
MEDITATIONS

THE UNDERGRADUATE JOURNAL OF PHILOSOPHY AT THE
UNIVERSITY OF CALIFORNIA, LOS ANGELES

SPRING 2025

ISSUE 12



MEDITATIONS

ISSUE 12

§

SPRING 2025

DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF CALIFORNIA, LOS ANGELES

MEDITATIONS

σύννοιαι

THE UNDERGRADUATE JOURNAL OF PHILOSOPHY AT
UCLA

ISSUE 12, SPRING 2025

Editor-in-Chief

Logan Cary

Editorial Staff

Kiran Thomas

Lucy Liu

Isabella Lam

Kiaris Charalampos

Journal Advisor & Design Editor Brandon Ward

Departmental Staff Assistants Ashna Madni

Editor's Letter

Dear Reader,

Our editors and authors have worked diligently to produce this edition of the *Meditations*. To the editors, I want to thank you all, Kiran, Lucy, Isabella, and Harris, for all of your time and effort you so graciously dedicated to this journal. It has been my honor to work with each and every one of you. To the authors, Steve, Will, Andrew, and Yiannos, on behalf of the editors, it has been our great joy getting to work and philosophize with you. And to our readers, I hope this journal, caringly and painstakingly crafted by each of our authors and editors, provides a space to meditate. A space to meditate on something new, on something insightful, and on something that I hope can provide a bit of respite from these trying times.

Without further ado, it is my pleasure to introduce the twelfth edition of *Meditations*.

Best,
LOGAN CARY
Editor-in-Chief

Contents

Knowing v. Deriving: The Formula of Humanity in Kant's *Critique of Practical*

Reason 1

STEVE NAM, *Brown University*

1	Introduction	1
2	Context to How the FH “follows of itself” from God’s Final End	2
3	How the FH “follows of itself” from God’s Final End	5
4	The Problem of Heteronomy and a Solution	9
5	Conclusion	14

Aristotle on the Nature of Our Intellect 17

ANDREW DE LA CRUZ, *The University of Utah*

1	Introduction	17
2	Aristotle’s <i>De Anima</i>	18
3	Caston on Theology and the Human Intellect	23
4	Gerson on Functional Unity and the Intellect	26
5	On Divine Thought and Human Particularity	30
6	Within Present Conditions	35
7	Final Words	38

Wagers, Bribes, and Bets: A Defense of Moderate Pragmatism 40

WILLIAM HASSETT, *Brown University*

1	Introduction	40
2	Motivating Pragmatic Encroachment	41
3	The Case Against Moderate Pragmatism	46
4	Broader Implications	53

Sophisticated Consequentialism and the Problem of Alienation 58

IOANNIS KRIZOPOULOS, *Franklin and Marshall College*

1	Introduction	58
2	Railton’s Account of Character Consequentialism	61
3	Juan and Linda	62
4	Juan’s Moral Dilemma	65
5	Conclusion	69
1	Introduction	70

2	Acknowledging Utility Maximization Without Directly Aiming for It	70
3	Juan as a Consequentialist Without Recognizing Consequentialism	72
4	Refuting Juan's Care as Merely Instrumental	74
5	Conclusion	75

KNOWING V. DERIVING: THE FORMULA OF HUMANITY IN KANT'S *CRITIQUE OF PRACTICAL REASON*

STEVE NAM

Brown University

1 Introduction

Immanuel Kant's Formula of Humanity, introduced in his *Groundwork of the Metaphysics of Morals*, makes an appearance in the *Critique of Practical Reason*. In the *Critique*, Kant determines that the highest good for rational beings is virtue and proportionate happiness. He then argues that the possibility of the highest good necessitates three postulates: the immortality of the soul, the existence of God, and freedom.¹ In Chapter II.V of the *Critique*, Kant explains why we must postulate the existence of God, but he concludes the chapter with a surprising assertion: "It now follows of itself that in the order of ends the human being...is an *end in itself*" (KpV 5:131).² Those familiar with Kant's ethics will recognize this as the Formula of Humanity (FH) formulation of the Categorical Imperative.

This paper will investigate the precise way in which the FH "follows of itself" from Kant's discussion of God. I argue for two points: (1) The FH "follows of itself" from Kant's discussion of God insofar as one can

1. One might argue that the postulates most relevant to the highest good are the first two, while the postulate of freedom is more relevant to the possibility of moral conduct in general.

2. References to works by Kant feature an abbreviation of the title followed by volume and page numbers from the "Akademie" edition of Kant's works. Abbreviations are as follows: KpV for the *Critique of Practical Reason*; G for the *Groundwork of the Metaphysics of Morals*. All translations are from Immanuel Kant, *Critique of Practical Reason*, trans by Mary Gregor, 2nd ed. of *Cambridge Texts in the History of Philosophy* (Cambridge University Press, 2015). and Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans by Mary Gregor and Jens Timmerman, 2nd ed. of *Cambridge Texts in the History of Philosophy* (Cambridge University Press 2012).

obtain the FH by reasoning from the claim that God's "final end in creating the world" is the highest good (KpV 5:130). I then suggest that my interpretation, in which God seems to have become the source of a supreme law of morality (i.e. the FH), raises the worry of heteronomy. This would be in tension with Kant's principle of the autonomy of the will, on which his ethical theory fundamentally rests. To resolve this tension, I argue that (2) the claim that the FH "follows of itself" should be understood as presenting a way of *knowing* the FH in the newly introduced framework of God, not as a derivation of the FH from it. A derivation of the FH would suggest that the existence of God is the *grounds* of the FH as a valid moral principle, in the sense that the FH has been given to rational beings by God and that rational beings subject themselves to the FH for this reason. This would be heteronomous. In my interpretation, Kant means only that the existence of God provides an alternative way of coming to *know* the FH, without claiming that God's existence is the grounds of the FH as a valid moral principle. My interpretation thus avoids the issue of heteronomy.

I proceed as follows. In section 2, I provide the context required for my interpretation of how the FH "follows of itself" from the existence of God and his final end. In section 3, I show how the FH "follows of itself." In section 4, I raise the worry of heteronomy in my interpretation, and I resolve it by arguing that the "following of itself" should be understood as presenting a way of knowing the FH, not as a derivation of it.

2 **Context to How the FH "follows of itself" from God's Final End**

The *Critique's* statement of the FH comes at the very end of Chapter II.V:

It now follows of itself that in the order of ends the human being (and with him every rational being) is an *end in itself*, that is, can never be used merely as a means by anyone (not even by God) without being at the same time himself an end, and that humanity in our person must, accordingly, be holy to

ourselves. (KpV 5:131-2)

Kant does little to explain exactly how the FH follows, and it is unclear which parts of the preceding discussion are relevant. My interpretation will show that the FH follows from the paragraph just prior, which begins: “From this it can also be seen that if one asks about *God’s final end* in creating the world, one must not name the *happiness* of the rational beings in it *but the highest good*” (KpV 5:130). Specifically, I show that the FH follows from the fact that God’s final end in creating the world is the highest good.

The preliminary step to my interpretation is to provide some context regarding Kant’s passage about God’s final end. The passage is found towards the end of Chapter II.V. In the beginning of the chapter, Kant first explains why we must postulate the existence of God. He reasons roughly as follows. The moral law commands that we have as our object the highest good (KpV 5:125). For Kant, this means that the highest good must be possible.³ The highest good is virtue met with proportionate happiness. Moreover, Kant thinks that there must be a necessary connection between virtue and proportionate happiness: virtue must be met with proportionate happiness *no matter what*.⁴ However, no such connection exists in the natural world. Rational beings do not always attain happiness in perfect proportion to their virtue, and it is not difficult to imagine how vicious people may be happier than virtuous people. So, for the highest good to be possible, one must postulate a cause from beyond the natural world which will provide such a necessary connection. That is, one must postulate the existence of God, who will allot happiness to rational beings according to their virtue (KpV 5:125).

According to Kant, God’s final end in creating the world is not just

3. This follows from the principle informally known as “ought implies can.” In the KpV, Kant’s endorsement of the principle can be seen in the gallows-man example, where he writes, “He judges, therefore, that he can do something because he is aware that he ought to do it...” (KpV 5:30). Other examples from the KpV appear in 5:114 and 5:122.

4. Another way to understand this *necessary* connection is that virtue must be met with proportionate happiness such that this connection *could not have been otherwise*. Contrast this to a contingent connection: snow is falling from the clouds right now (assume this is veridical). The clouds are certainly connected with its snowfall, but only contingently so. It certainly *could have been* the case that it were raining now instead.

the happiness of rational beings in it, but the highest good – the virtue and proportionate happiness of rational beings. As mentioned above, my interpretation shows that the FH “follows of itself” from this claim.

However, my interpretation requires drawing a detail from Kant’s reasoning for it. Therefore, for context, I provide a brief outline of Kant’s argument for the claim that God’s final end is the highest good.

The first step of the outline is to note the specific way in which Kant puts his conclusion: God’s final end in creating the world is *not the happiness of rational beings in it*, but the highest good (KpV 5:130). The fact that Kant includes the negative claim (that God’s final end is *not* the happiness of rational beings) sheds light on his argumentative strategy. Kant seems to be working within a framework in which the assumption *was* that God’s final end in creating the world is the happiness of rational beings (as suggested below, this may have been attributed to God’s beneficence). Kant’s strategy, then, is to accept this assumption but argue that a further component must be added to such happiness. The resulting combination turns out to be the highest good.

The main thrust of Kant’s argument lies in his conception of the highest good itself. Above, I mentioned that the two parts of the highest good are virtue and proportionate happiness. The more specific relationship between the two parts is that virtue is the necessary condition of happiness.⁵ For a rational being to attain happiness, they must first attain virtue. Moreover, since the highest good requires the attainment of both virtue and happiness, virtue also becomes the necessary condition of the highest good. This understanding of the highest good is essential for Kant in arguing that God’s final end is the highest good. He reasons as follows. God is the “highest independent wisdom,” from which it follows that he has cognized the highest good (KpV 5:130-1). That is, God knows what the highest good consists of, including its necessary condition, virtue. If so, it would be irrational for God to have as his end merely the happiness of rational beings without regard for its necessary condition, virtue. As Kant puts it, one cannot

5. See especially KpV 5:110 and 5:122.

conceive of the “effect of this beneficence” (i.e. the happiness of rational beings) as befitting God except under the “limiting condition” that the rational beings first attain virtue (KpV 5:131). Whatever happiness of rational beings God wills from being beneficent, he must also will the satisfaction of its limiting condition, namely the virtue of rational beings. Thus, God’s final end in creating the world cannot simply be the happiness of rational beings. Instead, it must be their happiness in addition to the satisfaction of its necessary condition, virtue. That is, God’s final end in creating the world is the highest good.

I mentioned above that the main thrust of the foregoing argument is Kant’s conception of the highest good itself – that virtue is the necessary condition of the highest good. As shown in the next section, this same detail is essential context for my interpretation of how the FH “follows of itself” from God’s final end.

3 How the FH “follows of itself” from God’s Final End

In this section, I present a way in which the FH “follows of itself.” Before proceeding, readers should note that I do not comment on the precise nature of the “following of itself” here. Later, in section 4, I argue that the “following of itself” should be understood as presenting a way of *knowing* the FH in the framework of God, rather than a derivation of the FH from it. The latter would suggest that the existence of God is the grounds of the FH as a valid moral principle, in the sense that God is in fact the source of the FH and rational beings subject themselves to it for this reason. The former, on the other hand, only suggests that the existence of God is an alternative way of coming to *know* the FH, without making any claims about the grounds of its validity as a moral principle. The reason for this distinction will become clear in the next section. This section first presents my interpretation of how the FH “follows of itself” from God’s final end (where this “following” is understood as presenting a way of *knowing* the FH from the existence of God).

Let us start by reviewing Kant’s introduction of the FH in the

Critique. He says, "...that is, [the human being] can never be used merely as a means by anyone (not even by God) without being at the same time himself an end" (KpV 5:131). If God attempted to use human beings as a mere means, *even he* would fail to do so. God must treat human beings at the same time as ends. I suggest that there are two key details in Kant's ensuing remarks that lead to a plausible interpretation of how the FH "follows of itself." Kant says that (1) the human being is the *subject of the moral law*, and (2) the human being is the subject of the moral law *by agreeing* to the moral law (KpV 5:132). With help from the *Groundwork*, the foregoing two points can show that, if God's final end is the highest good, human beings can never be used merely as a means, not even by God.

First, I turn to the *Groundwork* to draw the tools required for my interpretation. In the *Groundwork*, Kant discusses the notion of *agreement* in relation to the FH. After introducing the FH, he gives several examples wherein the FH is violated – that is, examples where one treats another as a mere means. The lying promise example is of particular interest, for in it Kant uses a metric to determine whether a rational being is treating another merely as a means or at the same time as an end (as well as a means). The metric is this: a rational being treats another at the same time as an end (and not merely as a means) if the latter is able to "contain" the end of the former's action (G 4:430). Moreover, Kant specifies a way in which a rational being can "contain" the end of another's action. He says, "For the one I want to use for my purposes by such a promise cannot possibly agree to my way of proceeding with him and thus himself contain the end of this action" (G 4:429-30). A rational being can contain the end of another's action if they can *agree* to the action (or "way of proceeding" with them).⁶ Thus, a rational being is treated as an end in themselves if they can contain the end of the other's action, and they can contain the end of the other's action if they can *agree* to the other's action.

Let us now return to the *Critique's* discussion of God to see how the

6. For a more detailed discussion of this "metric" of the FH, see Christine Korsgaard, "The Right to Lie: Kant on dealing with evil," In *Creating the Kingdom of Ends*, 133-158 (Cambridge University Press, 2015): 137.

FH “follows of itself.” Kant has argued that God’s final end in creating the world is the highest good. Recall a premise from this argument: the virtue of rational beings is the necessary condition of the highest good. Now, if God has cognized the highest good himself and therefore knows its necessary condition, then God, in willing the highest good, should also will the satisfaction of its necessary condition: the attainment of virtue by rational beings. Kant himself would have agreed with this line of reasoning. In the *Groundwork*, he says, “Whoever wills the end also wills (in so far as reason has a decisive influence on his actions) the indispensably necessary means to it that is in his control” (G 4:417). In this case, the “indispensably necessary means” to God’s end would be the attainment of virtue by human beings.

However, there is a potentially troublesome detail in Kant’s remark above. Kant says that whoever wills the end also wills the means to that end that are *in his control*. One might object that it is unclear whether the virtue of rational beings is in God’s control. After all, a central principle in Kant’s ethical theory is that moral duties are provided solely by the autonomy of their will, not by some external authority (like God). At the same time, there are clear moments in which Kant seems to grant precisely such authority to God. For instance, before introducing the FH in the *Critique*, Kant characterizes God as *commanding* the moral law to human beings: “...respect for his command, observance of the holy duty that his law lays upon us” (KpV 5:131). Here, moral duties are *laid upon* human beings by God. Such characterizations would seem to suggest that God does in fact demand or control the moral conduct of human beings. Thus, there appears to be a tension between Kant’s descriptions of God and his principle of the autonomy of the will. This is the tension to which I turn in section 4. In this section, I first wish to focus on how the FH “follows of itself” from God’s final end. For that purpose, I suggest that we accept Kant’s characterizations of God at face value and stipulate that God does have the capacity to command the moral law to rational beings. Doing so will allow us to see how the FH “follows of itself” from the existence of God, and, as I show in section IV, the tension with autonomy

can be resolved without undoing this stipulation, leaving my interpretation of how the FH “follows of itself” intact.

With this stipulation, an interesting picture emerges. God wills the highest good as his end. God also wills the “indispensably necessary means” to it, or the attainment of virtue by human beings. Further, God commands that human beings conform to the moral law. In so-commanding, is God not using human beings as a means to his end?

