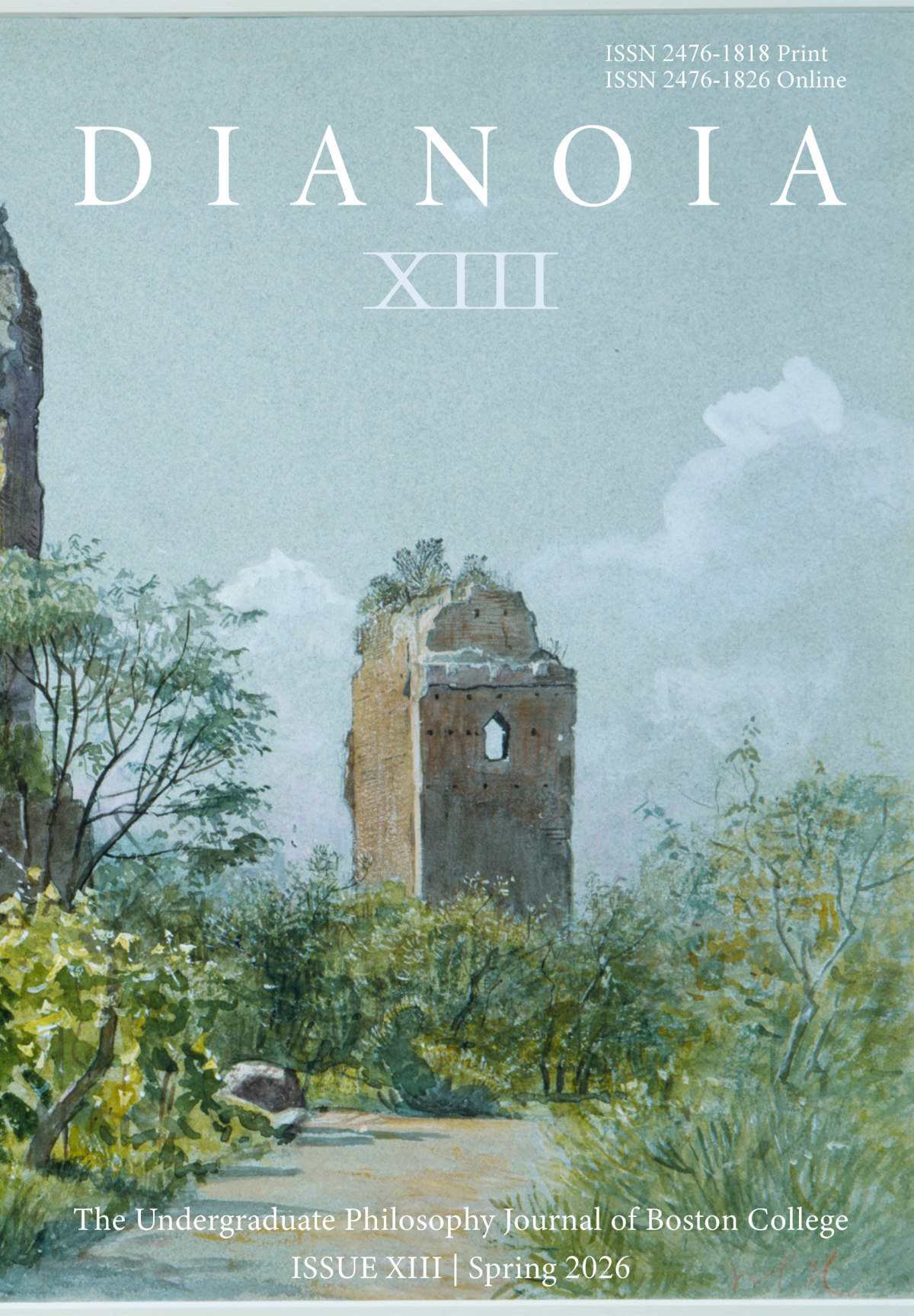
ISSN 2476-1818 Print
ISSN 2476-1826 Online

DIANOIA

XIII



The Undergraduate Philosophy Journal of Boston College

ISSUE XIII | Spring 2026

DIANOIA

THE UNDERGRADUATE PHILOSOPHY JOURNAL OF BOSTON COLLEGE

ISSUE XIII
Spring 2026

Cover Art Featuring:
Haseltine, William Stanley. Baths of Trajan. 1882. Obtained via The Met Museum.
https://www.metmuseum.org/art/collection/search/11021.

