Towards Enhancing Moral Agency through Subjective Moral Debiasing[1]

The capacity to act in accordance with one's morality (broadly construed) is constitutive of moral agency. This capacity can be undermined—in whole or in part—by for instance, hypnosis, addiction, or obsessive-compulsion. Another way this capacity can be undermined is through poor moral reasoning. Moral irrationality can frustrate one's capacity to act in accordance with one's morality and in turn, stunt one's moral agency. In a similar respect, improving moral rationality can strengthen this capacity and enhance moral agency. The empirical research program on (non-moral) *cognitive debiasing* inspires developing techniques to improve our moral rational capacities—i.e., *moral debiasing*. Yet, moral debiasing presupposes *moral biases*—that is, systematic *moral errors*. So, what are *moral errors* exactly? The pertinent kind is *subjective* moral errors. Ultimately, (adapting Connie Rosati's *two-tier internalism* regarding a person's good) A's φ-ing is a subjective moral error insofar as φ-ing deviates from A's *genuine* morality per *instrumental subjective moral rationality* (ISMR)—i.e., insofar as φ-ing frustrates A+'s morally-relevant ends, wherein A+ is a counterfactual idealization of A upon whom is bestowed those endowments that A considers authoritative under ordinary optimal conditions. The provision of an in-principle standard of subjective moral error lays important theoretical groundwork for future empirical inquiry into subjective moral debiasing.

The capacity to act in accordance with one's morality (broadly construed) is constitutive of moral agency. This capacity can be undermined—in whole or in part—by for instance, hypnosis, addiction, or obsessive-compulsion. Another way this capacity can be undermined is through poor moral reasoning. Morally irrationality can frustrate one's capacity to act in accordance with one's morality and in turn, stunt one's moral agency. In a similar respect, improving moral rationality can strengthen this capacity and in turn, enhance moral agency.[2]

How could moral rationality be improved? Perhaps we can draw from the empirical research program on improving *non-moral* rational capacities—namely, draw from the work on

[1] I am grateful to…for helpful discussions.

[2] (Moral) agency and rationality come in degrees.

cognitive debiasing.[3] *Cognitive biases* are systematic errors in judgment.[4] For example, we tend to overestimate the probability of dramatic events, such as airplane crashes (*availability bias*).[5] Cognitive debiasing employs techniques to reduce such errors. For instance, doctors tended to drastically overestimate the likelihood of disease-presence given positive test results when base-rates were influential (*base-rate neglect*).[6] Such errors are reduced by training doctors to re-represent probabilities in terms of natural frequencies—for example, (a) "If a woman does not have breast cancer, the probability that she will get a positive mammography is 9.6%," re-represented as (b) "Out of every 1,000 women without breast cancer, 96 will get a positive mammography."[7]

Perhaps we can adapt cognitive debiasing to the moral domain and yield empirically-supported techniques to improve our moral rational capacities—i.e., yield (what we can call) *moral debiasing*. Such moral debiasing presupposes *moral biases*—that is, systematic *moral errors*. Thus, the adequacy of moral debiasing depends upon the adequacy of this notion of *moral error*. So, what might such *moral errors* be exactly?

Recall that the morality in question is *one's* morality (as opposed to perhaps, *the true, best, and/or real* morality). In this respect, the morality in question is *subjective morality*. It is worth noting that accordance-with-subjective-morality constitutes merely a metric or standard of

[3] For overviews, see Larrick (2004) and Sanna (2013). For more familiar examples, many of the interventions advocated in *Nudge* (Thaler & Sunstein, 2008) constitute debiasing techniques.

[4] See e.g., Tversky & Kahneman (1974), Kahneman, Slovic, & Tversky (1982), Kahneman & Tversky (2000), Gilovich, Griffin, & Kahneman (2002), and Kahneman (2013). This sense of *bias* presupposes the cognitive psychological conceptual framework (or ontology) of inputs, cognitive processes, and outputs (e.g., judgments) (Marr, 1983, Cummins, 1983). *Bias* in the sense of systematic error refers to (a pattern within) a set of *outputs*. While a cognitive process may be *biased* in that it would yield a bias, a process cannot constitute a bias. While a bias can explain a judgment (*a la* Hempel & Oppenheim, 1948; *contra* Cummins, 2000), it cannot cause a judgment. In addition, this sense of *bias* differs from the sense that regards discrimination, such as that of *implicit biases* (e.g., Banaji & Greenwald, 2013), although cases may be instances of both.