I argue that in such a case God must be treating human beings at the same time as ends, not merely as a means. Recall the two details from the remarks following Kant’s introduction of the FH: (1) the human being is the *subject of the moral law*, and (2) the human being is the subject of the moral law *by agreeing* to the moral law. For Kant, (2) follows from the autonomy of the will. Kant thinks that a law can hold unconditionally – as the moral law does – only if the will has autonomy. That is, only if the will is viewed as itself having given the universal (moral) laws to which it is subject (G 4:432-3).⁷ Now, recall that God commands the moral law to human beings. Above, this raised the suspicion that God is treating human beings as a means to his end of the highest good. However, (1) and (2) provide that human beings cannot be subject to the moral law without also agreeing to it. Thus, if God commands the moral law to human beings and if human beings are therefore subject to it, they must also agree to it. In this way, human beings, by virtue of their autonomy, are always able to agree to God’s “way of proceeding” with them in his pursuit of the highest good (i.e., commanding the moral law). According to the earlier discussion of the *Groundwork*, then, human beings are always able to *contain* the end of God’s action. According to the same discussion, human beings are being deemed by God as ends in themselves. In this way, if God’s final end is the highest good, human beings are treated as ends in themselves. Not even God can use them as a mere means. This is the way in which the FH “follows of itself” from the

7. And a will agrees with a universal (moral) law it has itself given, which explains the claim in (2) that the subject of the moral law agrees to the moral law. Explanations of what Kant means by an “unconditional” law and why autonomy is required for this are provided later in the paper, when I discuss autonomy more directly.

existence of God and his final end.

4 The Problem of Heteronomy and a Solution

An interesting feature of this interpretation is that it is the *autonomy* of the human will that limits God's ability to use them as a mere means.

Autonomy ensures that human beings are always able to agree to God's "way of proceeding" with them, and therefore that human beings are always treated at the same time as ends. However, there remains an unresolved tension. The spotlight shone on autonomy appears to have been accompanied by a clear violation of autonomy. In concluding that God treats human beings as ends in themselves, I stipulated that God commands the moral law. Few things seem more antithetical to the autonomy of the will than God commanding the moral law. Yet my interpretation accepts both horns.

In addition to granting the premise that God commands the moral law, my interpretation raises the worry of heteronomy more generally. Namely, the claim that the FH, a supreme principle of morality, "follows of itself" from the existence of God and his final end may raise suspicions that God has become the grounds of the FH as a valid moral principle. As I explain below, this would be heteronomous. In this section, I suggest a way to resolve these tensions. I do so by employing a distinction between *knowing* and *deriving*. Specifically, I argue that the "following of itself" should be understood as presenting a way of *knowing* the FH in the framework of God, not as a derivation of the FH from it.

I begin by explaining the worry of heteronomy and why it is a worry at all. Heteronomy is an issue for Kant because of how central its opposite, autonomy, is for his ethical theory. In the *Groundwork*, he describes autonomy as follows:

Thus the will is not just subject to the law, but subject in such a way that it must also be viewed as self-legislating, and just on account of this as subject to the law (of which it can consider itself the author) in the first place. (G 4:431)

Autonomy is the idea that the will of a rational being must be viewed as *self-legislating*. That is, the will of a rational being must be viewed as itself giving universal laws and only for that reason being subject to the law (which it itself has given). Heteronomy, then, is a case in which a “law”⁸ has been given to a rational being in any way other than the will’s own legislation (i.e. in any non-autonomous way) (G 4:433). Moreover, Kant thinks that a rational being who subjects themselves to a heteronomous “law” does so by virtue of some interest (G 4:432-3). For instance, a rational being who subjects themselves to a “law” given heteronomously by a powerful ruler does so only because they have an interest in reaping the benefits provided by a powerful ruler. In a case of autonomy – that is, in a case where the law has been given by the will itself – such an interest is not required.⁹

The importance of autonomy (and thus the worry of heteronomy) for Kant lies in the fact that autonomy is the precondition for morality itself. If autonomy is not possible, then the concept of morality as a whole is merely an illusion. In the *Groundwork*, Kant explains that moral imperatives¹⁰ must be categorical imperatives, meaning that they must be imperatives which rational beings are compelled to obey no matter the circumstances (G 4:433). By contrast, imperatives which depend on the inclinations of rational beings for their effectiveness (much like heteronomous “laws”) are not categorical.¹¹ It follows that if categorical imperatives are impossible, then morality becomes an illusion for Kant. The possibility of categorical imperatives, in turn, depends on the autonomy of the will.¹² The stakes for Kant are therefore clear and

8. “Law” is in quotation marks because Kant seems to reserve the term for *unconditional* laws (also called categorical imperatives; G 4:416) which, unlike heteronomous “laws,” do not depend on the interests of a rational being in order for rational beings to be subject to them. This latter point is explained later in the paper, in conjunction with footnote 12.

9. Explanation as to why this is so appears later in the paper, in conjunction with footnote 12.

10. It is helpful to think of these as “ought”-commands. E.g. “You ought not lie” is an imperative.

11. For instance, the imperative “You ought to study in order to pass the exam” is not categorical, for it may lose its normative force for someone who no longer wishes (i.e. has no inclination) to pass the exam. Moral imperatives should be categorical in that they cannot be “canceled” in this way.

12. The possibility of categorical imperatives depends on autonomy because Kant thinks that an imperative can be categorical only if a rational being does not depend on their interests or inclinations in being subject to it (G 4:432-3). (Note that “categorical imperative” and “law” are interchangeable, cf. footnote

significant: morality is possible only if autonomy is. A subversion of autonomy – that is, heteronomy – would mean the loss of morality itself.

Now, if one takes the claim that the FH “follows of itself” from God’s existence to mean that God is the grounds of the FH as a valid moral principle, in the sense that the FH has been given to rational beings by God and that rational beings subject themselves to the FH for this reason, there would be a case of heteronomy. Given the conceptions of autonomy and heteronomy above, it is clear that a law has not been given by the will itself in such a case. Rather, a law has been given externally, and rational beings who are subject to such a law do so by virtue of some interest. In fact, at numerous points, Kant himself dismisses theological conceptions of morality as heteronomous and therefore inadequate (G 4:443; KpV 5:40-1). Given this position, Kant’s claims in the *Critique* that God commands our moral duties and that the FH “follows of itself” from God are puzzling. Such remarks appear at odds with his principle of the autonomy of the will.

The resolution to these tensions lies in a distinction between a way of *knowing* and a way of *deriving* something. To help illuminate the distinction, consider an example from math: the Pythagorean Theorem. One can come to *know* the theorem by encountering it in a math textbook and being told that the theorem is an arrangement of mathematical symbols that looks like this.¹³ Henceforth, one will be able to recognize the theorem when they see it. A *derivation* of the Pythagorean Theorem, which one can find by searching for one of its various proofs, is something quite different. More than just providing the ability to see the theorem and recognize it, such a derivation will provide grounds to

8.) And, for Kant, the only way a law can command categorically – without relying on any of the subject’s interests for its effectiveness – is if the will is viewed as self-legislating or itself giving the universal law (i.e. only if the will has autonomy). One way to understand this claim, as suggested by Reath, “Legislating the Moral Law,” is that those who have legislated a universal law themselves have already exercised their rational capacities to consider and endorse the law solely on its own merits (without appeal to any contingent circumstances, like the interests its subjects may have in certain empirical situations) as viable to hold universally, for every rational being. Therefore, these self-legislators do not require an additional interest in, say, not being punished, in order to be subject to the law; the law in itself is enough for them to subject themselves to it. Autonomy is thus required for categorical imperatives to be possible, and therefore for morality to be possible.

13. $a^2 + b^2 = c^2$, where c is the hypotenuse and a and b are the remaining two sides of a right triangle.

accept the theorem as valid.

By parallel reasoning, my interpretation of how the FH “follows of itself” only presents a way of coming to *know* the FH in the framework of God, not a derivation of the FH from it. If the claim that the FH “follows of itself” is understood as presenting a way of *deriving* the FH from the existence of God, then we would have an issue of heteronomy. A derivation of the FH, much like a proof of the Pythagorean theorem, would suggest that the existence of God provides the grounds for the FH as a valid moral principle, in the sense that God is in fact the source of the FH and that rational beings subject themselves to the FH for this reason. However, if the “following of itself” merely presents a different way of *knowing* the FH in the framework of God, then no claims would have been made as to the grounds of the FH, and there would be no worry of heteronomy. I argue that this is the understanding Kant intended.

There are clear signs earlier in Chapter II.V which suggest that the distinction between *knowing* and *deriving* is operative for Kant. Consider first the premise of my interpretation which initially raised worries of heteronomy: that God commands moral duties to rational beings. This worry can be dispelled by noting Kant’s earlier suggestion that he does not think God *in fact* commands moral duties, but only that rational beings must *consider* or *know* moral duties *as if* they had been commanded by God. This is a distinction between a way of knowing moral duties and the actual grounds of moral duties. That Kant is employing such a distinction can be supported by considering his discussion of religion in Chapter II.V more generally, where he mentions that the moral law requires “a step to religion” (KpV 5:130). The moral law requires that rational beings strive for the highest good, which implies that the highest good must be possible for rational beings.¹⁴ Yet a rational being cannot find the possibility of the highest good even thinkable unless they assume the existence of God, for God is required to furnish the necessary connection between virtue and proportionate happiness. Thus, in striving for the highest good, a rational being must ultimately take a “step to

14. Recall “ought implies can.” See footnote 3.

religion" (KpV 5:126). However, Kant stresses the point that one takes the "step to religion" only after virtue has been attained. He says:

But when morals...has been set forth completely, then – after the moral wish, based on a law, to promote the highest good...has been awakened, which could not previously have arisen in any selfish soul, and for the sake of this wish the step to religion has been taken... (KpV 5:130)

Kant says that one takes the "step to religion" (i.e. believes in the existence of God) after the wish for the highest good has been "awakened." Moreover, he says that the wish for the highest good "awakens" only after one attains virtue, and never before. The attainment of virtue occurs before the wish for the highest good has been awakened (and therefore before the "step to religion"). This suggests that the attainment of virtue or moral conduct generally does not require assuming the existence of God. It cannot be the case that God is required in order for rational beings to obtain moral duties, for Kant clearly thinks that the latter can be fulfilled before God enters the picture. Thus, it appears unlikely for Kant to maintain that God commands moral duties to rational beings. Kant makes this even clearer earlier in the chapter, where he says, "Moreover, it is not to be understood by this that it is necessary to assume the existence of God *as a ground of all obligation in general* (for this rests, as has been sufficiently shown, solely on the autonomy of reason itself)" (KpV 5:125-6). Moral obligation depends solely on the autonomy of the will, even after a rational being takes the "step to religion." It cannot be the case that God in fact commands moral duties.

What the "step to religion" does change, however, is that one must now have a *belief* in the existence of God. Otherwise, a rational being fails to even *think* of the highest good as possible.¹⁵ Thus, after acquiring a belief in God, a rational being *regards* their duties *as if* they were commands of God (KpV 5:129). Kant does not consider duties to be in fact

15. And they must be able to do so, since the moral law commands them to strive for the highest good. Recall "ought implies can." See footnote 3.

divine commands, which would be heteronomous. This is a distinction between a way of knowing and a way of deriving. Kant says only that one must know, regard, or consider moral duties as commands of God, where this is distinguished from a case in which moral duties are *in fact* commanded by God.

One might infer that Kant is employing the same distinction when he claims that the FH “follows of itself” from God’s final end. One might come to know the FH by reasoning from the belief in God’s existence, but this need not entail that the FH is in fact grounded in the existence of God (and so heteronomy). Kant’s distinction between the *ratio essendi* (ground of being) and *ratio cognoscendi* (ground of knowing) is helpful (KpV 5:5n). The existence of God and his final end might be a *ratio cognoscendi* of the FH, but the *ratio essendi* of the FH is still solely autonomy of the will. Once a belief in God’s existence has been acquired, one might be able to *know* the FH by reasoning about the consequences of such a belief, but this does not mean that the existence of God is the grounds of the FH as a valid moral principle. When Kant says that the FH “follows of itself” from his discussion of God, he means the former. Thus, Kant may avoid the worry of heteronomy.

5 Conclusion

A question arises as to why Kant may have wished to present a new way of knowing the FH in the framework of God. Some benefits of doing so are not difficult to imagine. For one, the fact that a supreme principle of morality can be known even after postulating the existence of God may lend a sense of logical consistency to his ethical theory. It would have been troublesome if, by contrast, postulating the existence of God somehow hindered one of his moral principles from being known, or, worse still, led to an outright contradiction of one of his moral principles.

As this paper has shown, however, Kant appears to have come close to such a contradiction. Initially, his claim that the FH “follows of itself,” in addition to the premise that God commands the moral law, seemed to

pose a clear threat of heteronomy to Kant's principle of the autonomy of the will. This paper took up the interpretive work required to resolve these tensions. First, I gave an account of how the FH "follows of itself": the FH follows from the fact that God's final end is the highest good and that humans have autonomy – that is, they are subject to the moral law only by agreeing to the moral law. Then, to show that this "following of itself" avoids the issue of heteronomy, I employed a distinction which appears operative earlier in Chapter II.V – a distinction between a way of *knowing* and a way of *deriving*. I argued that when Kant claims that the FH "follows of itself" from God's final end, he means only that God's existence presents an alternative way to *know* the FH, without any claims to the effect that God's existence is the grounds for the FH as a valid moral principle.

References

Korsgaard, Christine M. *Kant: Groundwork of the Metaphysics of Morals*. 2nd ed. Cambridge Texts in the History of Philosophy. Translated by Mary Gregor and Jens Timmermann. Cambridge University Press, 2012.

———. “The right to lie: Kant on dealing with evil.” In *Creating the Kingdom of Ends*, 133–158. Cambridge University Press, 1996.

Reath, Andrews. *Kant: Critique of Practical Reason*. 2nd ed. Cambridge Texts in the History of Philosophy. Translated by Mary Gregor. Cambridge University Press, 2015.

———. “Legislating the Moral Law.” *Noûs* 28, no. 4 (1994): 435–464.
<https://doi.org/10.2307/2215474>.

ARISTOTLE ON THE NATURE OF OUR INTELLECT

ANDREW DE LA CRUZ

The University of Utah

Abstract. Aristotle's philosophy on the nature of the soul and reason appears to offer two divergent strands of inquiry. Within *De Anima*, Aristotle establishes the soul as a fundamentally hylomorphic entity—a composite of matter and form—and divides soul into three distinct parts: the nutritive soul, the sensitive soul, and the intellectual soul. These parts, as presented, are established as only able to exist as components of a living, embodied being. Yet, within *De Anima*, Aristotle puzzlingly raises the question of whether the rational part of the soul can indeed be considered a true part of the hylomorphic composite, or whether it is instead separable and able to exist beyond. This inquiry regarding the separability of the agent intellect raises many questions about our understanding of the rational soul. It therefore necessitates a broader discussion regarding the nature of the soul, the nature of the divine, and even the precise formulation of the human good. Drawing on the works of Avicenna, this paper ultimately seeks to highlight these tensions and ambiguities, leaving room for a revised interpretation of the agent intellect informed by our distinct type and essence of human reason: reason that is uniquely characteristic of our hylomorphic existence.

1 Introduction

Within the early sections of *De Anima*, Aristotle establishes the intellect, or mind, as part of the hylomorphic soul-body composite, yet later in this work, he paradoxically claims that the mind is not besouled as originally thought but rather exists as an entity “confined by its present conditions.”¹ As such, when free from these conditions, the mind is

1. Aristotle, *De Anima*, in *The Basic Works of Aristotle*, ed. Richard McKeon (New York: Random House, 1941), Book 3, 430a22-26.

impassible, immortal, and eternal,² attributes which have traditionally been ascribed to the divine. What follows from these few lines within 3.5 of *De Anima* is the question of whether Aristotle, when referring to the agent intellect, is referring to a separate divine intelligence or discussing an aspect of human rationality.

Most of the scholarship surrounding Book 3 of *De Anima* adopts interpretive approaches in which competing readings are held as superior based on criteria of translational simplicity or fidelity to the book's governing analogies. This paper, rather than privileging one interpretation over the other, is guided by demonstrating how the two dominant approaches require sacrificing key Aristotelian principles from the rest of his corpus to resolve the tension. In examining the two prevailing interpretations—the fully divine and the semi-divine portrayal of the agent intellect—and weighing their respective costs, it will become evident that there is a third way forward: a path that does not sacrifice any of Aristotle's metaphysical or ethical commitments. Instead of demanding compromise, this alternative approach strengthens the coherence of Aristotle's system as a whole.

But first, before delving into the intricacies of the two dominant approaches, it is essential to outline the core tenets of *De Anima*, particularly how the separability of mind and interpretations of the agent intellect have lasting implications for Aristotle's *Nicomachean Ethics* and *Metaphysics*.