If you have questions regarding the Journal, would like to submit your work for review, or if you'd be interested in joining next year's staff, please contact the Journal at dianoia@bc.edu.

Dianoia
The Undergraduate Philosophy Journal of Boston College
Spring 2026 Issue XIII
Boston College
140 Commonwealth Avenue
Chestnut Hill, MA 02467
https://www.dianoiabc.org
https://www.bc.edu/bc-web/schools/mcas/departments/philosophy/undergraduate/dianoia.html

CONTENTS

ACKNOWLEDGEMENTS

The Editorial Staff of Dianoia: The Undergraduate Philosophy Journal of Boston College would like to extend our sincerest thanks to the following individuals for their assistance in making this issue of Dianoia possible:

Sean Haefner, Paula Perry, Chris Hanlon, Philosophy Department

Gabriel Feldstein, BC Libraries

We would also like to thank Dr. Ronald Tacelli, S.J., of the Boston College Philosophy Department, for his invaluable assistance as our advisor and editor of this issue's faculty interview.

Dear Reader, March 27, 2026

I am happy to present to you the long-awaited Issue XIII of Dianoia: the Undergraduate Philosophy Journal of Boston College. This year we received over 170 submissions from colleges and universities across three continents. Along with printing this edition, we received all five authors at our Annual Symposium where they presented their research to Boston College's students and faculty and celebrated their accomplishment with the Editorial Board.

Herein you will find five excellent papers, a short preview will, I hope, entice further reading. In thinking about the future, we present papers on population ethics and the nature of phronesis in Artificial Intelligence. You will also find a much needed interpretation of Nietzche's Eternal Return and an analytic analysis of self-interested action. Lastly, we present a convincing paper on an underappreciated philosopher and timeless topic: Iris Murdoch on love. We hope that these papers inspire new insights and new imaginations.

In addition, I am privileged to present two interviews in this issue. First, Editor Peini Feng and I paid a visit to Emeritus Political Science Professor Susan Meld Shell to discuss her time at Boston College and lasting scholarship on Kant. Second, Editor Claire Swanby and I graciously welcomed this year's Gadamer Chair Professor, Sara Heinämaa, where we discussed the importance of a phenomenological method for emotions and values.

The editorial board chose Baths of Trajan (Sette Sale, Villa Brancaccio, Rome) by William Stanley Haseltine for the cover art of this year's issue. This painting depicts the ruins of the Baths of Trajan overgrowth with lush vegetation. The ruins harken back to a forgotten history and the effects of civilizational decline, while the vital force of life symbolizes a thriving for the hope of a future not yet. Is this not the very task of philosophy? An oscillating ressourcement to ideas long forgotten and the breath of life that each reader can bring them. Dianoia continually seeks to enrich its readers with the sources of inspiration, the persistent questions, and the spring of wisdom that slowly nourishes our souls and civilizational decay. We acknowledge that these sources of inspiration come not from above, or below, but as it were, from the not so numerous center.

I offer four years' worth of thanks to our faculty advisor, Fr. Ronald Tacelli, S.J., and our graduate advisor, Sean Haefner. In addition, I want to thank our Senior Managing Editor, Michael Yost, Managing Editor Xavier Lafaire, and Paula Perry who assisted operations behind the scenes with eager aptitude and persistence. Finally, I am grateful to those members of the Editorial Board who devoted their time to this institutional endeavor.