[5] Tversky & Kahneman (1982).

[6] Casscells, Schoenberger & Grayboys (1978); Eddy (1982).

[7] Gigerenzer & Hoffrage (1995, p. 688); Sedlmeier (1997).

interest. Just as (a) interest in the extent to which a public policy maximizes utility does not entail endorsing utilitarianism, (b) interest in accordance-with-subjective-morality does not entail endorsing the *moral theory*, moral subjectivism (i.e., individual moral relativism or speaker subjectivism).

The rationality in question regards acting in accordance with one's (subjective) morality. We can formulate this as acting in accordance with one's subjective moral ends. In this respect, the rationality in question is a form of means-ends rationality that we can call *instrumental (subjective) moral rationality* (IMR). Acts (or decisions, judgments, or outputs) that deviate from those dictated by IMR constitute *subjective moral errors*. We can formalize such errors with the following provisional model:

Simple Model of Subjective Moral Errors: An agent A's performing some action φ is a subjective moral error insofar as φ-ing violates A's morality—i.e., frustrates A's moral ends.

For an example of a subjective moral error, suppose Rachael opposes racial discrimination—that is, she supports racial egalitarianism. Being racially egalitarian 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 racial 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 her moral end, being racially egalitarian; her doing so violates her morality. As such, Rachael's giving the job to Katie is a subjective moral error.

As this illustrates, one's actions can violate one's own morality. The most straightforward way to do this is to fail to identify the action that constitutes the means to one's moral ends. A simple way in which this can occur is by employing fallacious reasoning (e.g., affirming the consequent) when determining what action to take.[8]

Another way of violating one's own morality—a way with significant 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.[9] One manifestation of this consideration regards *internal reasons*. Internal reasons are reasons grounded in an agent's desires.[10] *Internal* reasons contrast with *external* reasons, which are grounded elsewhere, such as in objective goodness. Here's a *simple* model:

Simple Model of Internal Reasons: A has an internal reason to φ iff A *possesses* some desire whose satisfaction will be served by φ-ing.

This model of internal reasons is too simple, as it cannot handle cases of false beliefs. For example, suppose Gaston is thirsty.[11] 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 reductio of the simple model.

[8] In this respect, subjective moral errors and biases do not require generation by moral-domain-specific processes.

[9] E.g., Brandt (1979), Railton (1986a;1986b), Rosati (1996), Sobel (2017), Smith (1995), and Williams (1981).

[10] More precisely, internal normative reasons are distinguished by their being (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.

[11] *A la* Williams (1981, p. 78).

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*[12] that in turn, yielded 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.[13] 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 sentencing innocent Iona to death is not a subjective moral error is a reductio of the simple model of subjective moral error (i.e., a reductio of: A's φ-ing is a subjective moral error insofar as φ-ing violates A's morality—i.e., frustrates A's moral ends).

More refined models of internal reasons avoid such reductios by (a) utilizing counterfactual idealizations of the agent (e.g., endowing the agent with omniscience and perfect rationality) and (b) grounding internal reasons in the idealized agent's desires[14]. This yields:

[12] Railton (2007, p. 267).

[13] 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."

[14] For example, Williams (1981).

Idealized Agent Model of Internal Reasons: A has an internal reason to φ iff *A+ (i.e., idealized agent A)* possesses some desire whose satisfaction would be served by φ-ing.[15]

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.

Regarding subjective moral error, a sensible way to avoid analogous reductios (e.g., Justine's sentencing innocent Iona to death not being a subjective moral error) is to likewise employ counterfactual idealization. This is sensible because while subjective moral error (and IMR) concerns a person's moral ends, it (should) only concern 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 (broadly construed). In this respect, a moral end of A is *genuine* insofar as it would "survive the idealization" of A. This yields:

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 A+'s moral ends).

So, with respect to yielding an A+ regarding subjective moral error, what are the right counterfactual idealizations—i.e., the right *endowments* for A+? Before jumping into this question, it will be worthwhile to address some preliminaries.

[15] The "*idealized agent* model" is sometimes called the "*deliberative* model" (Arkonovich, 2011).