2 Aristotle's *De Anima*

Within the opening lines of *De Anima*, Aristotle declares that certain forms of knowledge are to be honored or prized above all, "either because of their greater exactness or of a higher dignity and greater wonderfulness in their objects."³ On both accounts, he argues, the inquiry into the soul deserves the highest regard. This declaration sets the stage for a

2. Ibid.

3. Aristotle, "De Anima," Book 1, 402a1-4.

three-part investigation that begins with two questions that are central to this work: Is the soul divisible, or is it without parts?⁴ Further, are the affections of the soul all affections of the composite body and soul, or is there any one peculiar to the soul by itself?⁵

Aristotle's conception of the soul is distinct from that of his predecessors, particularly in his insistence that the soul cannot be separated from the body. For Aristotle, the soul is the "actuality of a natural body having life potentially in it."⁶ It is the animating principle that brings the body's material potential for life into action. Although for some, to say that the soul is the actuality of the body is somewhat vague and therefore requires proper clarification.

Potentiality and actuality exist in a structured relationship. First potentiality is the capacity to acquire an ability, intellect, or skill.⁷ Second potentiality, which is also the first actuality, is the possession of that ability, though not yet its use. Finally, the second actuality is the active employment of that ability. Potentiality, then, is the state of readiness or, more fittingly, potential that exists within types of being. Actuality, on the other hand, is the fulfillment or realization of potential through existence or action.⁸ It is the state in which a being possesses its capacity or both possesses and employs it.⁹

Thus, when it is said that the soul is the animating principle of the body's potentiality, what is meant is that the soul actualizes the body's potential for life. Their relationship is one of form to matter. Without soul, body would be husk-like, empty, and lacking the dynamism that we attribute to living. Without body, soul would be akin to a melody without an instrument, or a thought without a mind to think it. This inseparability is central to Aristotle's hylomorphic framework:

"Whether the soul and the body are one: it is as meaningless

4. Aristotle, "De Anima," Book 1, 402b1.

5. Aristotle, "De Anima," Book 1, 403a1-3.

6. Aristotle, "De Anima," Book 2, 412a17-20.

7. Aristotle, "Metaphysics," in *The Basic Works of Aristotle*, ed. Richard McKeon (New York: Random House, 1941), Book 9, 1048a25-35.

8. Ibid.

9. Aristotle, "Metaphysics," Book 9, 1048a35-38.

as to ask whether the wax and the shape is given to it by the stamp are one”¹⁰

In *De Anima*, Aristotle asserts this notion repeatedly. For instance, consider these lines strewn throughout the work:

“From all this, it is obvious that the affections of the soul are mattered formulable essences.”¹¹

“But we must return from this digression and repeat that the affections of soul are inseparable from the material substratum of animal life.”¹²

Yet, puzzlingly, the question of whether any parts of the soul¹³ are themselves divisible from the complex of body and soul is raised constantly:

“A further problem presented by the affections of the soul is this: are they all affections of the complex of body and soul, or is there anyone among them peculiar to the soul by itself... The thinking seems the most probable exception.”¹⁴

“From this, it indubitably follows that the soul is inseparable from its body, or at any rate that certain parts of it are (if it has parts).”¹⁵

“We have no evidence as yet about mind or the power to think; it seems to be a widely different kind of soul, differing as what is eternal from what is perishable; it alone is capable of existence in isolation from all other psychic powers.”¹⁶

These passages conflict with the hylomorphic principle embedded within Aristotelian psychology: that anything referred to as a part or affection of the soul must be bound to the body. Separate existence of any of its faculties seems out of the question. It is for these reasons, taken

10. Aristotle, “*De Anima*,” Book 2, 412b6-7.

11. Aristotle, “*De Anima*,” Book 1, 403a24.

12. Aristotle, “*De Anima*,” Book 1, 403b17-18.

13. Given that the soul has parts.

14. Aristotle, “*De Anima*,” Book 1, 403a7.

15. Aristotle, “*De Anima*,” Book 2, 413a1-4.

16. Aristotle, “*De Anima*,” Book 2, 413b24.

together with what is said in Book 3.5 of *De Anima*, that the controversy arises.

Within Book 3.5, Aristotle seems to confirm his suspicions that the power of the mind is indeed “separable from”¹⁷ the complex of soul and body. He begins the text with the analogous relationship of the whole of nature to that of mind. In that just as there are two modalities of every class of being, the matter or potentiality and the form which actualizes it, these elements must “likewise be found within the soul.”¹⁸ The mind, presented as a faculty of soul,¹⁹ follows this structure and is distinguished into two senses: the potential intellect and the agent intellect.

The potential intellect, also called the patient intellect, exists in a state without character and is in virtue of receiving and “becoming” all things so that the mind may think of them. It possesses no inherent content of its own, but rather is of a character that is capable of receiving and being shaped by the forms it must think of.²⁰ The agent intellect, on the other hand, resides in a perpetual positive state of activity, and broadly speaking, is what remains when we abstract away all potentiality from thought. Yet, more precisely, it is also the necessary catalyst for actual thought. It is:

“A sort of positive state like light; for in a sense light makes potential colours into actual colours.”²¹

To recap briefly, the patient intellect exists in a state of pure potentiality capable only of receiving the forms of whatever is thinkable. It does not, however, actualize or “think” on its own. That role belongs to the agent intellect, whose nature is of pure activity, making the intelligible forms available and actual. The interaction between these two faculties gives rise to actual thought.

While the nature of the intellect and its role in producing actual thought is clear and uncontroversial, its precise function within human

17. Aristotle, “De Anima,” Book 3, 429b22-23.

18. Aristotle, “De Anima,” Book 3, 430a9-15.

19. Aristotle, “De Anima,” Book 1, 403a8-10.

20. Aristotle, “De Anima,” Book 3, 430a15-17.

21. Ibid.

psychology remains a subject of enduring philosophical debate. The key question is why the intellect is described as separable from the composite of soul and body yet equally construed as a faculty of the human soul in earlier sections?

A common interpretation holds that the patient intellect remains tied to the composite, as its nature is purely that of a destructible²² capacity whose existence ceases in tandem with the body's. In contrast, the agent intellect—described as impassible, eternal, and immortal²³—alone is capable of separate existence.

But taken at face value, this interpretation leaves many questions unanswered. If the intellect is divided into passive and active components, then claiming that one persists while the other perishes makes it difficult to regard the agent intellect as a genuinely human function. The continued existence of the agent intellect beyond the body means it cannot properly be considered part of the hylomorphic composite. How, then, can the intellect be both a human faculty and be able to exist independently of the soul and body?

Thus, the question must be raised as to whether *De Anima* 3.5 is introducing a second intellect distinct from human psychology. This questioning has given rise to two distinct interpretations. On one hand, the humanity of the agent intellect is maintained and can be understood as a collective semi-divine²⁴ human intellect, existing beyond any individual person. On the other hand, it is disregarded, and the agent intellect is taken as a reference to a separate intellect, synonymous with Aristotle's God, the Unmoved Mover. This paper shall therefore examine each interpretation in depth and begin with the more straightforward claim that Aristotle introduces an entirely distinct intellect in *De Anima* 3.5.

22. Aristotle, "De Anima," Book 3, 430a26.

23. Aristotle, "De Anima," Book 3, 430a23-24.

24. For the qualities attributed to the agent intellect cannot be discounted as only human and its divine qualities must be acknowledged in either interpretation.

3 Caston on Theology and the Human Intellect

While the association of *De Anima*'s agent intellect with Aristotle's God stretches from late antiquity through medieval commentaries, Victor Caston's 1999 work, *Aristotle's Two Intellects: A Modest Proposal*, provides a particularly compelling contemporary articulation. He holds that most of the scholarship has disproportionately focused on parsing the analogies within *De Anima* 3.5 and has ignored what is plainly articulated in the text. In his view, such approaches amount to a strategic circumvention—an effort to psychologize the agent intellect despite Aristotle's unambiguous²⁵ presentation of it in theological terms. Thus, when interpreters insist on treating the agent intellect as either a human faculty or as something ordered through empirical cognition, their task then becomes to “go on a wild goose chase trying to invent a role for the so-called agent intellect to play in human psychology.”²⁶

Caston rejects such textual parsing as unnecessary and holds that we ought to view Aristotle's discussion of cognition not qua human but from a cosmic perspective.²⁷ The text then does not examine cognition merely as an immanent human capacity but as a capacity actualized by the divine. God moves our intellect as a “beloved,”²⁸ a final cause that draws and actualizes our intellection towards perfection and completeness. Thus, in Caston's view, the agent intellect is the divine intellect itself²⁹ (Aristotle's god) externally actualizing human cognition, thus rendering *De Anima* 3.5 not a matter of psychology but theology.

Caston has his reasons for believing this view, as it aligns more straightforwardly with the outright divine qualities attributed to the agent intellect. Namely, of the 11 divine qualities ascribed to the Unmoved Mover:

25. Although my view does not hold that the nature of the agent's intellect is unambiguous, the word is used to give both due consideration and merit to Caston's classical theistic interpretation.

26. Caston Victor, “Aristotle's Two Intellects: A Modest Proposal,” *Phronesis* 44, no. 3 (1999): 199–227, 200.

27. Ibid.

28. Ibid.

29. Caston, “Aristotle's Two Intellects,” 222.

1. It is separated from sensibles. 2. It is impassible. 3. It is unalterable. 4. It is without matter. 5. It is pure actuality. 6. It is the most honorable. 7. It is identical to its object. 8. It is prior to capacity. 9. It is eternally thinking. 10. It is defined by its essence as thinking. 11. It is eternal. 12. It is the necessary condition of everything.³⁰

The agent intellect shares at least eight of these qualities in an exact match. (While three of them differ only slightly in principle and application to being and thought.)

Caston further motivates his position through a novel reading of a passage in *De Anima* that has historically presented difficulties for theological interpretations of the agent intellect:

“When mind is set free from its present conditions, it appears just as what it is and nothing more: this alone is immortal and eternal (we do not remember because, while this cannot be affected, the intellect that can be affected is perishable).”³¹

Notably, Aristotle’s usage of “we” in reference to the mind that is “immortal and eternal” suggests a direct association between the human and divine intellect. An association that implies integration³² rather than the speciation that Caston proposes.³³ At a minimum, such phrasing affirms a psychic quality to the intellect in question and does not suggest divine external guidance. However, this traditional reading may require reexamination. Caston proposes that it ought to be read as the following:

“We do not remember that while this cannot be affected, the intellect that can be affected is perishable.”³⁴

He interprets the phrase “we do not remember that” not as evidence for an association of the human and divine, but as serving a rhetorical and “philosophically urbane” purpose.³⁵ In his reading, Aristotle’s words

30. Caston, “Aristotle’s Two Intellects,” 212.

31. Aristotle, “De Anima,” Book 3, 430a23-25.

32. In the sense of psychic faculties.

33. Caston, “Aristotle’s Two Intellects,” 200.

34. Caston, “Aristotle’s Two Intellects,” 214.

35. Ibid.

should be read as a prompt towards philosophical reflection, thereby reframing the discussion as one concerned with human mortality, mortality that often clouds our perception and understanding of the divine. Thus, the passage operates dialectically, for it first concerns itself with human noetic perishability and then contrasts it with the immortal and eternal intellect. For Caston, the “we” thus serves a pedagogical, not an associative, purpose.³⁶

However, while commendable, this view inherits its own instabilities, for Caston’s theological reading of the agent intellect generates significant tension with Aristotle’s ethical commitments. At the heart of *Nicomachean Ethics* lies the fundamental premise that human flourishing (eudaimonia) must stem from an activity that is characteristic of our soul: rational activity in accord with virtue.³⁷ Yet if the actualizing principle of human thought does not essentially belong to us and stands only as an external mover completely disengaged from human temporality, then does this not threaten to alienate the highest good within us?³⁸ Rationality would paradoxically depend on an actualizing principle radically external to the soul—how then can we say the human good is an activity in accordance with it?

The implications intensify when we consider practical reasoning, for if the virtue explicitly tied to our embodied existence requires the divine intellect’s activation, this imposes theological constraints onto a domain that Aristotle takes pains to ground in immanent human experience. What emerges then is not merely an interpretive dilemma but an existential one, as Caston’s view renders the *human* good fundamentally heteronomous. As it stands, it is in direct opposition with Aristotle’s insistence that virtuous activity must belong to us intrinsically, or *kath’*

36. Caston, “Aristotle’s Two Intellects,” 214.

37. Aristotle, “Nicomachean Ethics,” in *The Basic Works of Aristotle*, ed. Richard McKeon (New York: Random House, 1941), Book 1, 1098a15-17.

38. Although the precise nature of the highest good within us is debated (particularly given Aristotle’s apparent ambivalence in Book 10 of the *Nicomachean Ethics* between practical and contemplative activity) it remains clear that the actualization of intellection holds greater significance than its mere potential. Setting aside that controversy, it follows that one primary, if not the primary, good within us is the activity of thinking, which is made possible only through the agent intellect.

hautō.³⁹ Thus, to make the divine the actualizer of our intellect is to undermine the autonomy that makes virtue ethically significant. Caston's proposal, if accepted without mediation, renders the human good as not even human.

4 Gerson on Functional Unity and the Intellect

In sum, Caston's fully divine conception of the agent intellect risks alienating human flourishing as fundamentally heteronomous. This tension becomes clearer when we turn to Aristotle's definitive account of human flourishing in the *Nicomachean Ethics*.

Within Book 1.7 of the *Ethics*, Aristotle begins his inquiry into the human good. He qualifies this good in several key respects. (1) It must be desirable in itself and never sought merely for the sake of something else.⁴⁰ (2) It must be characteristic of humans, not a good shared with other forms of life, for what is good and proper for a plant is not so for an animal.⁴¹ (3) What is characteristic of humans is grounded in the function of the soul,⁴² specifically, its rational capacities, for it is this that sets us apart from plants and lower animals. (4) The activity of rationality is of greater importance for our function, for it is responsible for exercising thought.⁴³ Thus, given these qualifications, the human good emerges as good rational practical activity, or rational activity carried out virtuously.⁴⁴

Let us now return to the concern at hand: why these premises compel us to uphold the agent intellect as a human faculty. As it has been established, the human function consists of good, rational, practical activity. Yet *De Anima* 3.5 complicates this picture, as it seems to assert that the only sense of rationality that can truly be considered as a "part of the soul" uncontroversially is the patient intellect. Under the common

39. Aristotle, "Nicomachean Ethics," Book 1, 1097b6-7.

40. Aristotle, "Nicomachean Ethics," Book 1, 1097a29-33.

41. Aristotle, "Nicomachean Ethics," Book 1, 1097b30-33.

42. Aristotle, "De Anima," Book 2, 412b6-7.

43. Aristotle, "Nicomachean Ethics," Book 1, 1098a5-7.

44. Aristotle, "Nicomachean Ethics," Book 1, 1098a15-17.

interpretation, the agent intellect is separated from the soul and, by extension, the human function. Thus, if the divinity of the agent intellect separates it from human psychology, how can it remain fundamental to our ethical fulfillment?

To resolve this, we require an account of the agent intellect that can simultaneously uphold its separability while maintaining its role in the human function. For anything less would disconnect “the most divine thing in us,”⁴⁵ the actuality of thought, from its manifestation in the human good, thereby severing Aristotle’s commitments in the *Ethics* and parts of *De Anima* from his broader system.

Lloyd Gerson offers exactly such a resolution to this dilemma in his work, *The Unity of Intellect in Aristotle’s De Anima*. He proposes a framework of a single intellect with two functional modes: (1) the “intellect proper,” what we call the agent intellect, which exists separately as pure actuality, and (2) the “thinking part of the soul,” or patient intellect, which is the intellect as it becomes operative within embodied cognition.

Rather than attributing independent causal power⁴⁶ to the patient intellect, Gerson interprets it as the relational state in which the intellect proper engages with the embodied soul through phantasmata (mental images formed from sense perception).⁴⁷ Crucially, the thinking part of the soul is not equivalent to the intellect itself, but rather, it is “the locus of the actualization of forms in the images available originally through sense-perception.”⁴⁸ Thus, when Aristotle describes the soul as the “place of forms,”⁴⁹ Gerson believes he is referring to this embodied, image-dependent mode of intellectual operation. Importantly, the intellect proper, functioning in or out of this mode of intellectual operation, is not identical with Aristotle’s God. Instead, it is an intellect

45. Aristotle, “Nicomachean Ethics,” Book 9, 1177a15-17.

46. “Causal” here denotes that the patient intellect exercises some form of genuine intellectual power in its own right, not as a derivative or byproduct of some other intellectual activity.