I hope that this issue's contents are intellectually stimulating and personally enriching and that you continue to offer your support for Issue XIV next spring.

Sincerely,

ELLIOTT R. JONES
Editor-in-Chief
Dianoia

AN INTERVIEW WITH GADAMER CHAIR PROFESSOR SARA HEINÄMAA

Elliott R. Jones: Professor Heinämaa, you were announced as the Gadamer chair professor in 2021 and we're very glad that you're finally here, though this is a very snowy winter. Your work deals with the phenomena of embodiment, intersubjectivity, temporality, normality, emotions, and sexual difference. You described to me in a former conversation that your previous research interests were heavily analytic and focused on debates concerning artificial intelligence as well as the philosophy of mind. And yet, it was in conversation with Martha Nussbaum that you said you decided to transition to new research interests. Could you describe her influence in prompting you to transition into these new interests and why you ultimately took her advice seriously?

Sara Heinämaa: Well, actually Professor Nussbaum has given me philosophical guidance about direction several times. I was a young PhD student, I think I was 27 or 28, and was working with the philosophical problems of artificial intelligence and cognitive sciences. And it was then that I met Martha for the first time in a workshop on Aristotle's reception. She has the delightful habit of asking students about their philosophical interests, and when she asked me, I told her that I was studying contemporary philosophy of mind, focusing on Jerry Fodor's hypothesis of a language of thought and the related idea that all thinking is computational. I happened to also mention that I was reading Simone de Beauvoir's *The Second Sex*, just for fun. When I mentioned Fodor my voice was quite apathetic, and then,

when I mentioned Beauvoir, it became full of enthusiasm. And so she said to me, "Why don't you just follow your intuitions and what calls you in philosophy. Why don't you see what comes out of it?" And I changed my PhD topic, immediately without hesitation. Luckily my supervisors, Professors Lilli Alanen and Leila Haaparanta were both positive and supportive about continental philosophy and feminist topics.

And then I had to do a lot of phenomenology—especially with topics of embodiment and intersubjectivity, in order get at the philosophical core of Beauvoir's argumentation in *The Second Sex*. For this purpose, I did a lot of philosophy of selfhood and studied the self-other relationship, how different philosophers make sense of it, not only ethically but also ontologically. Now Martha came to Helsinki every year for workshops and conferences organized by Finnish colleagues, originally and most importantly, Professors Simo Knuuttila and Juha Sihvola, who were experts of Aristotelianism and Stoic philosophy. The philosophical topics of these meetings ranged from philosophy of mind and ontology to ethics and political philosophy. By the turn of the millennium, our workshops started dealing with political emotions because that had become one of Martha's central research topics at that time. (She has since authored several volumes on political emotions, *Hiding from Humanity* 2004, *Political Emotions* 2006, *Anger and Forgiveness* 2016, *The Monarchy of Fear* 2018.) The speakers and participants of our meetings represented many different branches of philosophy, very broadly, from the history of philosophy and analytical philosophy of mind to pragmatism, phenomenology and Foucaultian archeology.

These discussions gave me a new interest in emotions. I had been studying emotions, independently of Martha's influence in the 1980s and 1990s, since both my teachers, Professors Alanen and Haaparanta, were historians of philosophy, Lilli specialized in Descartes and early modern emotion theories, Spinoza, Hume and Leibniz; and Leila specialized in classical phenomenology, and especially phenomenological accounts of logic and mathematics. And then, thanks to Martha, I found my way back to the topic of emotions and started to really study it, from the roots, both historical and systematic. But at this same time, I happened also to meet a young PhD student in Leuven, Tomas Sinkunas is his name. He was working on disgust and

related emotions. We had a small bar discussion, and what he told about his work on disgust—mainly based on his PhD—was deeply intriguing in its phenomenological tenor and political-philosophical implications. He published this work later. It deals mainly with the analysis of disgust by the Hungarian phenomenologist, Aurel Kolnai.