Firstly, while there is an established canon of such idealizations regarding subjectivist value theory and practical reason,[16] unfortunately, there is a paucity of philosophical work done on idealization regarding subjective morality (e.g., on idealization options regarding speaker subjectivism).[17] As such, much of the following will draw from adapting work in value theory and practical reason.

Secondly, an important tension overhangs 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 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[19]). 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 connection yield reasons,

[16] E.g., Brandt (1979), Railton (1986a; 1986b), Rosati (1996), Sobel (2017), Smith (1995), and Williams (1981).

[17] Some moral relativists do appeal to idealizations. In this respect, there is philosophical work that *utilizes* idealizations of 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 with (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) subjective *standards/metrics* and *subjective constructs of interest*. I'll just stick with the more reader-friendly normative version.

[19] For example, Finlay & Schroder (2012, 1.2.3), Railton (1986a, p. 46), and Rosati (1996).

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.

Having addressed these preliminaries, we can return to the question of what counterfactual idealizations are the right ones for yielding an A+ regarding subjective moral errors. 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" should these endowments be exploited to yield genuine ends? (e.g., by identifying 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.

One type of endowment is *informational endowments*. Due to space constraints, I will simply declare that the most promising informational endowment for subjective moral errors is *knowledge of the relevant (non-normative) facts*.[20]

Another type of endowment is *rationality endowments*. An important condition of rationality is preference-coherence (which includes consistency and arguably, unity).[21] Human beings often lack preference-coherence, including amongst their moral preferences. This 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 are stable,[22] known,[23] relevant,[24]

[20] *A la* Brandt (1969-1970, p.45-46); Kolodny & Brunero (2013).

[21] Regarding *unity*, see Joyce, 2001, p. 71. Per the conventions of the practical reason literature, I previously used the term, "desires," as a stand-in for *contingent conative set*, etc. (n. 4). 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.

[22] Zimbardo (2007).

[23] Lichtenstein & Slovic (2006).

[24] Haidt (2001).

accessible,[25] or present in any coherent form.[26] Achieving preference-coherence can require revising moral preferences.

The extent of revision required to redress coherence deficits varies. On the low-revision end of the spectrum is redressing inchoate (moral) 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.[27] For example, clarifying one's weighting of equality vis-à-vis liberty. On the high-revision end of the spectrum is redressing intransitive preferences. 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 non-equivalent solutions.

Especially when preference-incoherence cannot be resolved by mere clarification, one option is to privilege certain types of preferences. For instance, privileging second-order preferences above first-order preferences.[28] This is an attractive option as it provides a mechanism for discounting (arguably) less authoritative preferences, such as those that stem from akrasia, addiction, impulses, or whims.[29] Nevertheless, preferences 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 moral preferences according to the types of cognitive processes that yield them. For example, one might privilege moral preferences from deliberative (type 2) processes over those from intuitive (type 1) processes[30]—or at least, discount moral preferences that lack deliberative

[25] Fischhoff (1991).
[26] Churchland (1996).
[27] Milgram (1996, p. 504).
[28] *A la value-based Humeanism* (Radcliffe, 2012, p. 783).
[29] Hubin (2003).
[30] For example, Greene (2007), Singer (2005); *contra* Bartsch & Wright (2005), Kass (1998), and Railton (2014).

endorsement (or lack principled endorsement, *a la moral dumbfounding*[31]). 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 type of endowment regards the agent's state-of-mind or mental condition. Without such an endowment, the idealized agent would (presumably) inherit the state-of-mind of the agent at the time of the potential subjective moral error. This could include inheriting emotional disturbance, obsessive compulsion, mania, and hypnosis, and leave consequent preferences intact.[32] As such, a mental condition endowment is necessary. One such endowment is "psychological normalcy."[33] However, would that exclude all preferences traceable to conditions classified as disorders in the DSM?[34] Should it? Should preferences traceable to inflamed passions be excluded?[35] A common endowment is a state-of-calm.[36] 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—namely, care, fairness, loyalty, authority, sanctity/purity, and liberty.[37] Another endowment option is bestowing the "highest" stage of

Readers may be familiar with *type* 1 and 2 processes in terms of their theoretical predecessors, *system* 1 and 2 processes. For more on this development within dual-processing theory in cognitive psychology, see Evans & Stanovich (2013).

[31] Haidt, Bjorklund & Murphy (2004).

[32] Rosati (1996, p. 302).