47. Aristotle, “De Anima,” Book 3, 432a1-7.

48. Lloyd Gerson, “The Unity of Intellect in Aristotle’s De Anima.” *Phronesis* 49, no. 4 (2004): 348–73, 365.

49. Aristotle, “De Anima,” Book 3, 429a27-29.

capable of both separate and embodied thought, simply occurring under different contexts. Thus, in this way, Gerson provides an account of the intellect that sustains both Aristotle's psychologically informed ethical framework and his commitment to separability.

The rationale of this interpretation becomes clear when we examine how any reading that insists on the agent intellect's transcendent divinity must explain why "the divine intellect pops into and then out of the picture"⁵⁰ during Aristotle's extended analysis of human intellect spanning *De Anima* 3.4-7. Without carefully navigating this complexity, one is left with only two unsatisfactory alternatives: either accepting that the text is fundamentally and hopelessly fragmented or concluding that Aristotle inexplicably introduces the divine intellect in the midst of a psychological investigation.⁵¹

Gerson's reasoning hinges on Aristotle's decisive characterization of the agent intellect as both separate and unmixed—attributes that distinguish it radically from lower psychic faculties like imagination or sensation.⁵² Yet, he crucially acknowledges that *De Anima* 3.4's definition of thought's actualization from first potentiality to second actuality is contained within a single intellect for the subject of both states. To fracture this process across two intellects, as Gerson notes, would render Aristotle's psychology incoherent. This framework is why Gerson insists that the agent intellect can be "in the soul" as its noetic actualizer, functioning through phantasmata, but not of the soul, in the sense of a hylomorphic faculty bound to the body and therefore incapable of separability.⁵³ In this way he is able to postulate the agent intellect as capable in partaking in human psychology, yet equally as fundamentally separable and divine. Thus, in somewhat grounding the intellect in human immanence, he does not risk severing Aristotle's ethical commitments as Caston does.

Having now established what the intellect proper cannot be in

50. Gerson, "The Unity of Intellect," 349.

51. Ibid.

52. Gerson, "The Unity of Intellect," 357.

53. Gerson, "The Unity of Intellect," 365.

Gerson's account—a hylomorphic faculty or a fully ontologically divine substance—we are compelled to provide a positive characterization of its nature. From its characterization in *De Anima* 3.5, we can discern that the intellect proper is divine-like in nature. Yet, crucially, Gerson rejects its association with the fully divine intellect: Aristotle's God.⁵⁴ Equally, it cannot be a hylomorphic faculty of the soul *dependent* upon embodiment, as this would undermine its separability and impassability. Thus, for Gerson, the intellect emerges as a semi-divine (SD) intermediary between pure divine consciousness and human phantasmata. The intellect, then, is not *the* divine intellect, but *a* divine intellect that enables Aristotle's system to cohere.

Yet this argument, although commendable in the intermediary it proposes, encounters a significant difficulty when we consider that it achieves its coherence only through a strategic recalibration of terms. It merely displaces the original duality between divine and human into a carefully graded distinction. For, even if we accept the view, we are still forced to concede that there are two ontologically distinct intellects—one semi-divine (the human) and one fully divine (God). Moreover, the SD intellect, as Gerson conceives it, would possess a capacity in which Aristotle's God lacks: the ability to engage in both theoretical and practical activity. This, in turn, commits Aristotelian heresy, for the intermediary intellect, which is theoretically complete and practically qualified, could be portrayed as superior to Aristotle's God. Thus, in our pursuit of ontological unity of the intellect, we are given a hierarchy of divinity where the fully divine intelligence is potentially stratified below the semi-divine intelligence.

Moreover, if we accept the possibility of multiple divine-like intellects, then we must also accept that Aristotle's framework provides no principled constraint against further proliferation. This move burdens Aristotle's metaphysical framework by multiplying intelligences to explain phenomena that a unified noetic principle should account for. Without an explicit limit, the proposition of other intermediary intellects

54. Gerson, "The Unity of Intellect," 361.

or gradations of divinity remains possible. Although such a move might be predicated on Aristotle's mention of "47 or 55" eternal movers of the celestial bodies,⁵⁵ (a point further developed by Greek and Arabic thinkers who posited the existence of divine planetary intelligences, e.g., Alexander of Aphrodisias's *On the Cosmos* and Averroes's *Middle Commentary* on Aristotle's *Metaphysics*) it nevertheless introduces an unnecessary complexity into Aristotle's system. It seems then that the semi-divine conception of the agent intellect, although sound in its application, has glaring flaws that, if accepted, seem to leave us in a state of further confusion.

5 On Divine Thought and Human Particularity

We have now arrived at a crossroads between these two interpretations. On the one hand, Gerson's SD view offers the advantage of coherence with Aristotle's ethical work, preserving the agent intellect (just the intellect in his view) as a faculty properly aligned with the human function. Yet, this approach lacks definitive constraints on the intellect and its proliferation, straining the metaphysical integrity of Aristotle's work. What is more, positing two discrete intellects risks rendering human flourishing into a domain of intermediate goodness, forever limited by its exclusion from an intellect that is more divine and perfect, thus condemning the human good into a sort of limbo.

On the other hand, Caston's proposal circumvents these complications by aligning the agent intellect with Aristotle's God. But while this argument does not dilute the divine qualities inherent in the agent intellect nor risk subordinating divine intelligence, it achieves coherence at the cost of rendering the human good heteronomous. For if the highest fulfillment of our nature depends wholly on an external, divine actualizer, then Aristotle's insistence that eudaimonia must be "properly human"⁵⁶ becomes untenable.

55. Aristotle, "Metaphysics," Book 12, 1074a1-12.

56. Aristotle, "Nicomachean Ethics," Book 1, 1098a7-8.

Thus, we are left with an interpretive dilemma that forces us to prioritize Aristotle's metaphysical or ethical commitments—either of which comes at the cost of the other. Gerson's SD view preserves the autonomy of human flourishing at the cost of introducing an ontologically fractured account of intellect. While Caston's theological reading, though metaphysically coherent, achieves its unity by abandoning the *human* character of eudaimonia. Faced with this choice, it seems we must reframe and ask: Which view sacrifices the least? Perhaps, however, the dilemma is not as inescapable as it seems. In truth, the answer to interpreting the agent intellect lies in the excerpt at the end of *De Anima* 3.5:

“Actual knowledge is identical with its object: in the individual, potential knowledge is in time prior to actual knowledge, but in the universe as a whole, it is not prior even in time. Mind is not at one time knowing and at another not. When mind is set free from its present conditions it appears as just what it is and nothing more: this alone is immortal and eternal.”⁵⁷

What we see is that this excerpt admits two plausible readings, each reflecting the perspectives we have discussed in large up until this point. At the individual level, potentiality precedes actuality, i.e., a human must first have the capacity for language before exercising it. Yet universally, actuality is always prior to potentiality; Aristotle's God, the unmoved mover, is the eternal actuality setting all of being into motion.⁵⁸ The tension emerges in the ensuing sentence where this framework appears to be applied to the intellect as a whole. For if “mind is not at one time knowing and at another time not,” then the patient intellect (reminiscent of Gerson's view) could not coherently be called a separate intellect. It would exist dependently, arising or functioning when the agent intellect becomes entangled with the body and halts once the entanglement ceases. This is confirmed in the last sentence: “When mind is set free from its

57. Aristotle, “*De Anima*,” Book 3, 430a21-26.

58. Aristotle, “*Metaphysics*,” Book 9, 1050b20-25.

present conditions it appears as just what it is and nothing more.” The “nothing more” thus serves an exclusionary purpose; mind is not at one time one thing and at another something else. Mind must be ubiquitous. At the beginning of the discussion, Aristotle presents the agent intellect as temporally and ontologically prior to the universe, thereby explicitly aligning it with his God. Yet, the following sentence speaks of the agent intellect as operating within “present conditions” —undoubtedly a reference to human intellection.⁵⁹ Thus, to accept Gerson’s account, we must disregard the opening theological framing that aligns the agent intellect with Aristotle’s God. Conversely, Caston’s reading requires us to dismiss Aristotle’s assertion of the intellect’s operation within human temporality.

For many years, scholars have performed this bifurcation of the text without confronting its central paradox: Why is the agent intellect simultaneously attributed divine qualities yet constrained within mortal cognition? What is needed is not another translation of an already exhausted text, but a more cohesive interpretation that remains consistent with Aristotle’s insistence on the intellect’s divine nature and its essential role in human understanding without compromising either.

The agent intellect ought to be interpreted as neither strictly human nor wholly divine, but rather as Aristotle’s God itself in a self-limiting engagement—or emanation⁶⁰—with human cognition. When entangled with mortal conditions, what Aristotle terms “present conditions,” it actualizes human thought, though it never becomes merely a faculty of the soul (as both Caston and Gerson rightfully maintain). Yet in its unconstrained nature, it remains what it eternally is: pure, actual, and prior in time to all of being. The agent intellect, thus, is to be understood as the self-binding of the eternal within the temporal, as Aristotle’s God emanating and engaging with human cognition.

59. For up until 3.5, the discussion has been explicitly contained within a singular and human intellectual framework.

60. Olga Lizzini, “Ibn Sina’s Metaphysics,” in *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), ed. Edward N. Zalta, accessed April 20, 2025, <https://plato.stanford.edu/archives/fall2021/entries/ibn-sina-metaphysics/>.

But, for what reason would the divine intellect engage in such a self-binding emanation? It is precisely because of its divine nature. Aristotle's God is pure intelligence, thinking only of itself, being both the act and object of its thought.⁶¹ It is all that is intelligible, universal, and thus, in a sense, nothing particular. In its perfection, it requires nothing; in its completeness, it is unmoved.⁶² Yet this very completeness renders it inaccessible and isolates it from particularity. Even if one rejects the thesis of this paper, one must concede that this conception of divinity is, by its very nature, eternally self-limiting.

The greatest misunderstanding lies in the belief that the supremacy of good is the totality of it. For while such totality may encompass all goods in the abstract, it cannot grasp any good in its particularity. Pure completeness, by its nature, excludes distinction—if everything is good, then nothing is uniquely so. Thus, the excellence of goodness lies not solely in the divine that is in us, but in what is especially human. What we are especially is our temporality, our boundedness. Like light, Aristotle's God shines through all things, thinks all things, but we alone are capable of seeing those things individually. The value lies not in accessing the totality of all that is good, but in being able to receive the universal as particular. The self-binding nature of the divine intelligence is not merely a metaphysical function, but a kind of metaphysical necessity: an entry into the particular, not to lose perfection, but to know it more fully.

It becomes necessary, however, to ask how such binding is metaphysically possible, or even coherent, within Aristotle's system. This conception of the agent intellect finds parallel with Avicenna's (Ibn Sina) metaphysical framework, wherein the agent intellect emerges as an eternal emanation from the principal intelligence (Aristotle's God).⁶³

For Avicenna, the principal intellect's very nature, being perfectly identical with its act of self-thinking, or thought upon thought, necessitates an engagement in self-contemplation.⁶⁴ In contemplating

61. Aristotle, "Metaphysics," Book 9, 1074b20-30.

62. Aristotle, "Metaphysics," Book 9, 1074b25-27.

63. Lizzini, "Ibn Sina's Metaphysics."

64. Aristotle, "Metaphysics," Book 9, 1074b20-30.

itself both as the subject and object of intellection, an apparent duality is generated as the principal intellect becomes aware of itself both as the contemplator and the object of contemplation, contingent upon a prior existence.⁶⁵ Thus, in this reflective activity, a mirrored, yet separate, contingent intelligence is produced. An intelligence that is numerically distinct but formally continuous.

It is at this point that Avicenna claims a soul is, in effect, produced through the act of self-contemplation. For the subordinate intelligence's existence is necessarily through another,⁶⁶ thereby becoming contingent. Yet, because the subordinate intelligence shares the nature of the principal intellect, it is unable to sustain this potency within its unified essence.⁶⁷ It therefore externalizes its contingent aspect through the production of a soul.⁶⁸ The soul becomes the receptacle for potentiality while preserving its connection to actuality through its immaterial nature, thus mediating and sustaining the production of intellects, says Avicenna.⁶⁹

For our purposes, however, this moment of divine self-contemplation will not be said to generate a soul as such but rather a modality of potentiality within the emanation. Thus, for the view advanced within this paper, steps one and two remain consistent with Avicenna's metaphysical framework: the principal intelligence, through its eternal act of self-contemplation, produces a secondary intelligence.⁷⁰ Upon reflecting on its existence, this secondary intelligence recognizes that it is contingent, existing not by necessity but through another and therefore is made complex. In this act of reflection, it perceives that the divinity it has 'received'⁷¹ from the principal intelligence is marked by limitation. Here, the model departs from Avicenna's, for the moment the

65. Lizzini, "Ibn Sina's Metaphysics."

66. Ibid.

67. Aristotle, "Metaphysics," Book 9, 1074b25-27.

68. Lizzini, "Ibn Sina's Metaphysics."

69. Ibid.

70. Ibid.

71. "Received" may be a somewhat misleading term here, as the secondary intelligence does not receive divinity in any principled way. Rather, it simply exists alongside the principal intelligence, sharing in its qualities by necessity. Nevertheless, 'reception' serves to indicate the proper sense of contingency of the second intelligence.

secondary intelligence contemplates itself, something akin to a desire might be said to arise as it recognizes the limits of its divinity. Yet, being of pure actuality, the divine intellect can never admit desire, for Aristotle's God admits no privation or potentiality. What would become desire in any other being is, in the divine, an instantaneous self-actualization. The very appearance of longing for temporality is already resolved into emanation. The agent intellect, then, emerges not from desire, but as its annihilation, the only form such an "impulse" could take in that which is wholly actual.

Thus, the dilemma of Aristotle's agent intellect dissolves once we recognize that it is neither a divine principle nor a human faculty in its entirety but rather the self-limiting engagement of Aristotle's God within mortal cognition. The paradox of *De Anima* 3.5, wherein the intellect is ascribed divine qualities yet operative within "present conditions," is not the contradiction it appears to be, but rather reflects a necessary metaphysical function of the divine to engage in particularity.

It must be clarified, however, that the view advanced here does impose alteration upon the agent intellect, insofar as it operates between divine and human modes of being. The agent intellect is, at all times, a divine emanation that makes human thought possible, while the embodied soul particularizes actuality. Yet when it is freed from its present conditions, the intellect appears simply as "just what is, and nothing more."⁷²

6 Within Present Conditions

Of course, one might object that my view does not constitute a genuine resolution, but instead merely circumvents the concern of moral agency. It would still seem that rationality depends upon an actualizing principle external to the soul, thereby forfeiting the human good as uniquely human. Despite the divine intellect's emanative engagement, humans would, nevertheless, still be reduced to vessels for divine goodness rather

72. Aristotle, "De Anima," Book 3, 430a23-26.

than genuine moral agents. At worst, this conforms to the very problem I accuse Caston of committing, though merely having offset radicality and transplanting the locus of alienation from transcendent mover to entangled emanation.

Nevertheless, the objection succeeds in exposing a genuine tension: any account of a divinity actualizing human cognition must explain how virtue's rational dimension can be simultaneously derivative in actualization and authentic in ownership. I concede that if my view implied sole actualization without proper qualification, it would indeed commit the same oversight.

In truth, however, the emanation of the agent intellect, while ontologically prior, does not merely impose rationality upon the soul, but becomes proper to the soul when informed by the soul's temporality: i.e., phantasmata and habituation.⁷³ What we call the patient intellect is not a second intellectual activity, for this would imply either dualism or genuine causal force, but the soul's capacity to particularize the universals through its embodied state. If the agent intellect is like light, making potential colors actual,⁷⁴ the patient intellect and embodied soul are then like the prism, capable of refracting light into particular colors.

Thus, moral agency emerges not because the agent intellect becomes a faculty of the soul, but because the soul's faculties inform and condition divine intellection in a way that makes it functionally the soul's own until it separates upon death. The rational structure of virtue depends on the agent intellect's actualization, and morality emerges from the soul's non-rational components, making it neither a divine imposition (contra Caston) nor a hybrid faculty (contra Gerson), but the soul's own achievement through collaboration.

