guilt objection is unsuccessful.[3]

[3] There may be room for a counterargument here and subsequent replies that would take us far afield. Nonetheless, that debate is somewhat moot as in this dissertation, scientific admissibility is ultimately determined by what the scientific community *would* deem admissible and the preceding was merely my explaining my sympathy for the non-admissibility of *a priori* moral knowledge (e.g., my sympathy for

1.2. Adapt A Posteriori *Knowledge?*

Another unproblematic standard is that which grounds erroneousness regarding the *haze illusion* (Brunswik, 1943; Kahneman & Frederick, 2002). The haze illusion refers to the tendency to overestimate the distance of objects in hazy weather conditions. It stems from the fact that when we assess the distance of an object (e.g., a mountain), one perceptual cue we use is the perceived blurriness of the object's edges; the blurrier the edges, the farther away we assess the object to be. Hazy weather increases perceived blurriness in general, and as such, the edges of objects appear blurrier than they otherwise would. As such, we exhibit the perceptual bias of systematically overestimating the distance of objects in hazy weather conditions. Such overestimation constitutes an error. This erroneousness is grounded in the estimate's deviation from what is known to be the actual distance—or more generally, the estimate's deviation from the unproblematic standard, *a posteriori* knowledge.

Can *a posteriori* knowledge be adapted to ground *moral* erroneousness within the context of a scientific research program? Such adapting would yield a standard that provided purported *a posteriori moral* knowledge. Moral claims in the ballpark of such are those that are purportedly grounded in the observation of moral properties (or something of that sort). However, within the context of a scientific research program, such purported *a posteriori* moral knowledge would be inadmissible. In other words, while one may claim to observe moral wrongness in a particular act or event, nonetheless, the notion that the scientific admissibility granted to the observation of a mountain's distance be bestowed upon purported observations of moral properties is simply a non-

the scientific admissibility granted to the conjunction rule being extended to the categorical imperative being simply a non-starter).

starter. Once again, while the basis of scientific-admissibility is what the scientific community *would* deem admissible, I am highly sympathetic to not admitting purported observations of moral properties given that metaethics and moral philosophy simply have not progressed to the point of delivering sufficiently incontrovertible pertinent moral *knowledge* (whether *a priori* or *a posteriori*).

2. Viable Solution: Ideal Instrumental Moral Rationality

An additional unproblematic standard is ideal-instrumental-rationality. Someone exhibits *instrumental rationality* (*simpliciter*) -i.e., *means-ends* rationality- "insofar as she adopts suitable means to her end" (Kolodny & Brunero, 2013). In this respect, instrumental rationality regards the optimization of goal-fulfillment (Stanovich, Toplak & West, 2008, p. 252).[4] Notably, the agent's ends are accepted as the ends in question

[4] (1) An exemplar instantiation of the means-ends relation is an *action* (means) causing the fulfillment of a goal (end)—e.g., one's drinking water (means) causing the quenching of one's thirst (end). However, deviations from this exemplar model are common. For instance, since most ends can constitute means to a more fundamental end, most "things" that constitute ends (e.g., goals, preferences, states-of-affairs, etc.) can (with some minor adjustments) also constitute means. For example, the action of overthrowing a tyrant can constitute a means to the end of freeing a populace, the latter of which can *also* constitute a means to the more fundamental end of maximizing happiness. Ends can feature values (e.g., fairness, liberty, etc.) and can be expressed as a goal in terms of realizing/instantiating that value (e.g., realizing fairness). The means to such goals can consist of not only actions (and intentions, decisions, etc.), but also judgments, attitudes, etc. For example, a tolerant attitude towards a different race can constitute a means to the end of equality; an accurate judgment about a person can constitute a means to the end of fairness. Furthermore, attitudes, dispositions, relations, judgments, and actions can constitute ends in themselves (especially in the context of non-consequentialist value systems). For example, instantiating a compassionate attitude, courageous disposition, friendship, unbiased judgment, and right action can constitute ends in themselves. In sum, all sorts of "things" (e.g., actions, states-of-affairs, attitudes, etc.) can be formulated to constitute means or ends, and often, both. (2) For an example of *instrumental*-(ir-)rationality, consider the following. Suppose Cassie is sitting in the back of a classroom. Once the class ends, she will have only ten minutes to get to her next class, which is on the other side of campus. She wants to get to her next class on time—that is her goal—her end. "Therefore," once the bell rings and class is over, Cassie gets out of her chair, and walks in a complete circle around her desk. After circling her desk, she moves to the desk in front of her, and circles it. Finishing that circumnavigation, she walks to the next desk in front of her and circles it. And so on. She continues to do this until she finally reaches the desk closest to the door, circles it, and then exits. Cassie's route out of the classroom is not only bizarre but given her goal of getting to her next class on time, the route is *irrational*. It is a route (i.e., means) she has little reason to take given her goal of getting to her next class on time (i.e., end). The *rational* route or means, given her end is something like, the more direct route of walking down her row to the front of the class and walking straight out the door. In this respect, the circling route is instrumentally irrational.

(though this will get more complicated later). In this respect, the ends are *subjective* ends, and the standard, instrumental-rationality, is a *subjective* standard (or something of that sort). This contrasts with standards that appeal to *objective* ends, such as Aristotelean accounts of practical reason. Fellow travelers of such views include objective theories of a person's good or well-being (e.g., objective list theories) and external reasons.[5]

Ideal-instrumental-rationality contrasts with *pragmatic*-instrumental-rationality.[6] The former evaluates means in terms of their contribution to the *ideal* satisfaction of ends; the latter, evaluates means in terms of their contribution to the *pragmatic* satisfaction of ends (e.g., accommodates satisficing).[7]

[5] It is worth noting that while instrumental rationality's ends are subjective, the extent to which a means contributes to the satisfaction of those ends is an objective matter.

[6] Restating the above in Jonathan Baron's (2004) terminology: ideal-instrumental-rationality -that is, *normative*-instrumental-rationality- contrasts with pragmatic-instrumental-rationality -that is, *prescriptive*-instrumental-rationality.

[7] The ideal/pragmatic distinction gets a little tricky regarding *instrumental* rationality. It was fairly straightforward regarding *theoretical* rationality. For instance, for a computationally daunting problem with low stakes, it may be pragmatic to use a heuristic that quickly and easily yields close, but slightly *incorrect* answers. In this respect, the heuristic can straightforwardly be *pragmatically* rational, despite not being *ideally* rational. Regarding *instrumental*-rationality, suppose you have the goal of reducing the length of your drive home from work. Suppose the means at issue are what route to take. The ideal means would be the shortest route. However, suppose that the shortest route is also very complicated—it requires an extraordinary number of turns, highway transfers, and unfamiliar roads. Suppose an alternative route is slightly longer, but much simpler. Given normal human limitations, you are susceptible to missing your turn and getting lost. As such, the route that is best to prescribe to you -the pragmatic route- may be the slightly longer, but simpler route. In this respect, that route is *pragmatically*-instrumentally-rational. Though the shorter route might not be pragmatically-instrumentally-rational, there is still a sense in which it is the best route—the sense that is independent of your limitations. That sense is captured by describing the shorter route as being *ideally*-instrumentally-rational. It might be tempting to argue that the *pragmatically*-instrumentally-rational *is* the *ideally*-instrumentally-rational. However, the difference between the two is not only conceptually substantive, it is also useful. For instance, the difference facilitates recognizing opportunities for interventions that overcome normal human limitations and allow for the realization of the ideal. For example, if the difference between the ideal and pragmatic route is significant enough, it may be worthwhile for you to get a smartphone equipped with GPS navigation. The trickiness of applying the ideal/pragmatic distinction to instrumental-rationality stems from different senses of "can" and "possibility." To illustrate this, consider that regarding the route-home problem, the most "ideal" route is flying in a direct line at the speed of light. Of course, this is impossible for humans and so, it is nonsensical to identify this as the ideal means. As such, there is a sense of possibility that constrains what can count as the ideal. However, once this condition is introduced, we can ask whether there is a principled way of preventing this constraint from collapsing the ideal into the pragmatic. While articulating such a principle is beyond the scope of this dissertation, I believe we can rely upon intuitions to preserve this distinction and prevent such collapsing. Nonetheless, this a complication

We can now turn to unproblematic uses of ideal-instrumental-rationality to ground erroneousness in applications of the heuristics-and-biases model. The formal doctrine (if you will) of judgment-and-decision-making psychology, in general, and heuristics-and-biases psychology in particular, includes a commitment to (ideal)-instrumental-rationality as the standard of assessment whenever decisions, etc. are evaluated (Baron, 2008; Over, 2004, p. 5; Stanovich, Toplak & West, 2008, p. 252). For example, this standard is used to ground attributions of erroneousness to irrational gambles. The erroneousness ascribed to an irrational gamble depends upon the presumption that the goal or end of such gambling is maximizing expected financial payoff (or something of that sort). What grounds privileging this end as the standard from which deviations constitute errors? For instance, why not privilege the end, having fun? Appeals to (something like) purported objective values to ground such privileging would be scientifically-inadmissible. However, appealing to the gambler's subjective ends is scientifically-admissible. The subjectivity here can be obscured by (a) the near universal acceptance of such ends, and (b) the presence of the standard's objective aspects—for instance, the role of ideal-theoretical-rationality (which is objective) in assessing the extent to which the means (e.g., a particular gamble) contributes to the ideal satisfaction of the player's ends (e.g., maximizing winnings).[8]

worth noting.

[8] Despite the official commitment to instrumental-rationality, I have some doubts about whether the standards employed in actual research are subjective (in the aforementioned sense). Ultimately, the standards are probably insufficiently specified to determine this. Nevertheless, if push comes to shove, I ultimately believe that the standards employed are subjective (or at least, would be upon rational reconstruction). Nonetheless, this ultimately, does not matter for my purposes. The primary function of reviewing unproblematic groundings of erroneousness by variants of the heuristics-and-biases paradigm is to use them to develop standards that can be employed in the MBF program. If adapting instrumental-rationality yields a useful MBF standard, whether instrumental-rationality was ever an employed standard is not that important. In other words, the standard developed for the MBF program will sink or swim on its own merits, not those of the inspiration for its development.

The proceeding (with some simplification) yields: per ideal-instrumental-rationality, an action/decision/judgment φ by agent A is an error insofar as φ-ing frustrates the ideal satisfaction of A's ends.[9]

Adapting ideal-instrumental-rationality begins with replacing cognitive-domain-regarding or domain-general-regarding features with moral-domain-regarding features. Recall that the definition of 'ideal-instrumental-rationality' is a standard that evaluates one's actions, etc. (i.e., means) in terms of their contribution to the ideal satisfaction of one's goals (i.e., ends). Applying the replacement procedure yields the following: ideal-instrumental-*moral*-rationality (IIMR) is a standard that evaluates one's actions, etc. (i.e., means) in terms of their contribution to the ideal satisfaction of one's *moral* goals—i.e., one's *moral* ends—or generally speaking, one's *morality*. In this respect, the standard can be referred to as *subjective morality* (not to be conflated with *moral subjectivism*—see §2.1).

Given a standard, we can construe actions/decisions/judgments that deviate from it as *mistaken* or *incorrect* vis-à-vis that standard—that is, as *errors* vis-à-vis that standard. As such, we can construe actions that deviate from IIMR as errors vis-à-vis *IIMR*—that is, errors vis-à-vis *subjective morality*—that is, *subjective moral errors* (and when systematic, *subjective* moral biases). The proceeding (with some simplification) yields the following *simple* model:

Simple Model of Subjective Moral Error: an action/decision/judgment φ by agent A is a

[9] This formalization may be vulnerable to counterexamples. However, the qualifiers, etc. necessary to redress them get very complicated very quickly and would take us far afield. As such, I will stick with the simple formalization given, as it captures the basic idea, which is sufficient for our purposes.

subjective moral error insofar as φ-ing violates A's morality (i.e., violates IIMR—i.e., frustrates the ideal satisfaction of A's moral ends).