[33] Bjorklund, Bjornsson, Eriksson, et. al (2012, p. 125-127).

[34] American Psychiatric Association (2013).

[35] *A la* Brandt (1943, p. 487) and Rawls (1971, p. 47).

[36] For example, Rosati (1996, p. 305), Sobel (1994, p. 791), and Wallace (2014, §5).

[37] Haidt (2013).

Kohlbergian moral development.[38] Other options include bestowing impartiality (*a la* the ideal observer[39]), moral maturity,[40] selflessness,[41] and open-mindedness.[42]

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 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. The answer is (an adaption of) Connie Rosati's *two-tier internalism*.[43] Per such, 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 that endowment authoritative. This mechanism avoids alienation. Adapting this solution to subjective moral error yields: the endowments of A+ are those the agent would consider morally authoritative (under ordinary optimal conditions).[44]

[38] Kohlberg (1958); *contra* Gilligan (1982).

[39] Firth (1952).

[40] Bartsch & Wright (2005, p. 546); Brandt (1943, p. 486).

[41] Daniels (1979, p. 270).

[42] Richardson (1994, p. 31).

[43] Rosati's (1996) model regards a person's good, though it can be adapted for other applications. It includes several modifications to the ideal advisor model that are left aside as they do not apply to subjective moral errors—e.g., avoiding the conditional fallacy (Railton, 1986a, p. 53; Finlay & Schroeder, 2012, §2.4) -i.e., fragile reasons (Sobel, 2001)- and indirection (Rosati, 1996, p. 304). Rosati's internalism 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 for instance, an objective list), and (2) the makeup of that idealized agent—i.e., the idealization conditions or endowments—are also grounded in that agent (as opposed to declared from "outside").

[44] The appeal to *ordinary optimal conditions* avoids an infinite-regress of the authoritative endowments being determined by an idealized agent whose idealization is determined by an idealized agent, whose idealization is determined by an idealized agent, *ad infinitum* (Rosati, 1996, p. 305).

We are now in a position to state the final model of subjective moral error:

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

This model of subjective moral errors provides a solid foundation upon which to build. For instance, such errors, when systematic, instantiate *subjective moral biases*. Techniques to reduce such instantiations constitute *subjective moral debiasing*. Successful techniques would improve moral rationality, strengthen one's capacity to act in accordance with their morality, and enhance moral agency.

Additional work remains before subjective moral debiasing could be pursued via empirical research. For example, figuring out how to operationalize A+'s moral ends (though some tools might be harvested from the *deliberative model* of medical consultation[45]). Nonetheless, the provision of an in-principle standard of subjective moral error lays important theoretical groundwork for future empirical inquiry into subjective moral debiasing.

[45] Emanuel & Emanuel (1992).

References

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

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

Baghramian, M., & Carter, J. A. (2015). Relativism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from: http://plato.stanford.edu.

Bagnoli, C. (2017). Constructivism in metaethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from: http://plato.stanford.edu.

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

Bartsch, K. & Wright, J. C. (2005). Towards an intuitionist account of moral development. *Behavioral and Brain Sciences*, 28 (4), 545-546.

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

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

Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299, 999–1000.

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.

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.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

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.

Finlay, S., & Schroeder, M. (2012). Reasons for action: Internal vs. external. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from: http://plato.stanford.edu.

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

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

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684-704.

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

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

Gowans, C. (2015). Moral relativism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.

Retrieved from: http://plato.stanford.edu.

Greene, J. (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.

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

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

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

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

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

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

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

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.

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

Kolodny, N., & Brunero, J. (2013). Instrumental rationality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from: http://plato.stanford.edu.

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

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.

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

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

Millgram, E. (1996). Reviewed work: Practical reasoning about final ends by Henry Richardson. *Mind*, 105 (419), 504-506.

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

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.

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

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

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

Sanna, L. J. (2013). Debiasing. In H. Pahler (Ed.), *Encyclopedia of the mind.* California: Sage Publications Inc.

Sedlmeier, P. (1997). BasicBayes: A tutor system for simple Bayesian inference. *Behavior Research Methods, Instruments, & Computers*, 29, 328–336.

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

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

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

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.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT.: Yale University Press.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1982). 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.

Wallace, R. J. (2014). Practical Reason. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from: http://plato.stanford.edu.

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

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.

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