Now, the justification for this conclusion does not rest solely upon the present analysis; in fact, it finds its most compelling ground in Aristotle's contemplation of human virtue in Book X of the *Nicomachean Ethics*. It is here that Aristotle appears torn between two claims, the

73. Aristotle, "Nicomachean Ethics," Book 2, 1103b14-15.

74. Aristotle, "De Anima," Book 3, 430a15-17.

supremacy of theoretical reason and the indispensability of practical virtue. Both of which seem to be of primary importance to human happiness. On the one hand, the human good is that which is sought for its own sake:

“Of this nature virtuous actions are thought to be; for to do noble and good deeds is a thing desirable for its own sake.”⁷⁵

Yet Aristotle is no less emphatic in his treatment of the contemplative life as the human good:

“If happiness is activity in accordance with virtue, it is reasonable that it should be in accordance with the highest virtue; and this will be that of the best thing in us. Whether it be reason or something else that is this element which is thought to be our natural ruler and guide and to take thought of things noble and divine, whether it be itself also divine or only the most divine element in us.”⁷⁶

In essence, Aristotle’s treatment of the human good appears torn between asserting that human flourishing consists primarily through practical virtuous activity or with theoretical contemplation. If the reading of the agent intellect here is correct, then this very tension is precisely what one would expect Aristotle to struggle reconciling.

Yet, rather than demanding a choice between practical and theoretical virtue, this framework reveals their complementary nature, for the former constitutes the proper exercise of our embodied rationality, the latter its participation in that which is supremely good. This treatment of the agent intellect, then, helps re-conceptualize the enduring debate in Aristotle’s *Ethics* by showing how the human good is reflective of our composite nature: our embodiment and our divinity.

75. Aristotle, “Nicomachean Ethics,” Book 10, 1176b6-8.

76. Aristotle, “Nicomachean Ethics,” Book 10, 1177a11-18.

7 Final Words

Aristotle's exploration of the human intellect in *De Anima* presents a profound tension that challenges the coherence of his metaphysical and ethical systems. While the competing interpretations may resolve aspects of this tension, they do so at a significant cost. The alternative view proposed here, inspired by Avicenna's framework, offers a mediating path in which the agent intellect is viewed as the divine intellect's self-limiting emanation into human cognition. Thus, the view advanced in this paper preserves the divine nature of the agent intellect while simultaneously grounding its operation in human embodiment. Thereby, my view ensures the divine nature of the agent intellect without sacrificing virtue as properly human, even when it is enabled by the eternal.

Aristotle's ambiguity surrounding the nature of the agent intellect may be deliberate, perhaps even reflecting some irreducible complexity of besoulment. Nevertheless, in recognizing the intellect as both divine in origin and human in manifestation, the tension is resolved without severance.

References

- Aristotle. *The Basic Works of Aristotle*. Edited by Richard McKeon. Random House, 1941.
- Caston, Victor. "Aristotle's Two Intellects: A Modest Proposal." *Phronesis* 44, no. 3 (1999): 199–227.
<https://doi.org/10.1163/15685289960500033>.
- Gerson, Lloyd P. "The Unity of Intellect in Aristotle's 'De Anima'." *Phronesis* 49, no. 4 (2004): 348–373.
- Hirji, Sukaina. "Acting Virtuously as an End in Aristotle's Nicomachean Ethics." *British Journal for the History of Philosophy* 26, no. 6 (2018): 1006–1026. <https://doi.org/10.1080/09608788.2018.1454296>.
- Lizzini, Olga. "Ibn Sina's Metaphysics." In *The Stanford Encyclopedia of Philosophy*, Fall 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021.

WAGERS, BRIBES, AND BETS: A DEFENSE OF MODERATE PRAGMATISM

WILLIAM HASSETT

Brown University

1 Introduction

What does it mean to believe something? And how is a belief justified? Many of us might justify a belief by referring to the set of relevant evidence on which the belief is based. *Evidentialism*, a theory in which only evidence provides justification for one's belief, is built on this intuition. However, is evidence the only relevant factor for justifying belief? A brief quote from Richard Rudner raises questions: "[h]ow sure we need to be before we accept a hypothesis will depend on how serious a mistake would be".¹ While some hold steadfast to evidentialism, others conceive of an alternative: one in which pragmatic factors are considered in the framework of epistemic justification. For the remainder of this paper, views of this sort are labeled as a form of *Pragmatism*, as per Alex Worsnip.

Many accounts of pragmatism involve a form of encroachment, in which the stakes of one's situation become relevant in some, but not all cases. Importantly, evidence remains the basis of belief. In contrast to a view in which pragmatic factors apply in all cases of belief, with evidence not playing a role at all,² pragmatists hope to remain moderate, with the claim that pragmatic factors are not the *only* factor relevant for epistemic justification. However, crafting a theory in which pragmatic factors only

1. Richard Rudner, "The Scientist Qua Scientist Makes Value Judgments," *Philosophy of Science* 20, no. 1 (1953): 2.

2. See Susanna Rinard, "No Exception for Belief," *Philosophy and Phenomenological Research* 94, no. 1.

affect belief in a limited set of cases is difficult. While pure pragmatists and anti-pragmatists (such as evidentialists) can clearly define requirements for justification, moderate pragmatists must toe the line while maintaining a compelling and generalizable theory.

This paper begins with an introduction to moderate pragmatism by way of Jeremy Fantl and Matthew McGrath's *PC* principle, before moving on to a discussion of various objections raised by Alex Worsnip. In light of this discourse, I conclude with a brief exploration of how pragmatism might remain moderate when faced with other serious objections.

2 **Motivating Pragmatic Encroachment**

Fantl and McGrath (F&G) reject evidentialism outright. More specifically, they reject the notion that a number of subjects, presented with a common set of evidence regarding a proposition *p*, are all either justified or not justified in believing *p*. An evidentialist might support this assertion with the claim that a specific set of evidence should lead to only one correct conclusion, therefore two subjects with the same set of evidence but opposing beliefs cannot both be justified. F&G present two subtly different cases to illustrate their argument against this claim.

In both cases, you are a soon-to-be passenger on the commuter rail out of Back Bay Station in Boston. On the platform, you overhear another passenger mention your train is one that makes all the stops—one of which is Foxboro. In Train Case 1, you are on vacation and are excited to visit some friends in Providence, your final destination. Whether or not the train makes all the stops has no impact on you, besides your arriving in Providence a few minutes earlier or later. Therefore, based on the stranger's information, you form the belief that your train will stop in Foxboro. In Train Case 2, your final destination is Foxboro, and it is critical that you arrive as quickly as possible. Whether or not the train makes all the stops has an impact on you, and determines whether or not you will be able to get off at the correct station. Therefore, recognizing the stranger may not be experiencing the same high-stakes situation as you,

you decide to make sure for yourself before forming any beliefs that your train will stop in Foxboro.³

This example demonstrates how one's situation may affect whether or not a belief is justified. While the evidence—information overheard from a stranger—is constant across cases, the contrasting conclusions within each case both feel justified. However, the train cases do not necessarily spell doom for evidentialism. An evidentialist might claim that tying these actions to belief is simply conflating being responsible with being epistemically rational. In either case, evidence is minimal. It is only when faced with stakes that it becomes the responsible thing to do to verify what you overheard. Ultimately, how one reacts to a situation's pragmatic factors is an action and an action alone—irrelevant to belief formation and epistemic rationality. Therefore, F&G must seek to clearly define the role of pragmatic factors in epistemic justification through tying actions to belief.

Building on our intuitions regarding Train Case 1 and 2, F&G open with a claim: given your knowledge of a proposition *p*, you should be comfortable acting as if *p*. For example, when faced with a decision between actions *A* and *B*, if knowledge of *p* results in the clear choice of *A* whereas knowledge of not-*p* results in the clear choice of *B*, if you know that *p*, choosing *B* is unexpected. Therefore, if one knows *p*, and if knowledge of *p* results in the clear choice of action *A*, then one also knows that *A* is the clear choice. After restating that knowledge of *p* results in the understanding that *A* is the “best thing to do” and doing *A* given *p* would be rational, F&G arrive at the following principle: “*S* knows that *p* only if, for any act *A*, if *S* knows that if *p*, then *A* is the best thing she can do, then *S* is rational to do *A*.”⁴ Here, the intuition regarding decision making on the basis of knowledge is incorporated as a necessary condition for knowledge. From this starting point, F&G provide reinforcement with four amendments.

3. Jeremy Fantl, and Matthew McGrath, “Evidence, Pragmatics, and Justification,” *The Philosophical Review* 111, no. 1 (2002): 77.

4. JFantl and McGrath, “Evidence, Pragmatics, and Justification,” 73.

The first involves broadening the scope from decisions between actions to “any states of affairs.” Therefore, the preceding principle need not apply only to actions, and instead may refer to a wide range of possibilities, including—as we will see—belief states.⁵ This way, knowing the train will stop in Foxboro also rests on your decision between, say, believing or disbelieving that the train will stop in Foxboro, in addition to actions.

Second, to address cases in which what makes a state of affairs A or B “best” is not entirely clear, F&G introduce the idea of preference. When there is insufficient evidence to determine whether A or B is better, holding preferences towards one in relation to your belief of p is rational: “if you know that p, and if you are rational to prefer one state of affairs to another, given p, then you will be rational to prefer that state of affairs to the other in fact”.⁶ Here, “in fact” refers to your preferences given the evidence and stakes of the present situation, whereas “given p” refers to your preferences given p with absolute certainty. Imagine that on the train to Providence, you get hungry. Though you are unsure whether the cafe car is open, you feel it would be better to purchase a sandwich en route than to wait until you arrive, potentially making yourself late to meet your friends. Soon after sitting down, the conductor announces over loudspeaker that the cafe car is open for service. Once you know for sure the cafe car is open, you are certainly rational to prefer eating on the train. However, before the announcement, or “in fact,” you were also rational to prefer eating on the train. Therefore, being rational to do A is reframed as being rational to *prefer* A.⁷

So far, F&G’s principle has functioned only as a requirement for knowledge. A third amendment redefines it as a requirement for epistemic justification, yielding the following: “S is justified in believing that p only if, for any states of affairs A and B, if S is rational to prefer A to

5. Whether a subject holds any agency over their belief states is the subject of debate. However, F&G accept it here as does Alex Worsnip in his critique, as discussed later. Therefore, we proceed with the assumption that decisions about whether to believe or disbelieve may be up to the subject.

6. Fantl and McGrath, “Evidence, Pragmatics, and Justification,” 74.

7. Fantl and McGrath, “Evidence, Pragmatics, and Justification,” 73-5.

B, given p, then S is rational to prefer A to B in fact".⁸ Consider two otherwise identical cases with the only difference being that one subject holds knowledge of p while the other is merely justified in believing p—not having reached the threshold for knowledge. If given p, both subjects are rational in preferring A to B, both are also rational to prefer A within their respective situations.

The consequent of the principle above (all that follows "only if") is a conditional, stating that one's preferences given p must be reflected in one's preferences given their current situation. The final amendment redefines this relationship as a biconditional—wherein if one were rational to prefer A to B in fact, one must also be rational to prefer A to B, given p. This biconditional reads: "for any states of affairs A and B, S is rational to prefer A to B, given p, [if and only if] S is rational to prefer A to B, in fact".⁹ In the context of belief, this amendment means that if one is justified in believing p, their preferences in fact and given p are never different. Intuitively, this claim holds: if one is justified in believing p and prefers A to B in their current circumstances, there is no case in which one would be rational to prefer B to A, given p. The satisfaction of the biconditional is condensed as "to prefer as if p," and we arrive at *PC*, F&G's account of a necessary condition for epistemic justification as relating to pragmatic factors. The final version reads: "S is justified in believing that p only if S is rational to prefer as if p".¹⁰

Returning to Train Cases 1 and 2 provides an opportunity to observe *PC* in action. Consider the decision in either situation between boarding the train and double-checking, with both options standing in relation to the belief that the train makes a stop in Foxboro. Call this belief p. In Train Case 1, given your lack of stakes, the testimony of a stranger on the platform is enough evidence for you to board the train without a second thought. This decision represents your preference to board given your current circumstances, or in fact. Similarly, given p, you would also board

8. Fantl and McGrath, "Evidence, Pragmatics, and Justification," 77.

9. Fantl and McGrath, "Evidence, Pragmatics, and Justification," 77.

10. Fantl and McGrath, "Evidence, Pragmatics, and Justification," 77.

the train without a second thought. As your preference in fact and given p align, you are rational to “prefer as if p ” and may thus be justified in believing p . In Train Case 2, the testimony of a stranger on the platform is not nearly enough evidence given your high stakes, and you choose to double-check. Again, this decision represents your preference in fact. However, given p , you would immediately board, content with the fact that the train will stop in Foxboro. As your preferences in fact and given p do not align, you do not “prefer as if p ” and the biconditional is not fulfilled, leaving *PC* unsatisfied. As a result, in Train Case 2, you are not justified in believing p , whereas in Train Case 1, you may be. Given both cases include identical evidence in support of p , the difference in epistemic justification presents an example of *pragmatic encroachment*, where the relative high stakes in Train Case 2 make belief in p no longer justified.¹¹

PC addresses Train Case 1 and 2 well. However, it is important to note F&G arrive at *PC* through the development and execution of an argument. It is by no means created ad hoc to capture these cases specifically. And while *PC* delivers encroachment in Train Case 2, it is important to note that all pragmatic factors—not simply those included in the case—must be considered when determining whether one is rational to prefer as if p . Though double-checking in Train Case 2 is an intuitive decision to most, pragmatic factors are by definition subjective. To someone with little regard for the possibility of missing their stop, the good word of a stranger may be sufficient to board the train. For such an individual, their preference in fact would align with their preference given p , and pragmatic factors would not encroach on the justification of their belief. But for most, Train Case 2 presents a higher set of stakes than Train Case 1. By stripping away justification in Train Case 2, *PC* hopes to show that as stakes increase in magnitude, so too does the threshold for justified belief. It is this feature of *PC* and pragmatist accounts in general that present major issues for remaining moderate.

11. Fantl and McGrath, “Evidence, Pragmatics, and Justification,” 79-80.

3 The Case Against Moderate Pragmatism

Alex Worsnip defines moderate pragmatism as a theory that differentiates between two sets of cases: those in which pragmatic factors should affect epistemic justification and those in which they should not. However, he contends, drawing a line between these two sets of cases is difficult. In the previous section, pragmatic factors seemed to be an intuitive consideration when justifying beliefs, as demonstrated by Fantl & McGrath's *PC*. While situations like Train Case 1 and 2 are well met, Worsnip questions whether moderate pragmatist views are successful in filtering out the second set of cases, or those in which pragmatic factors should not affect epistemic justification. To illustrate, Worsnip introduces two examples: Risks and Bribe. In Risks—similar to Train Case 2—you have evidence to support that a job interview starts at 3pm. If you are wrong, you will miss out on the opportunity. In Bribe, you have evidence to support that a game starts at 3pm. If you can prevent yourself from believing it does, an eccentric millionaire will pay you \$1,000,000.¹² Accounts of moderate pragmatism, Worsnip writes, are those in which pragmatic factors affect epistemic justification in cases like Risks but not in cases like Bribe. Maintaining this delicate balance of inclusion and exclusion is critical to the success of theories like *PC*. However, success in achieving a balance is exactly Worsnip's worry: that moderate pragmatism represents an "unstable stopping point" between hardline pragmatism and evidentialism.

With respect to Risks and Bribe, F&G's theory delivers desired results. Recall *PC* and the concept of preference: "S is justified in believing that p only if S is rational to prefer as if p".¹³ In Risks, calling the interviewer to confirm the appointment time is what one would prefer in fact, or given the situation as presented. Conversely, given p—that the interview does start at 3pm—one would not bother to call. The agreement of preference in fact and given p is a necessary condition

12. Alex Worsnip, "Can Pragmatists Be Moderate?" *Philosophy and Phenomenological Research* 102, no. 3 (2021): 531–58. 531.

13. Fantl and McGrath, "Evidence, Pragmatics, and Justification," 77.

to “prefer as if *p*.” This condition is not met here. It seems irrational to prefer as if the interview starts at 3pm, and unjustified to believe as such. Therefore, pragmatic factors affect belief thresholds in Risks.