For an example of such erroneousness, suppose Rachael opposes racial discrimination—that is, she supports racial egalitarianism. Being racially egalitarian is a moral goal of hers; racial egalitarianism is a moral end of hers. Rachael must choose between two job applicants: Katie, who is white, and Shauntel, who is African-American. Had Rachael been unaware of the candidates' races, she would have given the job to Shauntel. However, Rachael is aware of their races and is prone to an implicit bias against African-Americans due to her unconsciously associating African-Americans with negative traits. Rachael selects the candidate to hire by "going with her gut" and choosing the candidate that "feels right." She gives the job to Katie. Rachel's doing so frustrates (the ideal satisfaction of) her moral end, being racially egalitarian; her doing so violates her morality. As such, Rachael's giving the job to Katie constitutes a subjective moral error.

2.1. Independence from The True, Best, and/or Real Morality

An important feature of ideal-instrumental-moral-rationality (IIMR) is its independence from any notion of being *the true, best, and/or real* morality (or "*the true morality*," for short). Alternatively framed, IIMR is a standard or more aptly, a metric, and *not a moral theory* (it is distinct from *moral subjectivism*). To elucidate this distinction, let's focus on actions and consider the following.

Instrumental rationality (*simpliciter*) is a standard/metric. It concerns the extent to which one's actions (i.e., means) contributes to the satisfaction of one's ends.

Instrumental-rationality can also be proffered as a normative theory of practical reason (loosely speaking, a theory of what one ought to do) (Kolodny & Brunero, 2013). This theory claims (something like): *one ought to act* such that one's acts (maximally) contribute to the satisfaction of one's ends. Nonetheless, instrumental-rationality can be utilized as merely a standard/metric *without* any claims about practical reason or what one ought to do all-things-considered.

Now consider utilitarianism. Utilitarianism is a moral theory—a theory of morally right action. Simplified, it claims: one's action is morally right if it maximizes utility. *Utility maximization* can also constitute a standard/metric independent of claims to rightness. For example, it is coherent to ask how utilitarian a policy is (i.e., to what extent does it maximize utility) without making any appeal to rightness.

The preceding *theories* have the following form: one's action (1) has a status (e.g., is morally right, is what one ought to do) per (2) some standard/metric (e.g., the extent to which it: maximizes utility, is instrumentally-rational). IIMR is a standard/metric; it concerns the extent to which one's action contributes to the ideal satisfaction of one's moral ends. One *could* proffer IIMR as a moral theory, yielding: one's action (1) *is morally right* (2) to the extent that it conforms with IIMR. Such "is morally right" amounts to something in the ballpark of identifying IIMR as *the true* morality. We could refer to this moral theory as IIMR-*ism*. Nonetheless, IIMR can be utilized as a standard/metric independent of any notion of its being *the true* morality (for why one might utilize it in this way, see §4).

Let's replace "IIMR" with "subjective morality." Conforming with subjective-morality is merely a standard/metric. One could proffer subjective morality as a moral theory—i.e., proffer that one's action *is morally right* to the extent that it conforms with subjective morality. This would be a version of the moral theory, moral subjectivism (i.e., *individual moral relativism* or *speaker subjectivism*). However, one can appeal to or employ subjective-morality as a standard/metric *without* claiming that acts that conform to this standard/metric are morally right. That is, one can utilize subjective morality as a standard/metric independent of any claims to rightness—i.e., independent of any notions of its being *the true* morality. Just as utilizations of the standard/metric, utility-maximizing, is distinct from utilitarian-*ism*; utilizations of subjective morality as a standard/metric are distinct from moral subjective-*ism*. In this dissertation, "subjective morality" is a synonym for IIMR and is proffered as merely a standard/metric, not a moral theory.[10]

The previous adaptions of ideal-theoretical-rationality and *a posteriori* knowledge yielded standards that provided purported moral knowledge (either *a priori* or *a posteriori*). As knowledge constitutes (something like) justified, true belief, the standards provided purported moral truth. In this respect, the standards claimed to constitute *the true, best, and/or real* morality. Such claims are scientifically-inadmissible. In short, metaethics and moral philosophy -at least thus far- have not progressed to the point of delivering sufficiently incontrovertible pertinent moral *knowledge* that can yield ascriptions of moral erroneousness with a comparable epistemic warrant to that held by ascriptions of erroneousness to either ranking "Linda is a feminist bank-teller" as more probable than "Linda is a bank teller" or deviating from (justifiably purported) *a posteriori* knowledge of a mountain's distance. IIMR's independence from any claim to

[10] More precisely, IIMR is a specification of subjective morality. In this respect, IIMR is a theory of subjective morality. Nonetheless, they can be considered interchangeable within the text.

constituting *the true, best, and/or real* morality avoids this pitfall. In this respect, IIMR avoids this route to inadequacy. However, does it achieve adequacy?

We lack sufficient reason to believe that deviating-from-IIMR -i.e., deviating-from-subjective-morality—i.e., subjective moral error- is causally-aligned with the human cognitive system. As such, we lack sufficient reason to believe that it is an SOK. Nonetheless, recall that caring about a standard (e.g., ideal rationality), is sufficient grounds for construing objects and creating kinds in terms of their relation to that standard (e.g., the practical kind, *cognitive bias*). As such, *moral bias* can achieve adequacy -i.e., constitute an adequate practical kind- by our caring about subjective-moral-errors or more generally, caring about subjective-morality. For the moment, let us assume we (should) care about subjective morality—i.e., IIMR (I will justify this in §4). (Upon payment on this promissory note) this constitutes an adequate grounding of subjective moral erroneousness. As such, IIMR and subjective-morality is a viable option for an MBF moral standard.

Such utilization of IIMR/subjective-morality with the explicit absence of any claim to its being *the true* moral standard is especially conducive to allowing one to reject subjective morality with respect to moral theory and still (a) consider it a standard/metric of interest, and (b) consider an MBF program grounded in subjective morality to be a worthwhile research program. As such, one can be for instance, a Kantian, a utilitarian, or even a moral nihilist and still endorse undertaking a subjective-morality-based MBF research program (for why they might, see §4).

Since IIMR/subjective-morality is not claimed to be *the true* moral standard, adopting it does not preclude the development of other MBF programs that adopt a different standard (so long as that standard is one we care about). For instance, one might develop an MBF program that adopted utilitarianism as the moral standard. However, as I will later argue (§4.2), subjective morality is the best standard to adopt for an MBF research program, and thus the one I will utilize. As such, we can specify the MBF paradigm proffered in this dissertation as the *subjective* moral biases and fallacies paradigm, which features *subjective* moral biases (i.e., systematic *subjective* moral errors) and *subjective* moral fallacies.[11]

3. Subjective Moral Error

We can further flesh-out subjective moral error by responding to the following worry: since IIMR/subjective-morality accepts a person's moral ends, how would a person's actions/decisions/judgments deviate from their *own* ends to yield an error? One straightforward way in which this can occur is for one to fail to identify the action that constitutes the means to their ends. A simple way in which this can happen is by employing fallacious reasoning (e.g., affirming the consequent) when determining what action to take.[12]

[11] Claiming that a standard constitutes *the true morality* does not preclude its use in a MBF program. However, what is precluded is grounding that standard's adequacy in its purportedly being *the true* morality. For instance, one might think that the Pope's morality is *the true* morality and that may motivate developing a MBF program that utilizes the moral standard, the Pope's morality. That is fine, so long as the standard's adequacy is grounded in for instance, our caring about the Pope's morality as a standard (assuming we do). Furthermore, such caring can be motivated by a widespread belief that the Pope's morality is *the true* morality, so long as it is the caring that grounds the standard's inclusion in scientific inquiry.

[12] In this respect, moral errors do not require their being generated by -or the otherwise involving- moral fallacies or moral-domain-specific processes. For instance, racially discriminatory hiring decisions need not be the result of moral reasoning to be a moral error (subjectively or otherwise). Tracing moral errors to (arguably) non-moral processing (*a la* arguably, implicit biases) holds practical promise, as such errors are especially susceptible to debiasing measures (e.g., instituting blind review of job applications). In other words, they are cases where progress can be achieved through the *relatively* easy work of instituting best practices, as opposed to the more challenging task of changing people's hearts. It is worth noting that while moral errors need not be the product of moral-domain-specific processes, if those errors are of particular practical or scientific interest, there is nothing preventing a focus upon that subset. Also note that such moral erroneousness is distinct from moral blameworthiness (elaborated in §3.3.1).

A way to deviate from one's own moral ends that has broad implications is suggested by a canonical consideration in subjectivist theories of normativity, such as internalist theories of practical reason and subjectivist theories of well-being and the good (e.g., Brandt, 1979; Railton, 1986a, 1986b; Rosati, 1996; Sobel, 2017; Smith, 1995; Williams, 1981). One manifestation of this consideration regards *internal reasons*. Internal reasons are reasons grounded in an agent's desires.[13] *Internal* reasons contrast with *external* reasons, which are grounded elsewhere, such as in objective goodness. A *simple* model of internal reasons is:

Simple Model of Internal Reasons: An agent A has an internal reason to perform some action φ iff A *possesses* some desire whose satisfaction will be served by φ-ing.[14]

This model of internal reasons is too simple, as it cannot handle cases of false beliefs. For example (*a la* Williams, 1981, p. 78), suppose Gaston is thirsty. He sees a glass of clear liquid and believes it is water. As such, he desires to consume the content of the glass. Unbeknownst to him, the glass is filled with gasoline. According to the simple model, he has a reason to pick up the glass and drink the gasoline. This is a

[13] More precisely, internal *normative* reasons -as opposed to *motivational* or *explanatory* reasons- are distinguished by (something like) a function of the agent's *contingent conative set* (Sobel, 2009, p. 337) or suitably connected to the agent's *subjective motivational set*, which includes not only desires, but "dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent" (Williams, 1981, p. 81). "Desire" will be used as a stand-in for *contingent conative set*, etc.

[14] This discussion of internal reasons largely draws from the literature defending *reasons-internalism*. Reasons-internalism is a neo-Humean view that denies the existence of external reasons (Finlay & Schroeder, 2012). As such, proponents of reasons-internalism treat their models of internal reasons as models of normative reasons, *simpliciter*. Reasons-externalists can accept the existence of both internal and external reasons; as such, they can accept models of internal reasons. This dissertation is agnostic on reasons-internalism versus externalism. As such, models of internal reasons are simply that, and should not be construed as models of normative reasons, *simpliciter*.

reductio of the simple model.

Upon adaption to our subject matter, this consideration becomes that one can have *mistaken moral ends*. Just as (a) the simple model of internal reason allowed false beliefs to yield *mistaken desires* (Railton, 2007, p. 267) that in turn, yield putative reasons, likewise, (b) the simple model of subjective moral errors allows false beliefs to yield mistaken moral ends that in turn, yield putative assessments of moral erroneousness.

For an illustration, suppose that Justine values retributive justice. Suppose that executing a murderer would further her moral ends and executing an innocent person would frustrate those ends. Suppose Justine *falsely* believed that Iona was a murderer. As such, Justine would adopt—and possess—the moral end of executing Iona. As such, according to the *simple* model of subjective moral errors, Justine's sentencing Iona to death would not be a subjective moral error. However, this cannot be right, as Iona is innocent.[15] As such, just as Gaston's having a reason to drink the gasoline was a reductio of the simple model of internal reason, the assessment that putting innocent Iona to death is not a subjective moral error is a reductio of the simple model of subjective moral error (i.e., A's φ-ing is a subjective moral error insofar as φ-ing violates A's morality—i.e., frustrates the ideal satisfaction of A's moral ends).