Kolnai was a Hungarian Jew who had to flee Hungary when the Nazis came to power. He wrote an early book, *The War Against the West* (1938), a political-philosophical analysis about the new type of totalitarian formations that started to happen in Europe; but he fled to England, and published there several important articles on ethics and emotions, including a small paper called, "The Standard Modes of Aversion: Fear, Hatred, Disgust." It came out posthumously and quite late (in 1998), in the highly esteemed journal *Mind*. It deals with three aversions: disgust, fear, and hatred. I was struck by that article, its originality and boldness. One of Kolnai's arguments there shows that we can't theorize emotions in an analogously general manner as we can theorize beliefs and cognitions. We have to look carefully into their individual intentional structures and compositions, so their mental makeup, so to speak. Fear, for example, is structurally very different from hatred and also from anxiety—even if it also has similarities to both. And disgust again is very different intentionally from the other two aversions discussed by Kolnai. So the richness of the intentional structures of emotions showed itself to me.

Kolnai writes that the three emotions have very different kind of targets as well as different manners of relating to their targets. But he also shows they differ importantly in their temporality. They have very different forms of duration: Hatred can stay with you for years, even decades. Disgust and fear, in contrast, are occasional: disgust strikes you when you are exposed to certain kinds of things or materials in perception or imagination or perhaps memory.

So, yes, from these two inspirational sources, Martha's discussions of the political aspects of emotions and then Kolnai's analyses of aversions, I started to work again on emotions, this time very seriously. When I was invited to Boston College, I had already developed a new lecture series on the phenomenology of emotions and given it and pieces of it at home in Finland and other European universities. I thought that perhaps I can present

these topics here, too, since they have such contemporary interest and broad implications; you can see their relevance for us in what's happening every day.

ERJ: What was the reception of Kolnai's work? Was it popular at first or were you part of the people who revived it?

SH: He has been largely forgotten. A great mistake! However, two philosophers, Barry Smith and Carolyn Korsmeyer, translated Kolnai's work on disgust in 2003 into English, and this drew some younger phenomenologists to that topic. There are some American philosophers who know about his analyses, Colin McGinn, for example, but he's not well known here. I do think his work is now becoming more and more attractive because the topic of political emotions is now very real and disturbing for us, both practically but also philosophically. So his analyses of emotions and volitions are perhaps coming to be appreciated—certainly more than it has been for a long time. Which is nice, because his approach can be combined with both psychoanalytical accounts, like Kristeva's on the abject, and with cognitivist accounts, such as Nussbaum's. Kolnai's account of aversions differs from both alternatives, but it's not incompatible with either, and I believe that new combinations will prove powerful.

ERJ: Who did he study under?

SH: Honestly, I don't know. He was quite an original thinker; he was not really a rebuilder of phenomenology, like Scheler or Heidegger, after Husserl. It was more the case that he used the phenomenological methods of analysis, and began using them in his unique way for new intriguing questions. Another topic he treats—a very different one—is games and gaming. This is central to his philosophy of the will and willing. Moreover, his analyses show that volitions have crucial constitutive roles in many emotions, such as hatred and love, but not in all.

Claire Swanby: Interesting, so this semester you're teaching a course titled "Political Emotions: Phenomenological Analyses of Disgust, Hatred, Desire and Love", and it sounds like you had a vision for this course after you were announced as Gadamer Chair in 2021. I was wondering if that vision changed at all between 2021 and now, and, if so, what influenced those changes.

SH: Well, since the 2010s, I'd been giving lectures on specific emotions in Europe—hatred and hate-speech, for example, and disgust and its implications, both ethical and political. At the same time, I have written quite extensively also on love, wonder and curiosity, and studied their roles in ethics and political life. When I started to put the individual lectures and papers together into a course, I realized how very fruitful it is to compare their intentional and temporal structures, and do so open-mindedly without pre-established categorizations. Such comparisons display the distinctness of emotions and allows us to understand their specific roles and powers in our life. I have also compared hatred with some near-by emotions, such as anger, rage, contempt, indignation... The focus was not on those, but for comparative reasons it's important to look at both nearby emotions as well as distant ones. Hatred, for example, turns out to be quite different from the whole anger family, so from anger, indignation, and rage, both in its intentionality and temporality. The latter are all occasional, but hatred is persistent, abiding, even patient.