In Bribe, doing one’s best not to believe the game starts at 3pm seems like the best option. This is what one would prefer in fact. Even given *p*—that the game starts at 3pm—believing the game does not start at 3pm remains the best option. As receiving the \$1,000,000 depends only on belief, one could still reasonably believe that the game does not start at 3pm despite strong evidence. The alignment of preference in fact versus given *p* leads to the conclusion that one would be rational to “prefer as if *p*.” Therefore, preferring as if *p* is to prefer to believe that the game does not start at 3pm. While this seems strange, it is a positive result for *PC*. This way, F&G can reconcile the motivation to disbelieve *p*, and pragmatic factors do not affect epistemic justification in Bribe. Though under *PC* you remain justified in disbelieving *p*, to do so in the face of strong evidence is to hold a patently false belief.

With Risks and Bribes well met by *PC*, Worsnip introduces new cases. One such is commonly referred to as Pascal’s Wager, attributed to Blaise Pascal. Originally, it was developed to demonstrate that those who do not believe in the existence of God are gambling (and poorly, at that!). Pascal argues that no matter what probability you assign to God’s existence, the infinite utility of heaven and infinite disutility of hell after death make belief in God the rational choice. Worsnip’s conception may be simplified in the following form, where *p* refers to the claim that God exists:

Believe *p* when *p* is true: Heaven

Believe *p* when *p* is false: Nothing

Disbelieve *p* when *p* is true: Hell

Disbelieve *p* when *p* is false: Nothing

There are a variety of factors which make this case controversial. Pascal’s Wager is structured not only to encourage belief in God, but specifically to show that belief in God is the only rational choice. Despite

this, many wagerers are left unconvinced. For one, changing one's beliefs with the snap of one's fingers may seem troubling.¹⁴ Others reject the importance of belief in determining assignment to afterlife, or doubt *p* in the first place. Critically, Pascal's Wager relies on an implicit assumption that there is some reasonable possibility that God exists to function. For now, we set these and other issues aside.

Returning to *PC*, direct application to Pascal's Wager does not result in pragmatic encroachment. Provided one accepts the terms of the wager, one would prefer to believe *p* in fact—no matter their prior consideration towards *p* or level of justification. Given *p*, one would certainly prefer to believe *p*. With consistency across preferences, you are rational to prefer as if *p* and pragmatic factors do not affect epistemic justification.

However, *PC* seems to run into issues with an altered version of Pascal's Wager—with one key difference. Whereas *p* previously referred to the claim that God does exist, here *p* refers to the claim that God does not exist.¹⁵ As shown above, we found pragmatic factors do not affect epistemic justification in Pascal's Wager under *PC*. However, Worsnip contends *PC* seems to return the opposite result in Pascal's Wager (Modified). Potential outcomes for the modified case are presented below:

Believe *p* when *p* is true: Nothing

Believe *p* when *p* is false: Hell

Disbelieve *p* when *p* is true: Nothing

Disbelieve *p* when *p* is false: Heaven

Again, provided one accepts the terms of the wager, one would prefer not to believe *p* in fact. However, given *p*, it would seem that one would prefer to believe *p*. Without agreement, one cannot be rational to prefer as if *p*, *PC*'s biconditional fails, and pragmatic factors affect epistemic justification. This result appears damaging to *PC* and moderate pragmatism in general. After *PC* finds pragmatic factors irrelevant to

14. For a broader exploration of this objection, see Nomy Arpaly, "Practical reasons to believe, epistemic reasons to act, and the baffled action theorist," *Philosophical Issues* 33 (2023): 22–32.

15. Worsnip, "Can Pragmatists Be Moderate?" 543 and 550.

justification in Pascal's Wager, the same pragmatic factors affecting justification in Pascal's Wager (Modified) is strange.

The original Pascal's Wager relies on an assumption that a wagerer assigns some probability—no matter how small—to the existence of God. The case then creates stakes with *infinite* benefit or detriment (heaven and hell) to deliver the result that belief in the existence of God is your only rational choice. In doing so, Pascal's Wager eliminates the relevance of the particular level of evidence one might have for God's existence. *PC*, though it accounts for pragmatic factors, is formulated to test a particular level of evidence against the particular circumstances of a situation. Without a specified level of evidence, Pascal's Wager is markedly different from previous cases.

To an extent, this makes sense. The existence or non-existence of God is possibly beyond human comprehension. It is certainly beyond any human consensus. Therefore, claims about one's preferences with respect to religious belief are difficult to engage with. Different people may assign different levels of evidence with respect to the existence of God. Perhaps more importantly, different people may assign different sets of stakes with respect to such a belief. Worsnip's variation of Pascal's Wager simplifies the outcomes of belief to an eternity in heaven or hell. However, when faced with a decision of this nature, it is intuitive that individuals may have other considerations to reckon with. First, we will examine how individuals at opposite ends of the belief spectrum may engage with Pascal's Wager. Then, we will return to Worsnip's argument in its strongest form—an individual who accepts the terms of Pascal's Wager as is.

The following claim is critical to Worsnip's result: given God does not exist, you are rational to believe he does not exist. Worsnip writes, "conditional on *p*—conditional on God not existing—it is rational to prefer that you do believe *p*".¹⁶ It is this claim which lays the foundation for an argument against *PC*. At first glance, this feels sound. Under the specific set of circumstances described by Worsnip, it is rational to

16. Worsnip, "Can Pragmatists Be Moderate?" 550.

disbelieve p given p . However, it is important to understand how the simplifications of the original Pascal's Wager—as well as Worsnip's altered versions—affect an argument which hones in on the importance of pragmatic factors.

Addressing relevant considerations beyond those suggested by Worsnip may shed light on *PC*'s unexpected result. Given the nature of such high stakes as heaven or hell, accounting for other pragmatic factors seems both reasonable and necessary. In the case of a highly religious person—say, a priest—payoffs may be different from those presented on the previous page. The claim that “if God doesn't exist, there's no cost to accurately believing that this is so”¹⁷ does not account for this sort of person, for whom religious belief is a facet of daily life. Instead, Pascal's Wager (Modified) might involve a decision between lives with or without religion:

Believe p when p is true: No Religious Benefit

Believe p when p is false: No Religious Benefit, Hell

Disbelieve p when p is true: Religious Benefit

Disbelieve p when p is false: Religious Benefit, Heaven

Like many others, the priest would prefer not to believe p in fact, given existing religious beliefs or the threat of being sent to hell. The priest might continue on preferring not to believe p even when given p , depending on the utility associated with a lifetime of religious practice. Notably, the introduction of other pragmatic factors here make the original stakes of heaven and hell entirely irrelevant to belief justification.

Worsnip makes a different assumption related to one's preferences in fact. One who accepts the terms of the wager would be “rational to prefer not to believe p , since the belief that God doesn't exist risks sending you to hell”.¹⁸ However, even with the high-stakes threat of hell, it seems plausible that some individuals would nonetheless prefer to believe p . In the case of a decidedly *non*-religious person—say, an atheist—payoffs would again be different:

17. Worsnip, “Can Pragmatists Be Moderate?” 550.

18. Worsnip, “Can Pragmatists Be Moderate?” 550.

Believe p when p is true: No Religious Disbenefit

Believe p when p is false: No Religious Disbenefit, Hell

Disbelieve p when p is true: Religious Disbenefit

Disbelieve p when p is false: Religious Disbenefit, Heaven

Depending on the disutility a person associates with a lifetime of religious practice, one might prefer to believe p in fact, despite the threat of being sent to hell. Given p , the same person would certainly prefer to believe p . Therefore, in the cases of a priest or a staunch atheist presented with Pascal's Wager (Modified), pragmatic factors do not seem to affect epistemic justification.

The above arguments highlight how simplification is key to the success of Worsnip's objection. The absence of other relevant pragmatic factors in Pascal's Wager (Modified), in favor of a specific set of outcomes, is critical to its success. But what of the person who accepts the terms of the wager as they stand? Such an individual would assign some reasonable probability to the existence of God, and their only relevant pragmatic factors would be an eternity in heaven or hell. Allowing for these specific circumstances, the strange result delivered by *PC* seems to stand.

As discussed above, Pascal's Wager is structured to encourage individuals to believe God does exist. This result is demonstrated by *PC* in application to the standard case. Worsnip's modified version does something interesting—application of *PC* instead encourages individuals to consider the claim that God *does not* exist. The result is that Pascal's Wager (Modified) pulls wagerers in opposite directions. On the one hand, the stakes of heaven and hell make belief in God the rational choice in fact. On the other hand, given that God does not exist, disbelief in God becomes the rational choice.

Before discussing Pascal's Wager, we set aside discussing a key assumption: wagerers assign some reasonable probability to the claim that God exists. To do so, a wagerer must hold some level of evidence. The relationship between evidence and religious belief is the subject of

contentious debate, though let's accept that one can rationally hold religious beliefs, for now. Further, whether different people may have different levels of evidence on which to found religious beliefs is unclear, though let's also accept this claim, for now. Both of these assumptions are implicit to Pascal's Wager.

Now, let's return to the individual in Worsnip's case, to whom the above assumptions apply and for whom the only relevant pragmatic factors are heaven and hell. In the original Pascal's Wager, while preferences in fact and given *p* align, a belief in God existing in fact only *happens* to be correct in the event that God does exist. Similarly, in Pascal's Wager (Modified), a disbelief in God not existing in fact only *happens* to be incorrect in the event God does not exist. Therefore, we see how the underlying assumptions of the cases affect the individual's beliefs. It is critical to the rationality of the wagerer's beliefs that God exists. On the other hand, if we were to allow for God not to exist, it is understandable that the stakes of the situation in fact would nonetheless pull a wagerer towards belief in God in fact. If we allow for all assumptions to stand, there appears to be an issue for *PC*. However, this seems an acceptable bullet to bite given the intricate and limiting nature of the assumptions.

Focusing on a simpler example may be helpful in examining moderate pragmatism. Returning to Bribe, Worsnip raises an altered version to continue applying pressure to *PC*. Whereas the original case involved \$1,000,000 payment to form a new belief, Modified Bribe involves a \$1,000,000 payment to withhold from belief. Allowing *p* to represent the claim that the game starts at 3pm, the bribe is contingent on both withholding from believing *p* and *p*'s falsity. This change follows a similar argument as Pascal's Wager (Modified), with the goal of introducing a nearly identical case which yields different results. While the initial Bribe case is handled well by *PC*, where the situation's pragmatic factors do not affect belief justification, Modified Bribe seems to draw out pragmatic encroachment. Accepting the terms of the case, it is rational for you to prefer to withhold from *p* in fact. However, given *p*, it is rational for you to prefer to believe *p*, as the \$1,000,000 payout is

contingent on p 's falsity. Therefore, it seems you are not rational to prefer as if p and are thus unjustified in believing p , under *PC*. Payoffs are represented below:

Believe p when p is true: Nothing

Believe p when p is false: Nothing

Withhold belief from p when p is true: Nothing

Withhold belief from p when p is false: \$1,000,000

Modified Bribe presents a much simpler objection than Pascal's Wager (Modified), while being free of the barrage of pragmatic factors and uncertainty surrounding religious beliefs. Instead, intuitions about preference in fact and preference given p are even stronger. This result highlights a potential weakness in this account of moderate pragmatism. When presented with high stakes or a set of pragmatic factors which strongly incentivize withholding belief, contingent on the falsity of a belief for which there exists strong supporting evidence, *PC* allows for pragmatic encroachment.

4 Broader Implications

A rough trend associated with pragmatic encroachment involves a higher and higher confidence in P being necessary to offset increasing stakes of believing P . In general, this feels intuitive and is a positive feature of moderate pragmatism. However, this affords high-stakes situations great power in creating skepticism, even with beliefs which may otherwise be considered infallible. Jessica Brown draws attention to the effects of high stakes on a given belief in the context of a pragmatic condition on epistemic justification.¹⁹ The example is simple. Take any belief for which you have strong, resounding evidence—something you would be comfortable claiming under oath, for example. Next, construct a bet with outcomes dependent on the truth of your belief. As you raise the stakes

19. Jessica Brown, "Subject-Sensitive Invariantism and the Knowledge Norm for Practical Reasoning," *Noûs* 42, no. 2 (2008): 167-89.

higher and higher, eventually, it seems entirely irrational to accept a bet with such undesirable odds.

Brown describes an example in which a subject, Liz, contemplates whether or not she was born in England. Liz may otherwise feel sure about her birthplace. But when faced with a bet, for which her reward for being correct is £1 and the outcome for being incorrect involves losing her home, she is less confident. Under *PC*, Liz's choice to reject such a bet in fact, to avoid the potential consequence of losing her home—clashes with how she might proceed given she was born in England, where she might prefer to accept the bet and make a quick buck (or pound sterling). Therefore, *PC* seems to yield an instance of pragmatic encroachment, and Liz is seemingly no longer justified in believing she was born in England.

Further issues emerge for *PC* with the suggestion that one is perfectly capable of offering such high-stakes bets on one's own. Whereas Brown's example with Liz is limited to a hypothetical case in which a bet is offered externally, the option to do so yourself effectively transforms any situation into a high-stakes one. While you may have been sure of a belief minutes earlier, does your unwillingness to offer up high-stakes bets truly negate your justification?

In the previous section, clearly defining the payoffs associated with each case steered us towards a counter argument. In the betting scenario above, payoffs fall along the lines of:

If P: You win \$1

If not-P: You lose \$1,000,000

where P is, say, a sound and justified belief. Application of *PC* to this case seems to lead to a verdict where you are unjustified in believing P. While you would decline to offer such a bet in fact, you would be more comfortable doing so given P. Given this set of outcomes, this case presents a serious challenge to *PC*. However, these outcomes are not entirely representative of the pragmatic factors in play. All pragmatic factors are relevant—not simply the monetary stakes. Even with low-stakes bets, it seems unintuitive that given you are justified in believing P, preferring as if P would result in you going around offering

bets. Very few people take this approach. While it may seem that a pragmatist view might license this sort of behavior if you are truly justified in believing *P*, providing a more well-rounded account of the outcomes associated with doing so provides potential answers:

If *P*: You win \$1, but you enter a prolonged process to solidify the bet in legal language, your family thinks ill of you, etc. . .

If not *p*: You lose \$1, and but you enter a prolonged process to solidify the bet in legal language, your family thinks ill of you, etc. . .

Initially, not taking this sort of bet signaled a lack of belief.

However, it is plausible that even given *p*, one might not want to enter into a situation like this. Applying the language of *PC*, one might choose not to offer a bet along these lines in fact, and even given *P*, one might choose not to offer such a bet. While one might adjust the payoffs to become more favorable, so long as one's preferences disincentivize offering bets, the threat to *PC* is neutralized, to some degree.

Though Liz hesitated when faced with stakes of a million dollars, it is not clear every person would. The encroachment of pragmatic factors in Liz's case does not necessarily mean justification evaporates for every belief in a high stakes case. A different person, who has devoted spare weekends to researching their family history, may hold more confidence in the belief that they were born in England and therefore be comfortable with taking the bet. Furthermore, if Liz were presented with a similar level of evidence that this individual possessed, might also be inclined to take the bet. Therefore, we see how *PC* does not lead to skepticism across the board but rather accounts for varying levels of evidence, even in scenarios with the highest of stakes. *PC* was developed to test a particular level of evidence against the particular circumstances of a situation—it seems to do just that here.

Every day, we each enter high stakes situations and each react differently. In the morning, I am often last to leave the house. I make sure to lock the door, and I make my way to work. I reach the end of the driveway, retrace my steps, tug on the handle, and am only then satisfied

with my belief that the door is locked. To me, the stakes of leaving the door unlocked, risking embarrassment and potential theft, is enough to encroach on my belief—despite my very recent evidence of having locked the door. Applying *PC*, my preferences in fact do not align with my preferences given the door is locked, and my belief is no longer justified. It is this intuition that drives me to double-check. On the other hand, a different person, with more self-belief in their locking abilities, might already have reached their personal threshold for evidence, and therefore identical preferences in fact and given *P*.

High stakes scenarios about commonplace beliefs seemed to present a massive challenge for *PC*. However, instead of backsliding towards hardline evidentialism and risking widespread skepticism, *PC* allows us to model how individuals respond to stakes. Everyday, not only are our beliefs tested by high stakes scenarios, but we offer them up—casting bets into the world. Leaving the house, we stake the risks of a break-in against our evidence that we locked the door. Driving to work, we stake the risks of the brakes giving out against our evidence that they worked in the past. *PC* accounts for how justification changes in each scenario, for any set of stakes, for each person, for any set of pragmatic factors—suggesting moderate pragmatism may hold some possibilities still.