3.1. Idealization

More refined models of internal reasons (e.g., Williams, 1981) avoid such reductios by (a) utilizing counterfactual idealizations of the agent (e.g., endowing the

[15] One might interpret "*moral* error" in a way that could indeed render sentencing Iona to death *not* a moral error. One might say, "It isn't a *moral* error; it's a *doxastic* error." However, what is meant by "moral error" is akin to what is typically meant by "morally wrong." One would say, "Killing innocent Iona is morally wrong." One would not say, "It's not morally wrong; it's doxastically wrong." The relation of "morally" to "wrong" is how one should interpret the relation of "moral" to "error."

agent with omniscience and perfect rationality) and (b) grounding internal reasons in the idealized agent's desires. This yields:

Idealized Agent Model of Internal Reasons: A has an internal reason to φ iff *A+ (idealized agent A) possesses* some desire whose satisfaction will be served by φ-ing.[16]

Given an idealization that includes omniscience, *idealized* Gaston (Gaston+) would know that the glass is filled with gasoline and thus, would not desire to consume its contents. As such, according to the idealized agent model of internal reasons, (non-idealized) Gaston would not have an internal reason to drink the gasoline. As such, the reductio is avoided.

A sensible way to avoid such reductios with IIMR (e.g., Justine's sentencing innocent Iona to death's not being a subjective moral error) is to likewise employ counterfactual idealizations. This is sensible because while IIMR regards a person's moral ends, it (should) only regard their *genuine* moral ends. That is, it should not necessarily regard whatever moral end a person believes she has or acts upon. In this vein, a moral end is *genuine* (or something of that sort) insofar as it would be possessed by A+ (i.e., agent A upon idealization)—that is, insofar as it is constitutive of A+'s morality. In this respect, a moral end of A is *genuine* insofar as it would "survive the idealization" of A. This yields:

[16] In the practical reason literature, what I am calling the "*idealized agent* model" is sometimes called the "*deliberative* model" (Arkonovich, 2011).

Idealized Agent Model of Subjective Moral Errors: A's φ-ing is a subjective moral error insofar as φ-ing violates *A+'s* morality (i.e., frustrates the ideal satisfaction of A+'s moral ends).

3.2. Extensional Adequacy & Non-Alienation

So, with respect to yielding an A+ for IIMR and subjective moral error, what are the right counterfactual idealizations (e.g., omniscience? perfect rationality?)? Before jumping into this question, it will be worthwhile to address some preliminaries.

Firstly, while there is an established canon of such idealizations regarding subjectivist value theory and practical reason (e.g., Brandt, 1979; Railton, 1986a, 1986b; Rosati, 1996; Sobel, 2017; Smith, 1995; Williams, 1981), unfortunately, there is a paucity of philosophical work done on idealization regarding subjectivist morality (in the relativistic sense—addressed shortly).[17] As such, much of the following will draw from adaption of the former.

Secondly, an important tension hangs over the consideration of idealization options. Subjectivism involves grounding an agent's normativity in an aspect of that agent (e.g., grounding her reasons in her desires).[18] Idealization involves hypothetically

[17] Some moral relativists do appeal to idealizations (e.g., Wong, 1984). In this respect, there is philosophical work that *utilizes* idealizations of subjective morality (at least with respect to idealizations of epistemic capacities bearing upon subjective morality). However, at least for the literature I have been able to find (e.g., Baghramian & Carter, 2015; Gowans, 2015; Prinz 2007; Wong, 1984; 2006), the idealizations, themselves, are not much explored. Indulging some sociological speculation, I suspect that this apparent hole in the literature stems from many philosophers rejecting speaker subjectivism (*qua* moral theory). This is akin to one neglecting specifications of utility because one is not a utilitarian. Insofar as one cares about utility, which is *far* from sufficient for utilitarianism, this would be unwise as a general practice. Perhaps there are enough utilitarians out there to pick up any slack. By contrast, there might be too few speaker subjectivists out there to get around to delving into this topic. This sociological dynamic could persist despite the consistency of (a) interest in idealized subjective morality and (b) a rejection of speaker subjectivism.

[18] Adapting this paragraph to subjective morality yields a "de-normativized" (or less explicitly normative) version, wherein for instance, subjective-*ism* and *normativity* are replaced with (something like)

changing the agent and deriving conclusions regarding the actual agent from the hypothetical agent (e.g., deriving an agent's reasons from the *idealized agent's* desires). Subjectivist idealization must thus thread the needle of (a) changing the agent enough to secure extensional adequacy (e.g., not yielding a reason to drink gasoline), while (b) not changing the agent in a way that threatens the resultant idealized agent's connection to the original agent (who ultimately, grounds the normativity—e.g., Finlay & Schroder, 2012, 1.2.3; Railton, 1986a, p. 46; Rosati, 1996). An illustrative (though possibly problematic) example of maintaining this connection is ensuring that the idealization process preserves the original agent's deepest values, final ends, or tie to the deep-self. Idealizations that lose this fundamental connection yield reasons, goodness, etc. from which the original agent is *alienated*. These dual desiderata of (a) extensional adequacy and (b) non-alienation are important to keep in mind whenever considering idealization options.

A further feature of idealized subjectivism's -or more precisely, neo-Humeanism's- commitment to non-alienation is (what we can describe as) the *individual-relativistic spirit*, which is an impulse to index to individuals. In other words, it is what drives adding the qualifier, "to you" or "for you," to claims such as "X is a reason." It is distinguished by its rejection of universalism. This spirit is most clearly manifested in the distance neo-Humean exemplars of subjective idealization maintain from Kantian views of practical reason (e.g., Arkonovich, 2013, p. 215; Bagnoli, 2017, sect. 3; Sobel, 2017, p. 6). While Kantian views satisfy some of the technical criteria of "subjectivism," these views contend that all idealized agents' desires, etc. converge (e.g., Korsgaard, 1986).

subjective *standards* and *subjective constructs of interest*. I'll just stick with the more reader-friendly normative version.

Given human diversity (let alone the diversity amongst possible rational agents), the individual-relativistic spirit manifests as strong skepticism that a view of practical reason could yield universal convergence without violating non-alienation—that is, without violating subjectivism (e.g., Joyce, 2001, p. 75). In other words, arriving at universal convergence shows that the view is "not really subjectivist." In this respect, such "real subjectivism" (implicitly at least) has an individual relativistic spirit.[19] Loosely speaking, the individual relativistic spirit is that which drives neo-Humeans to declare that Kantian views are members of a different camp.[20] In this dissertation, *subjectivism* can be presumed to be non-Kantian and non-universal subjectivism. The individual relativistic spirit also distances exemplars of idealized subjectivism from (some versions or interpretations of) ideal-observer theories (e.g., Firth, 1952; Kauppinen, 2014, §4.2). The most important takeaway for present concerns is that the individual-relativistic spirit clarifies the pertinent sense of 'non-alienation' invoked in the aforementioned desideratum.

3.3. Counterfactual Endowments

Having addressed these preliminaries, we can return to the question of what counterfactual idealizations are the right ones for yielding an A+ for IIMR and subjective moral error. We can split this question in two: (1) what should the idealized agent be endowed with? (e.g., omniscience? full rationality?), and (2) by which "mechanism" (if you will) should these capacities be exploited to yield genuine ends? (e.g., by identifying

[19] This divide between neo-Humeanism and neo-Kantianism reaches a head on the conceivability of an ideally rational sadist (e.g., Caligula) (Bagnoli, 2017, §3). Kantianism, which claims to derive morality from rationality, contends that an ideally rational Caligula is inconceivable. Neo-Humeans disagree; their position reflects the implicit individual relativistic spirit of neo-Humeanism subjectivism.

[20] The issue of whether Kantian practical reason counts as "real subjectivism" is not a surprising one given Kant's tendency to upend sweeping categories such as *the subjective*.

the unidealized agent's genuine ends with the idealized agent's ends?). To answer these questions, I will first lay out some options. These options are not exhaustive.

The first class of endowments to consider are informational endowments. One dimension on which informational endowments can vary is their domain. One such domain is *all non-normative facts*—that is, endowing the idealized agent with non-normative omniscience. (Henceforth, the qualifier, *non-normative*, will often be left off and should be considered implicit.)

One might think that such omniscience should be ruled out as a possible counterfactual idealization because it is too strong. One basis for thinking this might be something like: "Omniscience is grossly unrealistic for humans. As such, calling someone's action/decision/judgment a moral error due to their lacking something so grossly unrealistic is unfair." The problem with this line of thinking is that it conflates the commission of moral errors with blameworthiness.

To illuminate this distinction, consider a math problem whose solution requires mathematical abilities that are grossly unrealistic for any unaided human—for example, identifying the square root of 1,524,155,677,489. There is one correct answer to this problem: 1,234,567. Is a failure to provide this answer blameworthy? No. How about providing the incorrect answer: 1,234,500? Not only is that answer not blameworthy, it is praiseworthy for an unaided human to get so close to the correct answer. Nevertheless, the answer is still incorrect; it is still erroneous—providing that answer still constitutes making an error. Objections of the sort, "this is an unrealistic standard of error," would be confused. Mathematical correctness is not a standard that adjusts for realistic human expectations, and that is fine, so long as we keep in mind the distinction between mere error and blameworthiness.

A variant of the above objection that may get more traction is: “Omniscience is grossly unrealistic for humans. As such, calling someone’s action/decision/judgment *irrational* due to their lacking something so grossly unrealistic is unfair (and/or inaccurate)” (*a la* Joyce, 2001, §3.0, §3.7). The problem with this objection is that it conflates rationality *simpliciter* with *ideal* rationality. The standard at issue in this chapter is *ideal*-instrumental-moral-rationality. The pertinent generalization that encompasses the above example is *ideal*-rationality. Deviations from this standard are merely irrational *per ideal*-rationality. While it is unclear what “irrational [simpliciter]” amounts to, it cannot invalidate *ideal* rationality. Even if rationality *simpliciter* is identical with something like *realistic*, *pragmatic*, or *prescriptive* rationality, one still needs *ideal* rationality to make sense of it being wise to employ a calculator to answer the math question.[21] In short, rationality per pragmatic rationality, realistic rationality, etc. neither entails nor invalidates ideal rationality. Similarly, ideal irrationality (1) does not entail pragmatic, realistic, and/or all-things-considered irrationality and (2) does not invalidate pragmatic, realistic, and/or all-things-considered rationality. This was the fundamental confusion driving the aforementioned misguided “rationality wars” (n. 1). Just as the previous objection conflated deviations from accuracy (i.e., errors) with blameworthiness, this objection seemingly similarly conflates deviations from ideal rationality with blameworthiness.

In sum, once we recognize the distinctions between (1) erroneousness versus

[21] In a similar vein, recall (n. 6) the case of the driving-route. Though the easier-to-follow route was the pragmatically rational route to take, this cannot invalidate the ideal rationality of the shorter-but-more-complicated route, as evidenced by the possible worthwhileness of getting a GPS navigator in order to take the latter route.

blameworthiness, and (2) pragmatic, etc. rationality versus *ideal* rationality, we see that omniscience and other unrealistic endowments should not be ruled out as counterfactual idealizations for IIMR and subjective moral error.

A second domain of information is the *relevant facts* (Brandt, 1969-1970, p.45-46; Kolodny & Brunero, 2013). Having the-relevant-facts as an option renders omniscience obsolete. Any fact made accessible by omniscience that is relevant is covered by the-relevant-facts. The difference between the two is whether the idealized agent is endowed with access to irrelevant facts. Since those facts are irrelevant, such access does not matter. In other words, nothing that matters is lost by downgrading omniscience to the-relevant-facts. In addition, omniscience may open cans of worms and philosophical puzzles that the-relevant-facts do not. As such, the-relevant-facts is a preferable informational endowment over omniscience. I will return to informational endowments shortly.

In the gasoline case, we saw how false beliefs can yield defective desires that yield putative reasons. Such defective desires and, in turn, putative reasons, can also arise from failures of rationality. For instance, suppose Gaston correctly believes the following conditional: If the glass contains water, then the water bottle will be empty. Suppose he sees that the water bottle is empty and affirms the consequent, concluding: Therefore, the glass contains water. Believing such, he forms the desire to consume the liquid in the glass. Unbeknownst to him, the liquid is gasoline. As such, according to the simple model of internal reasons (which lacks idealizations), he would have an internal reason to drink the gasoline. As with informational idealizations regarding false beliefs, models of internal reasons can avoid such reductios from failures of rationality by bestowing rationality endowments.