So, I had already decided I would teach on the topic of political emotions, but instead of discussing just desire and love on the "positive side", I decided to bring in also curiosity and wonder. So the course became a discourse on disgust, hatred, curiosity, and wonder. Namely, I think—am really deeply convinced—that if we want to develop an emotive life which is able to respond to the "negative emotions" or aversions, which seem to win more and more social space in our time, then we definitely need to develop our skills of being curious. It is the curiosity family —wonder, surprise, admiration, marvel, generosity—that is most important for us today, and much more so than some other constructive emotions, say sympathy, compassion or tolerance.

So, yes, the idea of the course changed a little in development. I also think that if one builds from curiosity, then one's understanding of love and its obligations becomes less naive. Curiosity namely also entails the task of self-critique: with it, you can't just act out your loves and think that you're a largehearted person; you also have to take a critical look into yourself and ask why certain things and person remain out of your sphere of love, or our spheres of love. What might still be curiosity-raising in them… even if they are not lovable to you… to you as I know yourself and have learned to think about yourself? So as a cure

for our emotive weaknesses and stubbornness, I think curiosity is urgently needed.

ERJ: Do you have any doubts about it—the strength of curiosity?

SH : No, I think it is powerful emotion, if it is allowed to flourish. But how we use it depends, of course, on how we think about it, how we conceptualize it. There are many alternatives here: You can start from Plato and Aristotle, from Plato's *Symposium*, where *eros* is tied to curiosity and knowledge. Or alternatively look into Descartes's *Passions* and the other early modern sources where it entails a suspensive moment. Or perhaps Kant or Kierkegaard? But at the end, you should bring this all together, for the analysis of our present.

ERJ: I talked to one of the students who is attending your course, and, judging from his comment, I'm sad I did not get a chance to sit in on your lectures. One thing he said that struck me was that you often "pictorialize" the philosophies and ideas you present through works of art. This reminds me of your philosophical articles which are littered through with references to literature, movies, and music. Could you describe how you use art in your lectures and the role it plays in your pedagogy or your thinking?

SH: Well, I think my interest in this comes from the American philosopher Charles Sanders Peirce. He argued that not all thinking is discursive; there are important modes of thought which operate by non-linguistic processes and structures, such as diagrams and models. I believe that this holds also in philosophizing, and that there are different kinds of philosophers in this respect. I'd say that my thinking is basically a bit diagrammatic: diagrams, maps and geometrical and topological structures help me capture connections, make distinctions, and see larger wholes. Images and moving pictures, in particular, allow me to reflect on emotions, but then of course one needs to put it all into a discursive form, into an argument, a thesis or a train of thinking. Images serve as tools that advance and support the discursive process which one cannot avoid if one wants to publish articles and books. You can't send a series of pictures to a journal; you have to explain what they signify.

I don't think such tools help everybody. Rather one needs to find out if one is a discursive thinker rather than a pictorial thinker. Originally I wanted to become a graphic artist and almost became an architect, so I think there is a historical root

in my use of pictures, images, and diagrams. I ended up being a philosopher, but writing was horribly difficult for me at first. In high school I always got an extra hour for writing assignments. If we had to write an essay of five or two pages, my teacher (a very nice person) would say, "Okay, just stay and finish," because I couldn't produce on time. It was horrible. I hated it.

ERJ: About the relationship between art and emotions, you do think that music, say, or movies really train the emotional life….