References

- Arpaly, Nomy. "Practical Reasons to Believe, Epistemic Reasons to Act, and the Baffled Action Theorist." *Philosophical Issues* 33, no. 1 (2023): 22–32. <https://doi.org/10.1111/phis.12239>.
- Brown, Jessica. "Subject-Sensitive Invariantism and the Knowledge Norm for Practical Reasoning." *Noûs* 42, no. 2 (2008): 167–189. <https://doi.org/10.1111/j.1468-0068.2008.00677.x>.
- Fantl, Jeremy, and Matthew McGrath. "Evidence, Pragmatics, and Justification." *Philosophical Review* 111, no. 1 (2002): 67–94. <https://doi.org/10.1215/00318108-111-1-67>.
- Ganson, Dorit. "Evidentialism and Pragmatic Constraints on Outright Belief." *Philosophical Studies* 139, no. 3 (2008): 441–458. <https://doi.org/10.1007/s11098-007-9133-9>.
- Rudner, Richard. "The Scientist Qua Scientist Makes Value Judgments." *Philosophy of Science* 20, no. 1 (1953): 1–6. <https://doi.org/10.1086/287231>.
- Worsnip, Alex. "Can Pragmatists Be Moderate?" *Philosophy and Phenomenological Research* 102, no. 3 (2021): 531–558. <https://doi.org/10.1111/phpr.12673>.

SOPHISTICATED CONSEQUENTIALISM AND THE PROBLEM OF ALIENATION

IOANNIS KRIZOPOULOS

Franklin and Marshall College

Author's Note. This paper was inspired by Peter Railton's *Alienation, Consequentialism, and the Demands of Morality*. While engaging with the text, I found myself repeatedly questioning how the character of the sophisticated consequentialist was meant to function. Each time I discussed these questions, the conversation seemed to end in the same place: it doesn't work—at least according to Railton himself. I believe Railton came remarkably close to something powerful with his notion of sophisticated consequentialism—only to ultimately abandon it. In this paper, I embrace the challenge of defending that very notion. By taking up the project Railton abandoned too quickly, I argue for the value of cultivating a moral character that aligns with consequentialist ends. A character that can reconcile moral life with personal fulfillment. I recognize that the arguments I present are difficult to defend because of their controversial nature and the perceived contradictions that might arise. Even so, I believe there is something worth salvaging in this moral framework. Despite the tension it creates, I am fully content knowing that I tread a fine philosophical line, and I hope to continue doing so in the future. This is a topic I hope to continue exploring and refining in my future philosophical work.

Part I: The Case of Juan and Linda

1 Introduction

In his paper, *Alienation, Consequentialism, and the Demands of Morality*, Peter Railton makes a claim about the relationship between the practice of morality, as viewed through the consequentialist lens, and alienation. He claims that living up to the standards of morality through the consequentialist lens comes with a certain set of demands that can cause estrangement from personal commitments, emotional capabilities, other

people, and, most importantly, from morality itself, that is, alienation. Even though the highest demand of the consequentialist is to maximize overall goodness, the means often pursued to achieve this end are self-defeating to the goal. By the end of this paper, it will be clear that relentless, self-tormenting deliberation and calculation make consequentialism a theory that ultimately undermines its own goals.

In her pursuit of the maximization of overall goodness, the traditional consequentialist agent inevitably suffers significant losses in many crucial areas of life. A relentless commitment to what she perceives as utility maximization prevents her from being fully present for herself, her family, her friends, and personal ends. More importantly, when a consequentialist realizes that her ultimate objective—maximizing utility—cannot be fully achieved, she faces an even greater loss: her faith in a moral life and consequentialist ideals. This raises a key concern: Are consequentialist moral theories uniquely threatened by alienation? To explore this question, I will begin by presenting Railton's view of character consequentialism, then turn to his case of Juan and Linda as a central example. From there, I will argue that Juan's actions are not suboptimal and prevent significant damage to his personal relationships, and consequently himself. Moreover, I will demonstrate how character-based approaches offer a more sustainable model of moral life. Finally, I will expand on how this form of consequentialism addresses alienation while preserving the desired consequentialist goals.

Railton proposes several potential solutions in which consequentialists act morally without losing touch with their personal lives, values, and relationships. Among these, is his notion of "character/sophisticated consequentialism,"¹ a view that emphasizes establishing a moral character and adopting traits that align with consequentialist values, rather than narrowly focusing on the direct outcomes of individual actions. By fostering character traits that generally promote good outcomes, character consequentialism allows individuals to act in morally maximizing ways without the pressure of

1. Character and sophisticated consequentialism will be used interchangeably in this paper.

constantly calculating the best action in each situation. This approach harmonizes personal commitments with moral obligations, suggesting that a morally good life should naturally flow from one's character rather than requiring constant moral deliberation.

The conventional image of a moral individual is of one who is selfless, generous, and always puts themselves after others. In other words, to be a morally good person and to have a morally good life, you have to always sacrifice yourself for others. The fact that people believe they have to give up what they are passionate about, the things that make them feel alive, to be considered moral pains me. It is quite complex to reconcile strict moral principles with a life that strives to pursue passions. By separating personal passions with moral principles, acting morally appears to stand as an obstacle to living a "decent" life.

This paper seeks to identify a moral principle that reconciles the pursuit of a fulfilling personal life with the demands of generosity and moral consideration for others. A harmonious combination of these two aspects of life makes for a morally good life.

The unattainable standards set by society in order for one to be considered a moral individual are an alarming philosophical concern. Acting morally should be more than non-alienating: one should feel more than merely an absence of conflict with their moral lives, they should feel harmony with it, and that harmony should be accessible to everyone. Thus, a moral theory should pursue depth and explanatory power but also ensure that its practice remains accessible to all. In this paper, I aim to show that leading a morally good life is not only within reach but is also achievable in a way that allows for harmony with oneself, loved ones, morality, and philosophy.

I will focus on the argument presented in the "Juan and Linda" example as the central claim of this paper. Moreover, Juan will serve as the protagonist throughout this paper, embodying the character of the sophisticated consequentialist. The dilemma centers on Juan's choice between two morally compelling actions: visiting his wife Linda, who needs his support, or donating the cost of his trip to Oxfam, a charity that

can use the funds to address urgent needs and maximize utility in a measurable way. The ethical dilemma arises because visiting Linda appears to prioritize personal, relational values, while donating to Oxfam seems to align more directly with utilitarian principles of maximizing overall welfare. Additionally, by donating to Oxfam and disregarding his wife's needs, Juan is forced to compromise both his and his wife's needs, placing them as a lower priority, leading to an imbalance in their relationship. This tension raises questions about whether Juan's decision to prioritize a personal relationship over a seemingly more impactful charitable act constitutes a "suboptimal" choice from a consequentialist standpoint.

In the next section, I will be explaining Railton's character consequentialism and carefully examining the Juan and Linda dilemma in more depth and detail. By clarifying and expanding on Railton's original account of the Juan and Linda example, I aim to demonstrate that Juan's decision to visit Linda is not suboptimal, contrary to Railton's own conclusion.

Not only will I argue that his decision is not suboptimal, but also that by choosing to support his wife, he is fostering character traits that maximize utility in the long run – even if this choice appears less impactful than donating to Oxfam in the short term.

I will establish throughout the rest of this paper how the value of cultivating such traits does not only lie in their consequences. The fact that Juan's decision to support his wife seems to lead to utility maximization is not the only thing that makes these traits morally valuable. Rather, they possess inherent worth and deserve to be acted upon, even when there are perceived short-term shortcomings that will be ultimately shown to be misguided in the long run.

2 Railton's Account of Character Consequentialism

Before turning to Juan's case in depth, it's essential to understand the broader theoretical framework Railton draws from. In this section, I will

outline the three forms of consequentialism he presents, focusing on how they differ and why sophisticated consequentialism stands out.

First, he introduces subjective consequentialism, which holds that individuals must make moral decisions based on the best use of information available to them. After doing so, they are expected to determine which action is most likely to produce the greatest good and try to act accordingly. In contrast, objective consequentialism defines the rightness of an action by its actual consequences—specifically, its capacity to maximize utility—regardless of the agent’s knowledge, intentions, or subjective reasoning. This entails that even well-intentioned actions can be judged as wrong if they fail to produce the best outcomes, placing a heavy burden on moral agents to follow an intense selection process for their actions. Finally, the sophisticated consequentialist seeks to lead a life that aligns with the goals of objective consequentialism, but additionally incorporates flexibility in the decision-making process.²

Rather than simply contemplating how to maximize utility in each decision, the sophisticated consequentialist cultivates values and traits that generally promote utility, thereby achieving consequentialist ends over time without deliberate calculation at every turn. This approach allows for a moral life that fulfills consequentialist aims while also accommodating the complexities of human motivations and relationships.

3 Juan and Linda

In this section, I examine Juan’s character and motivations, not simply to recount the case but to reveal how his decisions reflect the virtues of sophisticated consequentialism that will be articulated later on in the paper. Understanding sophisticated consequentialism through the embodiment of a real person is profoundly illuminating. It’s one thing to grasp the theory in abstract terms and another to see how it might

2. Peter Railton, “Alienation, Consequentialism, and Morality,” *Philosophy and Public Affairs* 13, no. 2 (1984): 134–171, 152–153.

genuinely operate in a person's life. This matters because moral theories often fail not in logic, but in livability. Understanding how a sophisticated consequentialist would feel, deliberate, and act in complex real-world situations allows us to evaluate the framework more holistically. Like Juan, we all have people we care about, and most of us have faced moral dilemmas involving those relationships. I invite the reader not just to consider the theory, but to place themselves in Juan's position and inhabit the mindset of a sophisticated consequentialist. To ask, not only "what would I do?" but "what would it feel like to live this way?"

Let's turn to the case of Juan and Linda in a more detailed manner. Juan is presumably a model husband. When one inquires about his commitment to Linda, Juan responds, "I love Linda. I even *like* her. So it means a lot to me to do things for her. After all we've been through, it's almost a part of me to do it."³ Juan's response offers important insight.

First, Linda is not *just* a person he loves; she is someone he enjoys being around. In contrast to those who love their relatives or childhood friends but can't stand their company – despite wishing them well – Juan's relationship with Linda is both affectionate and fulfilling. Juan's love for Linda, in combination with his statement that it means a lot to do things for her, clearly indicates that it is profoundly beneficial for Juan's well-being to be around Linda, for Linda to be well, and for him to be a contributor to her welfare. Additionally, Juan and Linda clearly share a special connection; Juan even says that being a good husband to Linda feels like an essential part of who he is. From this, we can conclude that failure to be part of Linda's life and welfare can result in alienation from Linda *and* himself.

Now that their relationship and what it means for them to be in such a relationship is clearly stipulated, let's see the different side of Juan: The principled Juan. The Juan that wants to leave a positive footprint on the world. In other words, the part of Juan that desires important positive outcomes. To better comprehend the moral tension he faces, we must also understand this side of him, which depends upon his broader ethical

3. Railton, "Alienation, Consequentialism, and Morality," 150.

commitments and worldview. As Railton describes him, Juan has ambitious desires to change the world. However, Juan is also a realist. He understands that idealist visions are difficult to realize in the real world. One person alone cannot bring about radical change in the world; this is why Juan must cultivate meaningful relationships, not only to foster his own growth but also to support the growth of those around him. If everyone takes care of the people close to them instead of focusing on almost unattainable changes, then real change can arise from smaller actions. This is related to a phenomenon called the “Paradox of Scale.”⁴ This refers to the phenomenon where actions that are possible on a small scale feel meaningful, whereas large-scale problems feel overwhelming or paralyzing. As a result, engaging in everyday actions that are achievable and controllable like recycling, instead of attempting to single-handedly clean out the world’s ocean, endorses a feeling of control and resolution in big issues. Applying this to Juan’s case, he should focus on what he can control to make a meaningful difference, rather than choosing actions that are overwhelmingly big in scale and risk wearing down his motivation and hope.

Furthermore, Juan also thinks that people who believe they can make things better by acting in a “heroic” manner and sacrificing the things they care about are fools. He thinks that people should take care of themselves and their close circle in order to be capable contributors to a larger scale of societal problems. More than that, Juan takes pride in his ability to protect his self-confidence and trust in his abilities. He selectively takes on projects that he knows he can complete, ensuring that his efforts are well worth it. One thing Juan is certainly capable of, and that practically everyone could be, is caring for the people closest to us.

4. The paradox of scale helps explain why small, manageable actions are preferred over overwhelming large-scale solutions. The “small wins” approach shows how localized actions build momentum for broader change. See: Catrien J. A. M. Termeer, Art Dewulf, and Gerrit Biesbroek, “A Small Wins Framework to Overcome the Evaluation Paradox of Governing Wicked Problems,” *Policy and Society* 38, no. 2 (2019): 298–314, <https://doi.org/10.1080/14494035.2018.1497933>. Annelise Jolley, “The Paradox of Our Collapsing Compassion,” *John Templeton Foundation*, November 20, 2024, <https://www.templeton.org/news/the-paradox-of-our-collapsing-compassion>. Brian Resnick, “A Psychologist Explains the Limits of Human Compassion,” *Vox*, July 19, 2017, <https://www.vox.com/explainers/2017/7/19/15925506/psychic-numbing-paul-slovic-apathy>.

Finally, Juan does something brilliant that is often overlooked in traditional consequentialist frameworks: he includes himself in the calculation of the utility. Most people consider their actions as separate from themselves. In other words, their actions affect the external world, and they are not included in that world. The sophisticated consequentialist version explored in this paper holds a more holistic view which Railton failed to carry off. Here the consequences brought upon the agent matter. Not only do they matter, but they should be carefully considered since they have a direct impact on their ability as a benefactor of the world. Juan does not only include himself in the utilitarian calculation, he prioritizes himself. That not only leads to a better life for him and his “little circle” but also makes him a more stable and motivated benefactor of the “big circle.”⁵

4 Juan’s Moral Dilemma

With the essential background now established, we can examine how Railton presents the central dilemma. Juan, a sophisticated consequentialist, and his wife, Linda, have a commuting marriage, meeting only two to three times a month. During a week when they weren’t scheduled to meet, Linda became noticeably depressed and strained. Concerned for her well-being, Juan decides to make an additional trip to comfort her. However, the cost of this trip is significant — enough to fund a well in a village without access to clean water. This puts Juan in a moral dilemma: Should he prioritize comforting his emotionally distressed wife or allocate the funds to a cause that could save and greatly improve lives?

Before analyzing Juan’s options, it’s worth addressing some structural issues with the example itself. Key details are missing, which makes it harder to assess whether Juan should prioritize visiting Linda or donating to Oxfam. For instance, how severe is Linda’s situation? Is she

5. At the same time, he recognizes the severity of worldwide urgencies and that you can’t always “retreat to your little circle. But still, you need that little circle.” Railton, “Alienation, Consequentialism, and Morality,” 150.

simply sad or clinically depressed? Is she on the verge of a crisis or just upset over a minor inconvenience? These distinctions are crucial since they directly affect the expected utility of Juan's decision. The more urgent Linda's needs are, the more morally weighty Juan's presence becomes. Without clear answers to these questions, any consequentialist calculation remains incomplete and speculative.

For the sake of clarity and a balanced analysis, I'll assume a middle-ground scenario: Juan and Linda are navigating a challenging period in their marriage, partly due to the strain of a long-distance relationship. On top of that, Linda is experiencing a particularly difficult week professionally, socially, and emotionally. While her struggles are significant enough to cause emotional instability, she is not clinically depressed nor at risk of harming herself. She could benefit greatly from emotional support, and Juan is in a position to provide it. This framing makes the example more realistic and situates Juan in a nuanced moral dilemma.

Railton suggests that Juan, given his character that values personal relationships, will choose to visit Linda but feel guilty afterward. Through my analysis of the case, this guilt (which Railton claims is deserved) arises because, from the perspective of objective consequentialism, Juan's choice is suboptimal—he did not choose the state of affairs that aims directly at objectively maximizing utility. This raises a series of troubling questions: Is Juan's character flawed if it leads to feelings of guilt and value conflicts? Should he adopt a different character, one more aligned with strict utility maximization? Is his current character consistent with being a true consequentialist?