One such endowment is inferential-perfection. So endowed, idealized Gaston would not affirm the consequent, and, in turn, would not conclude that the glass is filled with water, would not form the desire to consume the liquid in the glass, and thus, would not have an internal reason to drink from the glass. As such, inferential-perfection is a basic endowment of idealization.

An important condition of rationality is preference-coherence (which includes consistency and arguably, unity—e.g., Joyce, 2001, p. 71).[22] Achieving preference-coherence can involve revising preferences, including morally-relevant preferences. Revising preferences *per se* is not a big deal. For example, the Gaston case involved revising a preference for consuming the liquid. However, such cases involve (what we can call) shallow preferences. What shallow preferences are will be easier to grasp by contrasting them with *deep preferences*. Two major dimensions of deepness: (1) intrinsic-ness and (2) higher-order-ness. Intrinsic preferences involve preferring the object for its own sake. This contrasts with instrumental preferences, which prefer their object as a means to satisfying another preference (which either is -or ultimately, is grounded in- an intrinsic preference). Higher-order preferences, or more specifically, second-order preferences, are preferences for having a preference. Formulated in terms of desires, second-order desires are desires to have a particular desire. This distinction illuminates cases of *akrasia* (i.e., weakness-of-will), wherein one has a first-order desire for say, a delicious fatty food or procrastination, but does not have a second-order desire for having

[22] Per the conventions of the practical reason literature, I previously used the term, “desires,” as a stand-in for *contingent conative set*, etc. (n. 12). Regarding rationality, I will follow the conventions of its discipline and use the term, “preferences.” “Preference” and “desire” can be considered generally interchangeable stand-ins.

that first-order desire. Second-order desires are a form of endorsement. *Values* are commonly analyzed as second-order desires/preferences (e.g., Joyce, 2001, p. 69). Intrinsic preferences and second-order preferences are deep preferences; such preferences' revision risks alienation. In this respect, opening the door to idealization revising deep preferences (e.g., *values*) is a big deal.

Our preferences' vulnerability to coherence-based revision stems from the general messiness of human psychology. In other words, if one digs into the human mind, one will not find a tidy utility function that obeys the axioms of decision theory. Furthermore, it is debatable whether our preferences (ends, values, etc.) are stable (Zimbardo, 2007), known (Lichtenstein & Slovic, 2006), relevant (Haidt, 2001), assessable (Fischhoff, 1991), or present in any coherent form (Churchland, 1996). Achieving preference-coherence can require revising morally-relevant preferences.

The extent of revision required to redress coherence deficits varies. On the low-revision end of the spectrum is redressing inchoate (morally-relevant) preferences. This can merely require assigning preference-orderings and weightings. Such redressing might even be described as merely *specifying or clarifying* preferences, as opposed to *revising* them (Milgram, 1996, p. 504). For instance, suppose a question at hand is whether to support a tax cut for the wealthy. Suppose one holds deep preferences for liberty and equality. *Prima facie*, this entails that it is both rational and irrational to support the tax cut. This *prima facie* conflict can be resolved by clarifying the relative weights one assigns to these preferences (it is more complicated than that, but this is the basic idea).

On the high-revision end of the spectrum is redressing intransitive preferences, wherein for instance, one prefers: A>B, B>C, and C>A (in other words, when one is vulnerable to becoming a *money pump*—Ramsey, 1928). For example, one might prefer: harm-avoidance > liberty, liberty > fairness, and fairness > harm-avoidance. How to render intransitive preferences coherent is not obvious, as there are multiple solutions. For instance, one can replace A>B with B>A, or replace B>C with C>B, or replace C>A with A>C.

Especially when preference-incoherence cannot be resolved by mere clarification, one option is to privilege certain types of preferences. This is already done by privileging intrinsic preferences over instrumental preferences; though this is uncontroversial since such privileging is built into the notions of 'intrinsic' and 'instrumental.' That is, it is by definition that instrumental preferences are authoritative only insofar as they contribute to the satisfaction of intrinsic preferences. More controversial would be privileging second-order preferences above first-order preferences (*a la value-based Humeanism*—Radcliffe, 2012, p. 783). Furthermore, this first/second order dichotomy may be too simple as valuing often comes in degrees—Radcliffe, 2012, p. 777). Nevertheless, it is an attractive option as it would provide a mechanism for discounting (arguably) less authoritative preferences, such as those that stem from akrasia, addiction, impulses, whims, or alien desires (Hubin, 2003). Nonetheless, desires that lack second-order validation can still be intrinsic and strong; as such, revising them raises alienation concerns.

Once privileging preferences is on the table, another option is to privilege morally-relevant preferences according to the types of cognitive processes that yield them. For example, one might privilege morally-relevant preferences from deliberative (type 2) processes over those from intuitive (type 1) processes (e.g., Greene, 2007; Singer, 2005; *contra* Bartsch & Wright, 2005; Kass, 1998; Railton, 2014) —or at least, discount moral preferences that lack deliberative endorsement (or lack principled endorsement, *a la moral dumbfounding*—Haidt, Bjorklund & Murphy, 2004).

Privileging second-order preferences segues into idealization procedures that subject preferences to second-order evaluation. Such evaluation is at the heart of reflection. The pursuit of coherence and the privileging of second-order evaluation (arguably) drives the method of reflective equilibrium (Rawls, 1971). Something in the ballpark of utilizing reflective equilibrium -or at least reflection- for grounding moral error within scientific psychology has been endorsed by some (e.g., Stein, 2005; Tetlock, 2005), but rejected by others, particularly out of distrust of intuitions (e.g., Baron, 2005; Singer, 2005b). In sum, endowing rationality and in turn, preference-coherence can open a can of worms regarding preference-revision. I will return to rationality endowments shortly.

An additional class of endowments regards the agent's state-of-mind (or mental condition). One option is to have no such endowment. Most likely, this would mean that the default desires, etc. that the idealized agent inherits are those the agent possesses at the time of the judgment/decision/action-at-hand.

To begin with an extreme case, suppose that Heidi is hypnotized (Rosati, 1996, p. 302). Under hypnosis she may be made to have deep desires that would otherwise be completely alien to her. Presumably, such desires would not be legitimate grounding for her normativity (e.g., would not determine what is constitutive of her good). Such desires, once transferred to idealized Heidi, might survive informational and rationality endowments (e.g., if under hypnosis, all of Heidi's desires and aversions are flipped). To preserve the normative grounding of desires, desires due to hypnosis must be excluded.

Taking the example down a notch, suppose Debra is depressed—perhaps even suicidal (Bjorklund, Bjornsson, Eriksson, et. al, 2012, p. 126). This may affect Debra's desires and yield various alienated desires or eliminate non-alienated desires. Other mental conditions with similar effects include apathy, exhaustion, mania, obsessive compulsion, and emotional disturbance. To preserve the normative grounding of desires, these desires will also need to be excluded. As such, there is a need for the idealized agent to receive a mental condition endowment.

Drawing from the literature on motivational internalism (i.e., moral judgment internalism) one such endowment is "psychological normalcy" (Bjorklund, Bjornsson, Eriksson, et. al, 2012, p. 125-127). While intuitively compelling at first, "psychological normalcy" opens a can of worms. A classical commitment of subjectivists -or more precisely, Humeans- is that an idealized sadist (e.g., Caligula) is conceivable (*contra* Kantians). It seems that sadism is not psychologically normal but idealizing away sadism would violate a core Humean subjectivist commitment to non-alienation. Would psychological normalcy exclude all conditions that are classified as disorders in the DSM (American Psychological Association, 2013)? Should it?

Another state-of-mind that could render preferences illegitimate sources of normativity is inflamed passions (e.g., Brandt, 1943, p. 487; Rawls, 1971, p. 47). For instance, desires while enraged might require exclusion. In accordance, a common state-of-mind endowment is a state-of-calm (e.g., Rosati, 1996, p. 305; Sobel, 1994, p. 791; Wallace, 2014, sect. 5). Such states are often presumed to be optimal for deliberation and reflection. However, arguably, reducing passions might hinder that which drives moral preferences.

Another type of endowment is (conspicuously) moral psychological endowments. For example, bestowing maximal compassion and/or empathy. Such endowments could include sets of virtues, attitudes, and/or capacities that enhance sensitivity to each of the foundations of moral *psychology*, as proffered by Jonathan Haidt (2013)—namely, care, fairness, loyalty, authority, sanctity/purity, and liberty. Another endowment option is bestowing the "highest" stage of Kohlbergian moral development (1958; *contra* Gilligan, 1982). Other options include bestowing impartiality (*a la* the ideal observer—Firth, 1952), moral maturity (Bartsch & Wright, 2005, p. 546; Brandt, 1943, p. 486), selflessness (Daniels, 1979, p. 270) and open-mindedness (Richardson, 1994, p. 31).

3.4. Two-Tiered Subjective Idealization

Prima facie, it is unclear which endowments are the right ones to select for subjective moral errors. Such selection risks significant alienation. For instance, if I simply declare that A+ should be endowed with maximal reverence for authority, that could yield preferences from which one was alienated. Regarding Haidt's (2013) foundations of moral psychology, people vary not only in the weight they assign to such foundations, but also in whether they consider certain ones such as loyalty (which includes in-group loyalty) to have any normative authority at all.

A solution to this problem can be found by turning to the question of which "mechanism" of idealization to use. Recall the *idealized agent* model of internal reasons: An agent A has an internal reason to perform some action X iff *A+ (idealized agent A) possesses some desire* whose satisfaction will be served by X-ing (*a la* Williams, 1981). In this case, the "mechanism" is the introduction of A+ (as the vessel for endowments) and the appeal to A+'s desires. The solution is adapting *two-tier internalism* (Rosati, 1996).[23] Per such -upon adaption- the endowments selected are those that the agent would consider authoritative. For instance, should your idealized agent be endowed with maximal reverence for authority? That depends upon whether you consider such an endowment authoritative.

However, this move -as currently stated- would confront challenges faced by the simple model of internal reasons—namely, the granting of authority to preferences that have not been cleaned-up by idealization. Suppose an agent is enraged; one might think that being in such a state would delegitimize her granting or denying of authority to particular endowments. One response to this problem is to create a counterfactually idealized agent to select the endowments—i.e., create an idealized endowment-selector.

[23] Rosati's (1996) model regards a person's *good*, though it can be adapted for other applications. Such internalism (or per our terms, subjectivism) is *two-tier* in that the normative object (a person's good) is (1) grounded in (the desires of an idealized version of) that agent (as opposed to an objective list), and (2) the makeup of that idealized agent -i.e., the idealization conditions or endowments- is also grounded in that agent (as opposed to declared from "outside"). With respect to (1), we left off at the Ideal Agent Model of Internal Reasons (i.e., A has an internal reason to perform some action X iff *A+ possesses* some desire whose satisfaction will be served by X-ing). By -and during- Rosati (1996), (1) changes. For one, the Ideal Agent Model succumbs to the reductio of being unable to yield an internal reason for A to improve her rational capacities if A+ is already endowed with perfect rationality (Railton, 1986a, p. 53). This is an instance of the *conditional fallacy*, wherein an agent's reason (e.g., to improve her rational capacities) stems from the agent's non-idealized state (e.g., her rational capacities' being imperfect), but having that reason is "blocked" because the model requires that such reasons be grounded only in desires that would occur under idealization (e.g., under perfect rationality) (Finlay & Schroeder, 2012, §2.4). Such "blocked" reasons -i.e., reasons that "break" upon idealization- are also referred to as *fragile reasons* (Sobel, 2001). This reductio is avoided by the *Ideal Advisor* Model, wherein: A has an internal reason to perform some action X iff *A+ would desire that A possess some desire* whose satisfaction would be served by X-ing (*a la* Railton, 1986a; 1986b). Another problem is ensuring that A+ is acting in A's interest; this yields: A+ is to assume A's place -including her lack of endowments- upon completing the idealization procedure (Railton, 1986a, p. 53). An additional problem is *indirection*, wherein desire for -or explicit pursuit of- an object frustrates attaining that object (e.g., the *paradox of hedonism*) (Rosati, 1996, p. 304). This is rectified by having A+ desire not that A possess a certain desire, but that A pursue a certain action, etc. A further problem is doubt that the accounts provide a sufficient condition; this is rectified by replacing "(A has an internal reason…) *iff* (A+…)" with "only if" (Rosati, 1996, p. 311). With these developments, (1) ends up as: A has an internal reason to perform some action X only if, were A+ contemplating the situation of A as someone about to assume A's position and non-idealized state, A+ would desire that A perform X (*a la* Rosati, 1996). More specifically, Rosati's formulation -i.e., regarding an internalist conception- is: "something X can be good for a person A only if [A+] would care about X for [A], were [A+] under appropriate [idealization] conditions and contemplating the situation of her actual self as someone about to assume her position" (p. 303-304).