SH: Yes, I do believe so, and I was convinced by Martha Nussbaum about the strength of imagination, productive imagination, especially in cultivating the emotive life. But for me personally it's more about the fact that I *need pictures for thinking* about anything. From the beginning, it's really been kind of normal business for me, and I hope it helps the students and doesn't distract them. One has to be a little careful with it.

ERJ: And now on to some more philosophical topics. Your articles that deal with Husserl, have been extremely clear and insightful for me, even helping me to analyze my own interior life. Your recent work focuses on the intersection of phenomenology and ethics—with a particular focus on emotions and values. Could you help me solve what seems to me like a paradox, which is this: What can phenomenology, as a 'descriptive science', say about ethics, which is a 'prescriptive science'?

SH: My phenomenological commitments and interests are very classical and existential, that is, Husserlian and Merleau-Pontian. I contribute to contemporary philosophy, to philosophy of intersubjectivity and embodiment, sexual difference and emotions, and philosophy of the sciences too, but I use classical phenomenological methods in all these enterprises and am very happy to do so. I find these methods invaluable, crucial for the purposes of contemporary thinking.

Phenomenology is a strange field in philosophy, because there are several different manners of doing it, and not everybody agrees on how it should be done. So the Heideggerians do their thing (though there is a common ground) and the Levinassians do another thing (and again there's common ground), but what I say holds primarily for the Husserlian approach. And in *that* the task is to understand and analyze experiences, to discover their intentional and temporal structures.

The first result here is that we distinguish conceptually between three modes of intentional consciousness. One form of intentionality is *axiological*, and it deals with valuing and with different kinds of values (e.g., personal values of love, objective values, the value of truth, beauty); another one is *practical* or *praxeological*, and it deals with different modes of volition and with norms, goals and means as their objective correlates. And then the third mode of intentionality is what Husserl calls *doxic* (which comes from *doxa*, the Greek word for opinion or belief). This is about believing and all forms of cognition, convictions, assumptions, presumptions etc. Most emotions, interestingly, combine all these three modes of intentionality: they relate us with what is good, beautiful or valuable in some other sense, but at the same time they detect such properties in realities and beings of other sorts (ideal and fictional); and on the top of this, they also motivate our goal-directed actions. Emotions in many cases are intricate combinations of axiological, practical, and doxic modes of intending.

So we can say that they have a robust structure of intentionally which brings together these three aspects. The task is to distinguish and describe these structures and their variations across all emotions, and to do for the purposes of ethics, political life, and epistemology. This brings us to the normative and practical implications of emotions. Some of them, curiosity and responsibility for example, concern ourselves also as scientists *and* philosophers—which is why, when you start to study the possibilities of emotive and normative-practical intending, you soon realize that your own attempts to make sense of all this is also guided by certain goals and motivated by certain valuations.

I think that such self-reflective studies brought Husserl to understand that his philosophy is not just a theoretical enterprise with practical implications, but also a practical and existential commitment to *theoria*, a certain kind of vision of the world *in toto*. On that level, the two dimensions, the theoretical and the practical, cannot be separated. Of course, for analytical reasons they have to be conceptually distinguished, but they need not be separated into two oppositional, parallel or alternative regions of life, so that we would neglect the practical consequences of our theorizing, or try to push theoretical things aside when involved in some practice.

So I think that for Husserl the practical *and* the theoretical, the prescriptive *and* the descriptive grew together step by step in his investigations; and when we come to the 20s and 30s, they are already…well, not *fused* together, but related in a new way, freshly intertwined. Not by contrast or opposition, but by mutual excavation and deepening. Due to their egoic underpinnings, the descriptive, theoretical and analytical *always* drag with them and motivate normative questions, questions that concern one's own commitments and responsibilities: What should I do now, today? In this way, phenomenology becomes a normative enterprise.

ERJ: Is a goal a value?

SH: No, goals are not values, but the goals that we are striving for are values *for us*. So, yes, there is valuing that motivates the striving for the goal and working for the goal. And, of course, the amounts we're willing to put into our attempt to realize goals depend on how much we value them.