However, for now, the above are the wrong questions to ask. The primary question should be: Was Juan's act truly suboptimal? I argue that all things considered, long-term and for everyone affected, Juan is performing the utility-maximizing action by visiting Linda and not donating to Oxfam.

But how can this be the case? How can Linda's emotional discomfort be weighed against providing fresh water for a whole village?

What follows is an attempt to examine Juan's case more holistically. When a fuller and more detailed context is provided, in contrast to the original framework of the dilemma in Railton's paper, one will discern between the consequences of both courses of action more reasonably. By considering the full range of relevant factors—character, relationships, and long-term well-being—I will demonstrate the utility-maximizing benefit of maintaining a stable moral character. When viewed from this widened scope, such a character has the makings of a more reliable utility maximizer, especially when compared to narrow, calculation-based approaches.

To make sense of the dilemma, I will frame it through two contrasting evaluative lenses. First, there is the short-term perspective, which centers on the immediate consequences of an action. The judge operating from this perspective tends to focus solely on the surface-level information available, without delving deeper into broader or long-term implications. This approach leads to quick verdicts, often made without considering the full context or possible future repercussions of the judgment. Let's call this the "narrow-sighted judge." In contrast, there is the overall, "all things considered" perspective, which considers the broader implications of each action. Here, the judge evaluates all short-term and long-term effects, weighs their significance, and makes a decision based on the full context. We can name this more comprehensive judge the "sophisticated judge."

Now, let's take on the role of the more discerning judge—the "sophisticated judge." In evaluating the situation, I will consider the short-term and long-term consequences of Juan choosing not to visit Linda and instead donating his money to Oxfam.

In the short term, the most obvious negative consequence is Linda's continued emotional distress without Juan's support. Juan, too, would experience guilt and disappointment for not being there for his wife. On the positive side, Juan's donation would contribute to providing clean water to a village, improving their quality of life, and potentially leading to other long-term benefits.

Next, let's consider the less apparent consequences. First, Juan faces alienation from several important aspects of his life. He is alienated from his wife, as he cannot be there to support her emotionally. He is alienated from himself, as he has a deep desire to be there for her, but chooses not to act on it. He is also alienated from his moral character and morality, as he resists desires that are intrinsic to his values and sense of self. This creates a disconnect between his actions and the person he aspires to be.

Furthermore, if Juan consistently makes choices that harm his character, this would erode his ability to be a stable, healthy, and happy contributor to the broader world. By neglecting his "small circle," with Linda at its center, he would diminish his capacity to be a better person, a more effective contributor, and a positive force in the world. Over time, this would undermine his motivation to act morally. The key to maximizing utility lies in positioning oneself to be a capable consequentialist, and by not supporting Linda, Juan would fail to do this. Whether Juan can still be considered a consequentialist, despite his failure to act by consulting a consequentialist moral framework before acting and by acting solely from his character instinctively, will be explored in Part II of this paper.⁶

Considering the above, much evidence indicates that Juan's decision to not visit Linda is not truly maximizing when all the details are taken into consideration and the broader picture surfaces. In Juan's consistency with such actions, we have good reason to predict a steady decline in his well-being, his connection with his character, himself, and his "close circle."⁷

6. A sophisticated consequentialist might, in extreme cases, be required to sacrifice personal relationships for the greater good—such as choosing between the death of a loved one or a larger group of people. However, these scenarios are rare and unlikely to occur in everyday life, so they aren't particularly helpful for understanding the practical application of consequentialism. Rather than focusing on these cases, it is more productive to consider the more typical moral decisions we face, where the consequences are more relevant to our daily lives, character development, and universal impact. These extreme cases, while theoretically possible, should not dominate the discussion of the theory's real-world implications.

7. Moreover, as I recognize that Juan is not an all-knowing, mind-reading, future-predicting individual, I propose that his character, while maximizing utility, may still make errors. It's essential to clarify that when I say he maximizes, I mean he is a realistic maximizer, not an ideal one. This means Juan has everything necessary to be the best consequentialist version of himself, even though he's not flawless. Mistakes, though rare, are likely in situations where it's difficult to predict outcomes or distinguish between different utility-

5 Conclusion

This first part has argued that Juan's decision to visit Linda is not a moral compromise, but an example of moral excellence under sophisticated consequentialism. More than this, we have explored Peter Railton's critique of alienation in consequentialism, focusing on the relationship between morality and personal fulfillment. Through the case of Juan and Linda, I argued that the decision to visit Linda, rather than donating to Oxfam, is not only consistent with Juan's character but also a morally maximizing act in the long term. This conclusion objects to the seemingly suboptimal choice from a strictly consequentialist perspective, demonstrating that maximizing utility involves more than simply calculating outcomes; it also requires sustaining the character and personal commitments that enable an individual to continue acting morally over time.

While it is true that Juan's decision might not align with the narrow-sighted judgment, a sophisticated consequentialist's broader perspective allows for a more holistic understanding. Juan's decision to prioritize his wife's well-being enhances his ability to maintain a fulfilling life and a strong moral character. By investing in his closest relationships and his own moral development, Juan not only improves his well-being but also positions himself to contribute meaningfully to the larger world over the long run. Ultimately, Railton's character consequentialism offers a viable path to accommodate moral action with personal happiness, showing that the demands of morality do not have to lead to alienation from oneself, one's loved ones, or morality itself. By adopting a more comprehensive view of moral efficacy—one that incorporates the complexities of human relationships and the long-term benefits of virtuous character—individuals can live a morally good life without

maximizing options due to insufficient information. Despite this, these gray-area cases are rare. Juan is highly capable of discerning the best course of action as he is dedicated to cultivating his decision-making skills through his character. In any case, even with this minor flaw, Juan's actions still represent the optimal utility-maximizing result in the context of his situation, making his character's actions the best possible outcome.

sacrificing personal fulfillment. This approach not only strengthens consequentialism as a moral theory but also provides a more practical framework for navigating the moral dilemmas we encounter in everyday life.

Part II: Indirect Aims and Sophisticated Consequentialism

1 Introduction

This section explores Juan's alignment with Peter Railton's sophisticated consequentialism. It begins by addressing how Juan can maximize utility not by consciously aiming for it, but by embodying values that naturally promote well-being.

This section proceeds in three stages. First, I argue that Juan can act as a consequentialist without explicitly endorsing consequentialism as his moral theory. Second, I defend the authenticity of Juan's care for Linda by showing it is not merely instrumental. Third, I address counterarguments, emphasizing the long-term value of character-based moral action over short-term calculation.

2 Acknowledging Utility Maximization Without Directly Aiming for It

Building on the previous section, this part turns to a deeper question: Should we aim to cultivate virtues and character traits for their own sake rather than constantly striving to maximize utility through explicit consequentialist calculation? Is valuing these traits only defensible from a non-consequentialist standpoint, or can they also serve consequentialist ends without being reduced to mere tools? I argue that when we take on meaningful, achievable tasks that align with our character and desires, we not only affirm those traits' intrinsic value but also promote practical utility more effectively than if we had fixated on outcomes.

It is evident that constant deliberation, aiming to maximize utility,

before every action has clear downsides. It puts immense pressure on individuals to conform to consequentialist standards, often at the expense of their own values and passions. Even outside consequentialist contexts, persistent deliberation toward achieving a specific goal often proves to be an inefficient and counterproductive strategy. Actions become robotic, driven by outcomes rather than a genuine endorsement of the end, which can be tremendously alienating. One merely acts on practice: the goal is lost. We become disconnected from both the activity and ourselves because the point and aim of the activity is lost. While one could argue that continuous calculation might maximize utility, this approach is not only exhausting but also counterproductive. Constantly assessing every move can lead to guilt, anxiety, and even depression—emotions that detract from well-being rather than enhancing it.

Railton illustrates this with the example of a tennis player who underperforms when too focused on winning (analogous to utility maximization) instead of the love of the game (analogous to nurturing values and relationships).⁸ Just as the player's overthinking prevents him from playing his best, perpetual calculation in life can prevent us from fostering meaningful connections and enriching personal values, ultimately undermining the very utility we aim to achieve.

Additionally, it's crucial to recognize that it's not utility maximization itself that brings satisfaction, but the qualities and actions that lead to it. In this case, it's not the ends for which we aim, but the means. In this way, we don't perpetually focus on calculating utility, but we cultivate a character that naturally leads us to better results. The ends fall into place without any hassle when the means flourish. This echoes Aristotle's suggestion that virtuous actions are not performed as mere steps toward happiness (*eudaimonia*) but are undertaken because they are inherently good. Virtue and relationships are pursued for their own sake, not as a calculated path to an external end, such as happiness or utility.

The tennis player, for instance, doesn't play solely to win; she plays

8. Railton, "Alienation, Consequentialism, and Morality," 144–45.

because she loves the game and wants to be the best version of herself. When she's driven by this love, she naturally becomes a stronger player, which often results in more wins. If she still doesn't win, it's likely due to external factors, like facing a stronger opponent. Either way, she is acting towards a realistic goal to improve instead of senselessly focusing on winning every match.

3 Juan as a Consequentialist Without Recognizing Consequentialism

As mentioned earlier, there seems to be something inherently self-defeating about intentionally trying to do or avoid something. A classic example is the instruction "Don't think about a red ball." Another example, of course, is the focus on winning in tennis. The takeaway here is that, generally in life, we accomplish our best goals when we aren't distracted by thinking about them. Consider the phenomenon of the "flow state" as an example.⁹

This connects to consequentialist considerations, which can feel inefficient when pursued directly. There seem to be some situations where intentional utility calculations would be more efficient and logical means than acting instinctively through character. However, such situations are extreme and isolated. Consequentialism works better indirectly in everyday life because our minds are not built to constantly calculate outcomes. However, in high-stakes, one-off moral decisions—like a trolley problem¹⁰—the calculation can be more efficient because it's cognitively isolated and outcome-focused.

Now that the distinction above is clear, let us turn to an interesting case where the indirect pursuit of consequentialist ends may raise concerns about terminology. Consider the case of someone who

9. The flow state is a mental state in which a person becomes fully immersed in an activity, while acting naturally in it. It leads to better goal achievement precisely because the person is not consciously fixated on the goal, allowing action to unfold smoothly and efficiently. Mihaly Csikszentmihalyi, *Flow: The Psychology of Optimal Experience* (New York: Harper Row, 1990), 54–75.

10. Judith Jarvis Thomson, "The Trolley Problem," *The Yale Law Journal* 94, no. 6 (1985): 1395–1415

consistently acts in ways that align with sophisticated consequentialist principles, not because they consciously adopt the theory, but simply because they possess a character shaped by upbringing, experience, or personal values that causes them to act in ways that align with consequentialist calculations. Such a person may not identify as a consequentialist, yet their actions regularly produce outcomes that a sophisticated consequentialist would endorse. This raises a compelling question: should we consider this person a consequentialist in practice, even if not in theory? More than that, might they be an ideal consequentialist? One who, by acting instinctively and without constant calculation, effectively realizes consequentialist ends in a way that is both natural and in line with character consequentialism?

Consider a parallel: if I am suffering from cancer but don't know it, am I sick only once diagnosed, or have I been sick all along? If I have been acting per consequentialist aims in a more successful manner than any other consequentialist, am I, the sophisticated consequentialist, something so distant to a consequentialist? It's important to note that an example of someone accidentally succeeding is not useful here as a parallel to my argument. A sophisticated consequentialist would still recognize utility maximization as an outcome of her lifestyle. However, it is not the central goal—positive outcomes may be a motivator and benefit, but even when it's unclear, this character acts based on her values, moving forward without hesitation.

Railton's sophisticated consequentialism allows for individuals to embody consequentialist principles without consciously identifying as consequentialists. Juan's action—visiting Linda rather than donating the travel expenses to Oxfam—exemplifies this possibility. His decisions reflect a nuanced understanding of utility maximization, even if he does not explicitly endorse consequentialism as a moral framework.

Apart from considerations of efficiency, this approach is undeniably more appealing because it is grounded in positive dispositions like affection, empathy, and courage. These qualities not only contribute positively to well-being but, more importantly, help avoid alienation from

our relationships, from ourselves, and from morality.

Juan's lack of explicit commitment to consequentialist aims allows his actions to arise from authentic care and moral character, which are central to Railton's view and are essential for fighting alienation. By prioritizing Linda's well-being as part of his own moral commitments, Juan indirectly promotes the utility that a sophisticated consequentialism values.

4 Refuting Juan's Care as Merely Instrumental

Juan's care for Linda may seem instrumental if viewed merely as a means to maximize utility. However, his care is not fundamentally about utility maximization; it stems from the affection and special consideration he has for Linda. That said, Juan does recognize the utility-maximizing aspect of his actions, which arises naturally from his character, and he is glad for that outcome.

This brings us to an important and challenging question: What happens when the sophisticated consequentialist fails to maximize utility? Should she reconsider her approach? As argued earlier, using the tennis player analogy, my answer is no. It's crucial to understand that developing such a character is not a mere methodology—it is a lifestyle. Morality is not like using a tool where, if one doesn't work, you simply switch to another. It's far more complex. Morality requires a multifaceted approach. It calls for the cultivation of various traits and the understanding of numerous concepts just to begin practicing it—let alone to become truly good at it. Living a moral life involves grappling with alienation, nurturing personal projects, embracing education, and developing a holistic view of the world, among other things. By continuing to affirm her position despite the challenges and perceived shortcomings that might arise, the practitioner of sophisticated consequentialism is avoiding many complexities that are entailed in the efforts to uphold a moral lifestyle. Even when utility is not maximized, the sophisticated consequentialist should first question whether external

factors, such as unusual circumstances or misleading evidence, disrupted her ability to foresee better outcomes. These occasional inadequacies, while unfortunate, are a small price to pay for maintaining the integrity of a character that is built to consistently maximize utility over time. Like the tennis player who accepts occasional losses while honing her skill, the sophisticated consequentialist should persevere in cultivating her character. Despite isolated failures, this approach remains the most effective for maximizing utility in the long run and deserves dedicated commitment.

5 Conclusion

In sum, I firmly believe that the best life is one lived by values, not through the perpetual calculation of utility maximization. Juan's character exemplifies how to live authentically, without alienation from oneself or others. He highlights the importance of nurturing close relationships, reminding us that the world would be a better place if more people maintained such meaningful connections. Juan also demonstrates how one can be an effective consequentialist, not by adhering rigidly to consequentialist theory, but by maintaining a healthy distance from it and avoiding the alienation that can arise from strict theoretical commitments. His actions show that it is possible to eliminate self-defeating habits, like overreliance on explicit consequentialist theory, and still be a utility maximizer with a balanced life.

This raises a significant question: Is consequentialism undermined if one does not adopt it as a theory but still accomplishes consequentialist ends in a practical sense? As I have argued, I tend to believe that practicality holds more weight than theory. It is more appealing and humane for a moral agent to focus on practical considerations rather than theoretical deliberation that leads to unsuccessful practical outcomes. While I acknowledge the tension between rejecting consequentialism as a moral theory and embracing a "practical," "accidental," or "sophisticated" consequentialism, I believe this approach ultimately

achieves the goals of consequentialism most effectively. The challenge to the counterfactual conditional, though philosophically puzzling, is, in my view, a necessary and worthwhile sacrifice for a more practical and fulfilling moral life.

Ultimately, I believe that my adjustments and additions to Railton's view help restore hope in the possibility of practicing morality in a way that is both effective and appealing from a consequentialist perspective. Morality need not alienate us from life, but can guide us toward it.

References

Jolley, Annelise. "The Paradox of Our Collapsing Compassion." *John Templeton Foundation*, 2024. <https://www.templeton.org/news/the-paradox-of-our-collapsing-compassion>.

Railton, Peter. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13, no. 2 (1984): 134–171.

Resnick, Brian. "A Psychologist Explains the Limits of Human Compassion." *Vox*, 2017.
<https://www.vox.com/explainers/2017/7/19/15925506/psychic-numbing-paul-slovic-apathy>.

Termeer, C.J.A.M., and Art Dewulf. "A small wins framework to overcome the evaluation paradox of governing wicked problems." *Policy and Society* 38 (August 2018): 1–17.
<https://doi.org/10.1080/14494035.2018.1497933>.