But which endowments should be bestowed upon that idealized endowment-selector. One could create a counterfactually idealized agent to select the endowments for that idealized endowment-selector—i.e., create an idealized endowment-selector for the idealized endowment-selector; however, this could continue *ad infinitum*. To avoid both this infinite regress and having non-cleaned-up preferences determining endowments, Rosati's solution is to state that the endowments selected be grounded in the agent under *ordinary optimal conditions*.

> We need a way to rule out alienated [endowments], without allowing the appropriateness of [endowments] to depend upon how they now strike a person, whatever her present state might be [e.g., enraged]. We can do so by requiring that a person regard counterfactual [endowments] as appropriate not in her present state, but under ordinary optimal conditions. These would include that a person not be sleeping, drugged, or hypnotized, that she be thinking calmly and rationally, and that she not be overlooking any readily available information. This list of conditions is not intended to be exhaustive. By 'ordinary optimal conditions' I mean whatever normally attainable conditions are optimal for reflecting on questions about what to [care about or do]….Thus, we might say, counterfactual [endowments] are appropriate only if a person would so regard them under ordinary optimal conditions. (p. 305)

Adapting this solution yields: the endowments of A+ are those the agent would consider *morally* authoritative under ordinary optimal conditions.

We are now in a position to more fully flesh out subjective moral error. The major ingredients are the following. (1) An action/decision/judgment φ by agent A is a subjective moral error insofar as φ-ing deviates from A's genuine morality—i.e., φ-ing violates ideal-instrumental-moral-rationality (IMMR)—i.e., X-ing frustrates the ideal satisfaction of A's genuine moral ends. (2) A morality -and moral end- is *genuine* insofar as it would survive idealization—i.e., is constitutive of *A+'s* morality (i.e., the morality of agent A upon idealization)—i.e., is a moral end possessed by A+. (3) A counterfactual idealization of A yields the relevant A+ insofar as the endowments bestowed upon A+ are those that A considers authoritative under ordinary optimal conditions. This yields:

Two-Tier Model of Subjective Moral Errors: A's φ-ing is a subjective moral error insofar as φ-ing deviates from the A's genuine morality (i.e., frustrates the ideal satisfaction of A+'s moral ends), wherein A+ is a counterfactual idealization of A upon whom is bestowed those endowments that A considers authoritative under ordinary optimal conditions.[24]

4. Why Subjective Morality?

We can now turn to justification. We have no reason to think that genuine IIMR/subjective-morality's derivative kinds (e.g., *moral-bias-vis-à-vis-subjective-morality*) constitute system-specific objective-kinds (SOKs). As such, their scientific adequacy -or more precisely, scientific admissibility- depends upon our caring about them, or more generally, our caring about (genuine) IIMR/subjective-morality.

4.1. Species of Instrumental Rationality

As evidenced by the cognitive-heuristics-and-biases program, we care about the extent to which people are ideally-instrumentally-rational. Since ideal-instrumental-*moral*-rationality (IIMR) is a species of ideal-instrumental-rationality, we have reason to care about the former. To care about ideal-instrumental-rationality regarding prudential matters, but not moral matters, is unjustifiably myopic. This simple point is sufficient for

[24] While I contend that this account is sufficient for communicating the basic idea, I recognize that it may have difficulty with certain cases and be susceptible to counterexamples (*a la* those raised in n. 23).

caring about IIMR (i.e., subjective morality) in-and-of-itself.

Nevertheless, an objection that might be raised is that IIMR can yield positive evaluations of racist actions and negative evaluations of egalitarian actions. This stems from IIMR's acceptance of an individual's (genuine) moral ends. For example, David Duke has racist moral ends (i.e., a racist moral code, wherein "moral" is meant in a descriptive and not normative sense). As such, (*ceteris paribus*) the more racist his actions are, the more he satisfies his racist ends, and the more he conforms with ideal-instrumental-moral-rationality. Per IIMR, Duke's racist actions warrant a positive evaluation, and his racially egalitarian actions constitute *errors*. Yielding such evaluations can appear to be a reductio of IIMR and the MBF research program.

It is not. This is because IIMR neither is nor claims to be a standard of *all-things-considered* evaluation. IIMR is merely one dimension by which an action/decision/judgment can be evaluated. It is not the only evaluative dimension that we care about; but it is one that we do care about. Likewise, ideal-*theoretical*-rationality is also merely a single dimension of evaluation. Suppose Duke performs a racist act that conforms to ideal-theoretical-rationality. Per that standard, the act also receives a positive evaluation. Nonetheless, this is not a reductio of caring about ideal-theoretical-rationality, which grounds much of the cognitive heuristics and biases program. All-things-considered, we may have good reason to prefer that Duke violates IIMR so that the satisfaction of his racist ends is frustrated. However, this does not impugn the value of IIMR any more than all-things-considered, preferring that Duke is theoretically-irrational impugns the value of theoretical rationality. For instance, it may be better if Duke affirms-the-consequent (as it would frustrate the fulfillment of his racist ends); however, this does not preclude caring about valid inference and construing such conclusions as errors. For example, if Duke's racist ends are frustrated by instantiating the availability bias, this does not impugn construing the availability bias as an error. All it means is that evaluating an action/decision/judgment as theoretically irrational is insufficient for determining its all-things-considered evaluation. In sum, just as theoretical rationality is one evaluative dimension among many, so is IIMR. In this respect, we still have (*pro tanto*) reason to care about IIMR (i.e., subjective-morality) in-and-of-itself.

4.2. The Best Standard for the MBF Program

An assumed premise of this dissertation is that we have reason to care about the MBF program's potential benefits (Ch. 1, §4). As such, we have reason to care about having an MBF program as a means to realizing those potential benefits. In other words, the MBF program is instrumentally valuable with respect to realizing the potential benefits. More specifically, we have reason to care about having the best version of an MBF program and in particular, the best moral standard (*de dicto*) with respect to realizing those potential benefits. One especially important benefit is the potential practical benefit of decreasing immoral acts/decisions/judgments—i.e., decreasing moral errors. This requires changing behavior. As such, a practical criterion of the best moral standard is its conduciveness to changing behavior. A standard that lacks this conduciveness is practically inadequate. A standard that (best) has such conduciveness is practically adequate (at least in this respect—addressed in §4.3). As such, it is a standard we have reason to care about (as a means to -best- realizing the practical benefit of reducing moral errors). As will be shown, (genuine) IIMR/subjective-morality is that practically adequate best moral standard.

Let us now turn to different possible standards and consider how they fare with respect to conduciveness to changing behavior. As endorsed by Jonathan Baron (2005), one possible standard is utilitarianism.[25] For illustrative purposes, let's focus on *simple*-utilitarianism. Simple-utilitarianism can entail very controversial judgments (e.g., involuntary organ redistribution being morally obligatory). If simple-utilitarianism is adopted as the standard, then opposing any such judgments constitutes a moral error. However, what would happen, in practice, when someone opposes such judgments? The research program would identify -or better yet, *declare*- that the person's judgment was morally erroneous, and the person would disagree—perhaps vehemently. What is the practical point of this result? Where do things go from there? Is the person expected to just accept the researchers' attribution of error? Why should they? In short, the standard, simple-utilitarianism, has an authority problem. People who do not accept the standard lack reason to accept its entailments. As such, they do not have reason to accept the research program's conceiving of their judgment as a moral error. Given this, how is the research program going to change behavior? For instance, how is debiasing going to work? Will it be manipulative? Coercive? The answer to how debiasing is going to work is that it simply won't. In short, dissenters from simple-utilitarianism will not have reason to change their behavior. As such, the standard, simple-utilitarianism, is not conducive to changing behavior and thus, is practically inadequate.

This problem is not unique to simple-utilitarianism. Controversial entailments arise with many moral theories (e.g., stereotypical-Kantian-ethics entailing it being wrong to lie to the murderer at the door). One response to such entailments is to select a

[25] More precisely, if we are heeding independence from *the true, best, and/or real morality* (§2.3), the standard/metric would merely be *maximizing utility* (i.e., shorn of normative claims).

less contentious standard. Cass Sunstein (2005) proposes the purportedly uncontroversial standard that he calls *weak consequentialism*.[26] Weak consequentialism features a utilitarian default that is muted by allowing the violations of rights to count among the consequences that matter. However, no matter how broadly acceptable this standard might be, it inevitably would not be accepted by everyone, as people have irreconcilable moral views. Such purportedly uncontroversial standards will inevitably clash with some person's morality and confront the authority problem. As such, we are back where we were with simple-utilitarianism. The clash between the research program and those who do not accept its standard may occur less often with weak consequentialism than it would with simple-utilitarianism, but the obstacle to changing behavior remains the same and likewise, renders the standard practically inadequate. Perhaps the most broadly acceptable standard, by definition, is cultural-relative-morality (i.e., not cultural relative-*ism*, but cultural-scope subjective-morality shorn of appeal to rightness per §2.1); however, this standard also confronts the same problem of inevitable disagreement yielding practical inadequacy.

The above standards have difficulty with changing behavior due to the authority problem. The crux of this problem is the potential divergence between these standards and a person's morality. This problem does not confront (genuine) IIMR/subjective-morality, as it accepts each person's (genuine) moral ends. As such, divergence does not arise. In other words, we do not have this issue of the person not accepting the standard because it is their own standard. As such, the agent already accepts it as authoritative (in

[26] Sunstein proposed a MHB framework that was groundbreaking, but full of problems including equivocations on "weak consequentialism" (Adler, 2005) and "heuristic" (Hahn, Frost & Maio, 2005), and dismissing Kantian ethics by not much more than simply pronouncing it a mere heuristic (Fried, 2005).

the relevant respect). If deemed necessary, we could even stipulate that the relevant moral ends for attributing erroneousness are those the agent accepts as authoritative (under ordinary optimal conditions). As such, the person (at least under optimal conditions) agrees with the researchers' attributions of moral erroneousness; that is, the agent accepts the researchers' conceiving of certain acts/decisions/judgments as moral errors. As such, she accepts that she has some reason to change her behavior. If necessary, we can stipulate these conditions as constituents of the pertinent sense of "accepting as authoritative." As such, debiasing is not imposed. Subjective-morality is the only type of standard that can achieve this (or at least, it is the type of standard that best achieves this). In this respect, IIMR/subjective-morality offers a conduciveness to changing behavior that other standards do not, and as such, is the best moral standard for the MBF program and is practically adequate. As such, we have reason to care about IIMR/subjective-morality as a means to (best) realizing the potential benefits of the MBF program.

Another standard worth mentioning is any-plausible-moral-theory. This standard is typically appealed to when a moral judgment is shown to be a function of an obviously irrelevant factor (e.g., such as ordering effects). It often manifests as statements such as, "no plausible moral theory would endorse this." For each appeal to the standard, any-plausible-theory, the person either agrees with the standard's assessment or does not. If the person agrees, then IIMR/subjective-morality would yield the same result; if the person disagrees, then the authoritativeness problem arises (and thus, the failure to be conducive to changing behavior). As such, subjective morality is still the better standard for the MBF program to adopt.

In addition to being the best moral standard for the MBF program, there are other potential benefits -or more aptly, applications- of (genuine) IIMR/subjective-morality. For instance, consider moral education, especially in public schools. In such contexts, appeals to *the true, best, and/or real* morality may constitute inappropriate indoctrination and a violation of liberal values (e.g., Hand, 2018[27]). Grounding moral education in IIMR/subjective-morality avoids these problems. In a similar vein, IIMR/subjective-morality may have applications in public reason liberalism (e.g., Rawls, 1996).

In all, we have reason to care about IIMR/subjective-morality both in-and-of-itself (§4.1) and as a means to (best) realizing the potential benefits of an MBF research program (§4.2). As such, the standard -and the kinds it yields, such as *subjective moral biases*- are practically adequate. As such, those kinds are (worthwhile) practical kinds and so long as they are recognized and utilized as such (as opposed to mistaking them for SOKs warranting causation-oriented unification), they are also scientifically adequate.

4.3. Subjective Morality and Moral Improvement

The big payoff of having IMMR/subjective-morality as the moral standard for the MBF paradigm is scientific and practical adequacy. Such adequacy is enabled by IMMR/subjective-morality's independence from *the true, best, and/or real* morality. More specifically, such independence avoids scientific-inadmissibility (unlike the adaptions of ideal-theoretical-rationality and *a posteriori* knowledge) (§1-2) and enables practical adequacy by averting the *authority problem* (§4.2).

However, one might be interested in a moral biases and fallacies program only insofar as it would yield moral improvement in terms of (what one considers to be) *the true, best, and/or real* morality. For instance, the initial presentation of the benefits of the

[27] Hand (2018) ultimately proffers a standard that constitutes (being encompassed by) *the true, best, and/or real* morality.

program (Ch. 1, §1) spoke of decreasing *immoral* acts/decisions/judgments (with issues of moral truth set aside until this chapter). The transition from *immoral* acts/decisions/judgments to *moral errors* to *subjective moral errors* might seem like a bait-and-switch. One might care about the practical criterion of conduciveness-to-changing-behavior (§4.2) only insofar as such behavioral changes align with the (what one considers) *the true, best, and/or real* morality (in other words, that practical criterion may be insufficient for practical adequacy). Furthermore, the reason to care about IIMR/subjective-morality in-and-of-itself was as a species of ideal-instrumental-rationality (§4.1). This might yield a conception of the practical moral payoff of an MBF program as something closer to enhancing moral agency than achieving moral improvement (per what one considers *the true, best, and/or real* morality).

Unfortunately, unless one is a speaker-subjectivist/individual-moral-relativist (which few are), one cannot have an MBF program that satisfies the conjunction of (1) scientific-admissibility (i.e., scientific-adequacy), (2) practical adequacy, and (3) a moral standard per (what one considers) *the true, best, and/or real* morality. To illustrate this, recall that direct appeals to *the true, best, and/or real* morality are scientifically-inadmissible. Nevertheless, one can seemingly circumvent this constraint by going outside of the scientific context, specifying (what one considers) *the true, best, and/or real* morality, formulating it in terms of scientifically-admissible properties (e.g., via reductive naturalism, identifying the Pope's morality as a standard-of-interest, etc.), and then entering the scientific context and introducing that standard as a standard-of-interest (assuming sufficient caring about it). While such moves can secure scientific-admissibility and utilize (an equivalent of) a moral standard per (what one considers) *the* *true, best, and/or real* morality, such moves sacrifice practical adequacy insofar as the standard runs into the authority problem (§4.2). That is, rendering (what one considers) *the true, best, and/or real* morality scientifically-admissible also renders it unauthoritative.

Nonetheless, one can get an MBF program with (1) scientific-admissibility, (2) practical adequacy, and *a degree of* (3) a moral standard per (what one considers) *the true, best, and/or real* morality, insofar as individuals' genuine subjective morality aligns with (what one considers) *the true, best, and/or real* morality. Furthermore, an MBF program might focus upon relatively uncontroversial cases, which would yield high rates of alignment. Focusing on uncontroversial cases can still be very productive when there is significant susceptibility to committing subjective moral errors and biases. For instance, if one considers racial egalitarianism to be a part of *the true, best, and/or real* morality, then insofar as racial egalitarianism is often a part of individuals' genuine morality and individuals are susceptible to committing subjective moral errors regarding such, an MBF program would yield moral improvement per (what one considers) *the true, best, and/or real* morality. In this respect, if one is interested in moral improvement per (what one considers) *the true, best, and/or real* morality, then insofar as there are cases of alignment with individuals' *genuine* morality, one should be interested in an MBF program.

Some areas, such as racial egalitarianism, may be prone to alignment, and thus yield moral improvement per (what one considers) *the true, best, and/or real* morality. However, such alignment and moral improvement are not inevitable. For instance, consider people's tendency to give more money to identifiable victims than statistical victims. So long as individuals do not *genuinely* consider identifiability morally-relevant, this constitutes a subjective moral bias. An MBF program (including subjective moral debiasing) might identify, explain, and ultimately, rectify that bias—i.e., rectify that distortion per identifiability. However, whether that distortion would be rectified by increasing donations to statistical victims or *decreasing donations to identifiable victims* depends upon people's genuine subjective morality.[28] As such, if increasing donations to victims constitutes moral improvement per (what one considers) *the true, best, and/or real* morality, then it is possible that people's genuine subjective morality would rectify the distortion by decreasing donations to identifiable victims and thus, subjective moral debiasing would yield moral regression per (what one considers) *the true, best, and/or real* morality.

In this respect, the extent to which pursuing an MBF program would yield moral improvement (per what one considers *the true, best, and/or real* morality) is sensitive to the extent to which individuals' *genuine* morality aligns with one's own. In other words, if one is only interested in moral improvement per (what one considers) *the true, best, and/or real* morality, then perhaps one's supporting an MBF program is ultimately, a bet that individuals' *genuine* morality more aligns than misaligns with one's own.

[28] Some debiasing measures regarding the identifiable victim effect have yielded greater deductions in donations to identifiable victims than increases in donations to statistical victims (Small, Lowenstein, & Slovic, 2007).

CHAPTER 5: CLOSING REMARKS

This dissertation pursued scientifically and practically adequate moral analogues of cognitive heuristics and biases via adapting the CHB paradigm (Ch. 1, §1). This yielded the concept, 'subjective moral bias' (Ch. 4), with the meaning: systematic subjective-moral-error (Ch. 4, §2), wherein (per the two-tier model—§3.4) A's φ-ing is a subjective-moral-error insofar as φ-ing violates A's genuine morality—i.e., violates genuine-ideal-instrumental-moral-rationality—i.e., frustrates the ideal satisfaction of A+'s moral ends (§3.1), wherein A+ is a counterfactual idealization of A upon whom is bestowed those endowments (§3.3) that A considers authoritative under ordinary optimal conditions (§3.4). Such biases are scientifically adequate—or more precisely, scientifically-admissible (§1). This stems from firstly, scientific-inadmissibility being avoided via the biases' standard of error's being adapted from ideal-instrumental rationality (§2), which -unlike ideal-theoretical-rationality (§1.1) or *a posteriori* knowledge (§1.2)- is subjectively-grounded (if you will), and thus, upon adaption, yields a standard -namely, IIMR/subjective-morality (qua standard/metric, and not moral theory)- that is independent of appeals to *the true, best, and/or real* morality (§2.1), which moral philosophy and metaethics have yet to render sufficiently incontrovertible for admission in scientific contexts (§1.1). *Subjective moral bias* secures scientific adequacy, despite not being causally-aligned -i.e., not constituting a system-specific-objective-kind (SOK) (§2)- by constituting an adequate practical kind per its relation to a standard -namely, genuine IIMR/subjective-morality- about which we have reason to care (§4). Such reason-to-care includes reason to care about genuine IIMR/subjective-morality, intrinsically, per its being constitutive of ideal-instrumental-rationality, *simpliciter* (even if such care is not necessarily sufficient for all-things-considered evaluations) (§4.1). Such reason-to-care also includes reason to care about genuine IIMR/subjective-morality, instrumentally (§4.2), in that it is the best standard for realizing the potential benefits of the MBF program (Ch. 1, §1 & §4) due to its satisfying the practical criterion of conduciveness-to-changing-behavior due to its authoritativeness for individuals (i.e., avoiding the *authority problem*) stemming from its subjectivity (Ch. 4, §4.2). An MBF program per genuine IIMR/subjective-morality (including subjective moral debiasing) would yield moral improvement with respect to (what one considers) *the true, best, and/or real morality* insofar that standard aligns with individuals' *genuine* morality (§4.3).

The adaption also yielded 'moral fallacy,' with the meaning: reasoning that deviates from the ideal moral reasoning procedure (i.e., possesses DEVIMP) (Ch. 3, §7). This stems from the only way (in which we have reason to believe) that *cognitive heuristic* is scientifically-useful is neither per making a causal explanatory contribution (§2), nor constituting a (system-intrinsically objective—§3.1) system-specific objective kind (SOK) (§3), nor contributing to why-have explaining (§4), but only per *contrastively* explaining biases and only insofar as SCIRA entails deviating-from-the-ideally-rational-algorithm (DEVIRA) (§5-6). Since we lack sufficient reason to believe that either DEVIRA contributes to any of the other discussed ways of *cognitive heuristic* being scientifically-useful, and since the other conjuncts, MEIRA and intuitive process, do not contribute to explaining cognitive biases (§6), only DEVIRA-explaining-cognitive-biases should be imported, yielding deviates-from-the-ideal-moral-reasoning-procedure (DEVIMP) as the central property of "moral heuristic," or more aptly named, "*moral* *fallacy*" (§7).

A moral reasoning procedure is *ideal* insofar as it yields only morally correct judgments/decisions/actions—i.e., does not yield moral errors.[1] Given that the moral errors in-question are subjective-moral-errors per violating genuine IMMR/subjective-morality, the ideal-moral-reasoning-procedure in-question is per genuine IMMR/subjective-morality. As such, the moral fallacies in-question are *subjective* moral fallacies. In this respect, adaption ultimately yields 'subjective moral fallacy' with the meaning: reasoning that deviates from the ideal moral reasoning procedure per genuine IMMR/subjective-morality. Given the contrastive explanatory relation that *subjective moral fallacy* is poised to have with *subjective moral bias* (which is a practically adequate kind), *subjective moral fallacy* is poised to be an explanatorily adequate kind, and in this respect, a scientifically adequate kind. Furthermore, it is presumably of practical interest as well, and thus, a practically adequate kind in its own right. As such, it is unnecessary to unify *subjective moral fallacy* via identification with a distinct cognitive mechanism.

As for the identification of a reasoning procedure as *moral*, I will defer to current standards, such as those that govern the designation of processes as *moral* processes.[2]

Recall the hypothesized process that used cuteness as a proxy for moral status (Ch. 1, §2). Suppose that Erin's execution of this process yields a judgment that grants

[1] Basing the ideal-ness of a moral-reasoning-procedure upon the correctness of its *judgments/decisions/actions* may seem off-target; it may seem that the definition should more directly focus on the procedure, itself. However, such focus on the judgments/decisions/actions -i.e., the output-per-inputs mediated by the procedure- is akin to defining inferences' validity in terms of premises-conclusion (i.e., input-output) truth-preservation.

[2] Delineating the moral psychological domain is beyond the scope of this dissertation. Nonetheless, I can say that I am hesitant to infuse the concept of 'moral fallacy' with unnecessary presuppositions, such as conceptually presupposing the existence of moral-domain-specific processes. It is worth noting that the existence of such processes is not required for scientific inquiry into moral reasoning.

low moral status to rats (e.g., vis-à-vis squirrels) due to the former's low levels of cuteness. Is this an instantiation of a *subjective moral bias*? Yes, *if* it is a subjective-moral-error that is systematic (either with respect to Erin's judgments across time or between Erin and other individuals).[3] Let's stipulate that such judgments are systematic. Are such judgments *subjective moral errors*? Yes, *if* Erin is such that the judgment violates her genuine morality—i.e., violates Erin+'s subjective morality—i.e., violates ideal-instrumental-moral-rationality with respect to Erin+—i.e., frustrates the ideal satisfaction of Erin+'s moral ends, wherein Erin+ is a counterfactual idealization of Erin upon whom is bestowed those endowments that Erin considers authoritative under ordinary optimal conditions. Another way to frame this is whether Erin+ would consider cuteness an inapt proxy in this case (§1.3). I presume that few people -upon idealization- would endorse cuteness as a criterion of moral status; however, that someone would endorse such cannot be ruled out *a priori*.[4]

Is that process an instantiation of a *subjective moral fallacy*? Yes, *if* it deviates from the ideal moral reasoning procedure (possesses DEVIMP)—i.e., *if* it is a procedure that yields subjective moral errors.[5] If the process yields a subjective moral error in this case, then it is necessarily not ideal. Cutting to the crux of the matter, whether the process

[3] We can stipulate that the standards of systematicity that yield a moral bias are the same as those that yield a cognitive bias. Such binary designations of *constituting-a-bias* or *not* papers over various continuous metrics of systematicity. It is certainly not necessary for cognitive errors to be universal between individuals or completely constant within an individual to constitute a cognitive bias.

[4] The operationalization of idealization is a project for future research that is beyond the scope of this dissertation. Nonetheless, it could resemble other attempts at yielding privileged value-laden judgments, such as medical consultation per the *deliberative model* of the physician-patient relationship (Emanuel & Emanuel, 1992). Such operationalizations will of course, be far from perfect; nonetheless, providing an in-principle definitive result is theoretically important.

[5] This constitutes a fairly simple model of subjective moral fallacy. Nonetheless, it communicates the basic idea and reflects significant strides from moral-analogue-of-cognitive-heuristic. Just as the simple model of internal reasons required modifications, similar modifications will presumably be required for this simple model of subjective moral fallacy—though this is beyond the scope of this dissertation.

is a subjective moral fallacy will depend upon whether cuteness is an inapt proxy—that is, whether assessments of cuteness perfectly correlate with assessments of moral status (*a la* a synthetic entailment between variables). As proxies can vary in the extent to which they deviate from perfect correlation, processes can vary in the extent to which they constitute subjective moral fallacies—i.e., the extent to which they are subjectively morally fallacious. Perhaps some intuitive threshold will emerge for binary assessments of whether a process constitutes a subjective moral fallacy or not, but the more precise reality will be continuous.

The proffered meanings of 'subjective moral bias' and 'subjective moral fallacy' provide a conceptual foundation for a *subjective moral biases and fallacies* paradigm that could undergird fruitful moral psychological research. This includes (1) illuminating various subjective moral errors by identifying and unifying them as instantiations of subjective moral biases, (2) (contrastively) explaining many such biases with subjective moral fallacies, and (3) yielding subjective moral debiasing techniques to reduce the incidence of subjective moral errors. Of course, much work remains, but the preceding provides a solid groundwork to build upon.

REFERENCES

Adler, M. D. (2005). Cognitivism, controversy, and moral heuristics. *Behavioral and Brain Sciences, 28*(4), 542-543.

American Civil Liberties Union. (2014). Racial disparities in sentencing. Retrieved from https://www.aclu.org/sites/default/files/assets/141027_iachr_racial_disparities_aclu_submission_0.pdf

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington D.C.: American Psychiatric Pub.

Anderson, J. R. (2005). *Cognitive psychology and its implications* (6th ed.). New York, NY: Worth.

Arkonovich, S. (2011). Advisors and deliberation. *The Journal of Ethics, 15*(4), 405-424.

Arkonovich, S. (2013). Varieties of reasons/motives internalism. *Philosophy Compass, 8*(3), 210-219.

Associated Press. (2007, Oct 2). Roombas fill an emotional vacuum for owners. Retrieved from https://www.nbcnews.com/

Baghramian, M., & Carter, J. A. (2018). Relativism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2018/entries/relativism/

Bagnoli, C. (2017). Constructivism in metaethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2017/entries/constructivism-metaethics/

Balota, D. A., & Marsh, E. J. (2004). Cognitive psychology: An overview. In D. A. Balota & E. J. Marsh (Eds.), *Cognitive psychology: Key readings*. Psychology Press.

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. New York, NY: Delacorte Press.

Baron, J. (2004). Normative models of judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 19-36). Malden, MA: Blackwell.

Baron, J. (2005). Biting the utilitarian bullet. *Behavioral and Brain Sciences, 28*(4), 545.

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York, NY: Cambridge University Press.

Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists.* New Jersey: Lawrence Erlbaum Associates.

Bartsch, K., & Wright, J. C. (2005). Towards an intuitionist account of moral development. *Behavioral and Brain Sciences, 28*(4), 546.

Bazerman, M. H., & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what's right and what to do about it.* Princeton, NJ: Princeton University Press.

Bechtel, W., Abrahamsen, A., & Graham, G. (1999). The life of cognitive science. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science* (pp. 1-104). Malden, MA: Blackwell.

Bechtel, W., & Wright, C. D. (n.d.). What is psychological explanation? Retrieved from http://mechanism.ucsd.edu/~cory/psychexpln.doc.

Bermudez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. New York, NY: Routledge.

Bird, A. (1998). *Philosophy of science*. Routledge.

Bjorklund, F., Bjornsson, G. Eriksson, R., Olinder, R., F., & Strandberg, C. (2012). Recent work on motivational internalism. *Analysis Reviews, 72*(1)*,* 124-137.

Borgi, M. Cogliati-Dezza, I., Brelsford, V. Meints, K., & Cirulli, F. (2014). Baby schema in human and animal faces induces cuteness perception and gaze allocation in children. *Frontiers in Psychology, 5,* 1-12.

Botterill, G., & Carruthers, P. (1999). *The philosophy of psychology*. Cambridge, UK: Cambridge University Press.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*(34)*,* 1-21.

Brandt, R. B. (1943). The significance of differences of ethical opinion for ethical rationalism. *Philosophy and Phenomenological Research, 4*(4)*,* 469-495.

Brandt, R. B. (1969-1970). Rational desires. *Proceedings and Addresses of the American Philosophical Association, 43,* 43-64.

Brandt, R. B. (1979). *A theory of the good and the right.* London: Oxford.

Braun, D. (1995). Causally relevant properties. *Philosophical Perspectives, 9,* 447-475

Bruers, S. (2013). Speciesism as a moral heuristic. *Philosophia, 41*(2), 489-501.

Brownstein, M. (2015). Implicit Bias. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2015 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2015/entries/implicit-bias/

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50*(3), 255-272.

Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor production framework. *Journal of Risk and Uncertainty, 19*(1-3), 7-42.

Cherniak, C. (1994). Rationality. In S. Guttenplan (Ed.), *A companion to the philosophy of mind* (pp. 526-531). Malden, MA: Blackwell.

Cho, S., Gonzalez, R. & Yoon, C. (2011). Cross-cultural difference in the preference for cute products: Asymmetric dominance effect with product designs. *Proceedings of IASDR*.

Chomsky, N. (1980). *Rules and representations*. New York, NY: Columbia University Press.

Churchland, P. M. (1996). The neural representation of the social world. In L. May, L. Friedman & A. Clark (Eds.), *Mind and morals* (pp. 91-108). Cambridge, MA: MIT Press.

Civil Aviation Authority. (2002). Fundamental human factors concepts. Safety Regulation Group. Retrieved from http://www.caa.co.uk/docs/33/CAP719.PDF

Connolly, T., Arkes, H. R., & Hammond, K. R. (Eds.). (2000). *Judgment and decision making: An interdisciplinary reader* (2nd ed.). New York, NY: Cambridge University Press.

Cooper, J.R. (2005). Curing analytic pathologies: Pathways to improved intelligence analysis. Center for the Study of Intelligence. United States Central Intelligence Agency. Retrieved from https://www.cia.gov/csi/books/curing_analytic_pathologies_public/

Cosmides, L., & Tooby, J. (2006). Evolutionary psychology, moral heuristics, and the law. In G. Gigerenzer & C. Engel (Eds.), *Heuristics and the law* (pp. 175-205). Cambridge, MA: MIT Press.

Craver, C. (2006). When mechanistic models explain. *Synthese, 153*(3)*,* 355-376.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy, 72,* 741-764.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, MA: MIT Press.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-144). Cambridge, MA: MIT Press.

Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy, 76*(5)*,* 256-282.

Dennett, D. (2002). True believers: The intentional strategy and why it works. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings.* (pp. 558-566). Oxford: Oxford University Press.

Doherty, M.E. (2003). Optimists, pessimists, and realists. In L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 643-679). Cambridge University Press.

Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Dubljević, V., & Racine, E. (2014). The ADC of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds, and consequences. *AJOB Neuroscience, 5*(4), 3-20.

Ehrenfreund, M. (2015, Feb 2). The FBI director just quoted from Avenue Q's "Everyone's a Little Bit Racist." That's huge. Wonkblog. The Washington Post. Retrieved from http://www.washingtonpost.com

Emanuel, E. J., & Emanuel, L. L. (1992). Four models of the physician-patient relationship. *Journal of the American Medical Association, 267*(16)*,* 2221-2226.

Evans, J. St. B. T. (2008). Dual processing accounts of reasoning, judgment, and social cognition. *Annual Review, 59,* 255-278.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual processing theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3)*,* 223-241.

Eysenck, M. W. (2001). *Principles of cognitive psychology.* Psychology Press.

Finlay, S., & Schroeder, M. (2012). Reasons for action: Internal vs. external. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2012 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2015/entries/reasons-internal-external/

Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research, 12*(3), 317-345.

Fischer, R. W. (2016). Disgust as heuristic. *Ethical Theory and Moral Practice: An International Forum, 19*(3)*,* 679-693.

Fischhoff, B. (1991). Value elicitation: Is there anything in there? *American Psychologist, 46*(8), 835-847.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.

Fleischhut, N., & Gigerenzer, G. (2013). Can simple heuristics explain moral inconsistencies? In R. Hertwig & U. Hoffrage (Eds.), *Simple heuristics in a social world* (pp. 459-485). New York, NY: Oxford University Press.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fried, B. H. (2005). Moral heuristics and the means/ends distinction. *Behavioral and Brain Sciences, 28*(4)*,* 549-550.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103,* 592-596.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology, 8*(2), 195-204.

Gigerenzer, G. (2008). Moral intuitions = fast and frugal heuristics? In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 2): *The cognitive science of morality* (pp. 1-26). Cambridge, MA: MIT Press.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York, NY: Oxford University Press.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Gilovich, T., & Griffin, D. (2002). Heuristics and biases: Then and now. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1-17). New York, NY: Cambridge University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, NY: Cambridge University Press.

Glocker, M. L., Langleben, D, D., Ruparel, K., Loughead, J. W., Gur, R. C., & Sachser, N. (2009). Baby schema in infant faces induces cuteness perception and motivation for caretaking in adults. *Ethology, 115*(3)*,* 257-263.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75–90.

Goldstein, W. M., & Hogarth, R. M. (1997). Judgment and decision research: Some historical context. In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (pp. 3-68). New York, NY: Cambridge University Press.

Gowans, C. (2015). Moral relativism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2015 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2015/entries/moral-relativism/

Green, S. (2015, Mar 6). Training police departments to be less biased. Harvard Business Review. https://hbr.org/

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (Vol. 3): *The neuroscience of morality* (pp. 35-81). Cambridge, MA: MIT Press.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review, 69*(4), 623-638.

Hahn, U., Frost, J., & Maio, G. (2005). What's in a heuristic? *Behavioral and Brain Sciences, 28*(4), 551-552.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814-834.

Haidt, J. (2013). The righteous mind: Why good people are divided by politics and religion. New York: Vintage Press.

Haidt, J., Bjorklund, F. & Murphy, S. (2004). Moral dumbfounding: When intuition finds no reason. Unpublished manuscript, University of Virginia.

Hammond, K.R. (1990). Functionalism and illusionism: Can integration be usefully achieved? In R.M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn*. Chicago: University of Chicago Press.

Hand, M. (2018). *A theory of moral education.* London: Routledge.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. London: Sage.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2)*,* 135-175.

Hovland, C. I., & Sears, R. R. (1940). Minor studies of aggression: Correlation of lynchings with economic indices. *Journal of Psychology, 9*(2)*,* 301–310.

Hubin, D. C. (2003). Desires, whims, and values. *The Journal of Ethics, 7*(3), 315-335.

Jia, H., Park, C. W., & Pol, G. (2015). Cuteness, nurturance, and implications for visual product design. In R. Batra, C. Seifert & D. Brei (Eds.), *The psychology of design: Creating consumer appeal* (pp. 168-179). New York, NY: Routledge.

Joyce, R. (2001). *The myth of morality*. Cambridge: Cambridge University Press.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel Prize Lecture, 8*, 351-401.

Kahneman, D. (2013). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). New York, NY: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982a). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982b). Preface. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. xi-xiii). New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-291.

Kahneman, D., & Tversky, A. (1982a). On the psychology of prediction. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 48-68). New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982b). On the study of statistical intuitions. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 493-508). New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982c). Subjective probability: A judgment of representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 32-47). New York, NY: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582-591.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames.* New York, NY: Cambridge University Press.

Kaplan, R. (2015, Mar 4). Eric Holder: “Implicit and explicit racial bias” in Ferguson policing. CBS News. Retrieved from http://www.cbsnews.com/

Kass, L. R. (1998). The wisdom of repugnance. In L. Kass & J. Q. Wilson (Eds.), *The ethics of human cloning*. Washington, DC: American Enterprise Institute.

Kauppinen, A. (2014). Moral Sentimentalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2014 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2014/entries/moral-sentimentalism/

Kellogg, R. T. (2002). *Cognitive psychology*. London, U.K.: Sage.

Keren, G., & Teigen, K. H. (2004). Yet another look at the heuristics and biases approach. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 89-109). Malden, MA: Blackwell.

Koehler, D. J., & Harvey, N. (Eds.). (2004). *Blackwell handbook of judgment and decision making*. Malden, MA: Blackwell.

Kohlberg, L. (1958). *The development of modes of moral thinking and choice in the years ten to sixteen* (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.

Kolodny, N., & Brunero, J. (2013). Instrumental rationality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2013 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2013/entries/rationality-instrumental/

Korsgaard, C. (1986). Skepticism about practical reason. *Journal of Philosophy, 83*(1), 5-25.

Kringelbach, M. L., Stark, E. A., Alexander, C., Bornstein, M. H., & Stein A. (2016). On cuteness: Unlocking the parental brain and beyond. *Trends in Cognitive Science, 20*(7)*,* 545-588.

Kripke, S. (1980). *Naming and necessity*. Oxford, UK: Basil Blackwell.

Lakatos, I. (2000). Falsification and the methodology of scientific research programs. In T. Schick, Jr. (Ed.), *Readings in the philosophy of science: From positivism to postmodernism* (pp. 20-25). Mountain View, CA: Mayfield.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316-338). Malden, MA: Blackwell.

Lerner, M., & Simmons, C. (1966). Observer's reaction to the 'innocent victim': Compassion or rejection? *Journal of Personality and Social Psychology, 4*(2), 203–210.

Lichtenstein, S., & Slovic, P. (2006). *The construction of preference*. New York, NY: Cambridge University Press.

Lifton, R. J. (1967). *Death in life: Survivors of Hiroshima*. New York: Random House.

Lindstrom, B, Jangard, S, Selbing, S., & Olsson A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General, 147*(2)*,* 228-242.

Little, D. (1991). *An introduction to the philosophy of social science.* Boulder, CO: Westview.

Lopes, L. L. (1991). The rhetoric of irrationality. *Theory and Psychology, 1*(1), 65-82.

Loughnan, S., & Piazza, J. (2018). Thinking morally about animals. In K. Gray & J. Graham (Eds.), *Atlas of moral psychology* (pp. 165-174). New York, NY: Guilford Press.

MacDonald, H. (2008). Is the criminal justice system racist? Spring 2008. Retrieved from https://www.city-journal.org/

Mallon, R. (2017). Social construction and achieving reference. *Nous, 51*(1)*,* 113-131.

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: W. H. Freeman.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

McKitrick, J. (2005). Are dispositions causally relevant? *Synthese, 144*, 357-371.

Mullainathan, S. (2015, Jan 3). Racial bias, even when we have good intentions. New York Times. Retrieved from http://www.nytimes.com/

Murphy, D. (2015). Concepts of disease and health. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2015 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2015/entries/health-disease/

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Over, D. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 3-18). Malden, MA: Blackwell.

Perring, C. (2010). Mental illness. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2010 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2010/entries/mental-illness/

Petersen, M. B. (2009). Public opinion and evolved heuristics: The role of category-based inference. *Journal of Cognition & Culture, 9*(3-4), 367-389.

Piaget, J. (1970). *Genetic epistemology*. New York, NY: Norton.

Piazza, J., McLatchie, N., Olesen, C., & Tomaszczuk, M. (2016). *Cuteness promotes moral standing and reduces appetite for meat among women*. [Unpublished raw data].

Prinz, J. (2007). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Putnam, H. (1975). The meaning of meaning. *Minnesota Studies in the Philosophy of Science, 7,* 215-271.

Radcliffe, E. S. (2012). Reasons from the Humean perspective. *The Philosophical Quarterly, 62*(249), 777-796.

Railton, P. (1986a). Facts and Values. *Philosophical Topics, 14*(2), 5-31.

Railton, P. (1986b). Moral realism. *Philosophical Review, 95*(2), 163-207.

Railton, P. (2007). Humean theory of practical reason. In D. Copp (Ed.), *The Oxford handbook of ethical theory*. New York: Oxford University Press.

Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics, 124*(4), 813-859.

Ramsey, P. F. (1928). Truth and probability. In R. B. Braithwaite (Ed.), *Foundations of mathematics and other logical essays.* London: Routledge & Kegan Paul.

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

Rawls, J. (1996). *Political liberalism.* New York: Columbia University Press.

Richardson, H. S. (1994). *Practical reasoning about final ends*. Cambridge: Cambridge University Press.

Rieskamp, J., & Otto, P.E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*(2), 207-236.

Ritov, I. (2005). Cognitive heuristics and deontological rules. *Behavioral and Brain Sciences, 28*(4), 558-560.

Rosati, C. (1996). Internalism and the good for a person. *Ethics, 106*(2), 297-326.

Samuels, R., Stich, S., & Faucher, L. (1999). Reason and rationality. Rutgers University Research Group on Evolution and Higher Education. Retrieved from http://ruccs.rutgers.edu/ArchiveFolder/Research%20Group/Publications/Reason/ReasonRationality.htm/

Samuels, R., Stich, S., & Bishop, L. (n.d.). Ending the rationality wars: How to make disputes about human rationality disappear. Rutgers University Research Group on Evolution and Higher Education. Retrieved from http://ruccs.rutgers.edu/ArchiveFolder/Research%20Group/Publications/Ending_the_rationality_wars.pdf/

Sauer, H. (2018). *Moral thinking fast and slow*. New York: Routledge.

Schneider, S. L., & Shanteau, J. (Eds.). (2003). *Emerging perspectives on judgment and decision research*. New York, NY: Cambridge University Press.

Selton, R. (2001). What is bounded rationality? In G. Gigerenzer & R. Selton (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 13-36). Cambridge, MA: MIT Press.

Sheehan, K. B., & Lee, J. (2014). What's cruel about cruelty free: An exploration of consumers, moral heuristics, and public policy. *Journal of Animal Ethics, 4*(2), 1-15.

Sherman, G. D., Haidt, J., & Coan, J. A. (2009). Viewing cute images increases behavioral carefulness. *Emotion, 9*(2), 282-286.

Simon, H. A. (1990). Alternative visions of rationality. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches*. New York: Cambridge University Press.

Singer, P. (2005a). Ethics and intuitions. *The Journal of Ethics, 9*(3-4), 331-352.

Singer, P. (2005b). Intuitions, heuristics, and utilitarianism. *Behavioral and Brain Sciences, 28*(4), 559-560.

Skinner, B. F. (1953). *Science and human behavior.* New York, NY: Macmillan.

Slovic, P. (2007). “If I look at the mass I will never act”: Psychic numbing and genocide. *Judgment and Decision Making, 2,* 79-95.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment (pp. 397-420). New York, NY: Cambridge University Press.

Small, D.A., & Loewenstein, G. (2003). Helping *a* victim or helping *the v*ictim: Altruism and identifiability. *Journal of Risk and Uncertainty, 26*(1), 5-16.

Small, D.A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes, 102*(2), 143-153.

Smith, M. (1995). *The moral problem*. Malden, MA: Wiley-Blackwell.

Sobel, D. (1994). Full information accounts of well-being. *Ethics, 104*(4), 784-810.

Sobel, D. (2001). Subjective accounts of reason for action. *Ethics, 111*(3), 461-492.

Sobel, D. (2009). Subjectivism and idealization. *Ethics, 119*(2), 336-352.

Sobel, D. (2017). *From value to valuing: Towards a defense of subjectivism.* New York: Oxford University Press.

Stanford, P. K., & Kitcher, P. (2000). Refining the causal theory of reference. *Philosophical Studies, 97*(1), 99-129.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. In *Advances in Child Development and Behavior* (pp. 251-285). Elsevier.

Stein, E. (1996). *Without good reason*. Oxford: Clarendon.

Stein, E. (2005). Wide reflective equilibrium as an answer to an objection to moral heuristics. *Behavioral and Brain Sciences, 28*(4), 560-561.

Sunstein, C.R. (2005). Moral heuristics. *Behavioral and Brain Sciences, 28*(4), 531-542.

Tetlock, P. E. (1997). An alternative metaphor in the study of judgment and choice: People as politicians. In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (pp. 657-680). New York, NY: Cambridge University Press.

Tetlock, P. E. (2005). Gauging the heuristic value of heuristics. *Behavioral and Brain Sciences, 28*(4), 561-562.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the rationality of choice. *Science, 211*(4481), 453-458.

Tversky, A., & Kahneman, D. (1982a). Availability: A heuristic for judging frequency and probability. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 163-178). New York, NY: Cambridge University Press.

Tversky, A., & Kahneman, D. (1982b). Judgments of and by representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84-99). New York, NY: Cambridge University Press.

Tversky, A., & Kahneman, D. (1982c). The belief in the “law of small numbers.” In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 23-31). New York, NY: Cambridge University Press.

Tversky, A., & Kahneman, D. (2002). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 19-48). New York, NY: Cambridge University Press.

Tversky, A., & Kahneman, D. (2003). Judgment under uncertainty: Heuristics and biases. In T. Connolly, H. R. Arkes, K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (2nd Ed.) (pp. 35-52). New York, NY: Cambridge University Press.

[Victor metal pedal rat traps M200, photograph]. (2019). Retrieved from https://store.doyourownpestcontrol.com/catalog/product/gallery/id/4099/image/7979/

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

Wallace, R. J. (2014). Practical Reason. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2014 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2014/entries/practical-reason/

Williams, B. (1981). Internal and external reasons. In B. Williams, *Moral Luck* (pp. 101-113). Cambridge: Cambridge University Press.

Wilson, T. D., Houston, C., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125*(4), 387.

Wong, D. B. (1984). *Moral relativity*. Berkeley, CA: University of California Press.

Wong, D. B. (2006). *Natural moralities: A defense of pluralistic relativism.* Oxford: Oxford University Press.

Wright, L. (1973). Functions. *The Philosophical Review, 82*(2), 139-168.

Zimbardo, P. (2007). *The Lucifer effect: Understanding how good people turn bad*. New York, NY: Random House.