CS: I had the privilege of attending your Gadamer chair lecture, which was focused on the phenomenological method of *Besinnung*—which seems very similar to what we were just discussing—and its potential to ground and orient the philosophy of science. Could you please clarify for me and for our readers what *Besinnung* is? And also: Is *Besinnung* something that philosophers engage with at certain critical moments, or is it an ongoing attitude?

SH: That's an excellent question. I would say that it's *both*. It's a praxis, which is an ongoing process. But Husserl also interestingly argues that in times of crisis (be they theoretical crises or crises of other kinds, like political or ethical), we are as if *called* to take care of our commitments in a *different* way than usual. Oftentimes it's just business as usual: we know what to do and how to do it; what our specific goals are, and which one should be taken care of first and which later. But in times of crisis the whole praxis, which consists of several different goals, tasks, and subtasks, and connects several different agents, is brought into an unexpected and uncontrolled movement, and those times, he says, call us to reflect (*besinnen*) the praxis as a whole.

This means that we need to think through quite deeply to what we actually are committed to, and what might be additional assignments that during the years of working have been connected

to and associated with our deepest commitments, sometimes challenging them or hindering their realization. These additions might have served our work at some point, might have been part of business as usual for us, but in times of crisis we have to find our *core* goals in order to be able to save the praxis—if it's worth saving. So *Besinnung* is both a mode of critical reflection and self-reflection that stays with us as a capacity, a cultivated and habituated capacity, but it is also something that needs to be performed at certain moments, if the project or the praxis that we are sharing falls into crisis.

Let me add just add one thing. Because Husserl started with deep epistemological goals, his work at first was focused on questions on knowledge and the sciences; he started by asking about the structures and forms of theories, theorems, and sciences—to study how these are structured by sense. So phenomenology, at the beginning, was a theory of theories. But in the 20s and 30s, which is the period I'm looking into, it has become, as we've already talked about, also a critical study of *practical* intentionality, that is, intentionality of willing and valuing. Here we need to ask how our scientific, scholarly, and theoretical goals can be practically realized, if they can be, and on what conditions. This emerged as a very concrete question for Husserl in the 20s, because various kinds of quasi-scientific and bureaucratic discourses started to bloom within German universities and these gave rise to many pseudo-scientific activities which….

ERJ: Sorry, what would be an example of these 'quasi-sciences'?

SH: What comes to mind here are examples from Nazi Germany: eugenics and craniometry, for example. There were projects that measured the human skull and facial bone features and tried to classify, rate, and improve people on that basis… But this was not just some Nazi invention. Anthropology in the 20s was developing in that way more generally, in many countries, also here, trying to measure and classify—trying to make a hierarchical evolutionary typifications of the people we find in the world. So something was happening in the biosciences and humanities, something that went beyond their proper aims, something strange. And Husserl was aware of that.

ERJ: Something just came to mind. Could transhumanism fit with what you're thinking, trying to sort of modify the....

SH: Yes, I think that's a very good example, from our own time. So, yes, it seems to me that *Besinnung* is an invaluable method for sciences in times of crisis: also our own time. And Husserl gives us a kind of rulebook for it. He shows how to do it, but he also gives us concrete examples: critical reflections on the goals and norms of logic, mathematics, psychology, ethics... And his followers, Merleau-Ponty, for example continue by reflecting on the norms and goals of the biosciences and life sciences.

CS: I was wondering, in your view what is it that's *new* within phenomenology, and why is it especially important and relevant today—this question could be about phenomenology in general, or your own work in phenomenology in particular—and what do you think might change in philosophy or public thought if this approach were more broadly integrated into research and into debate?

SH: I think there are several things to say here. When I came to phenomenology, what specially interested me and my colleagues at the time—the mid-90s—was intersubjectivity—not just what the subject is, or how the ego relates to this or that objectivity (which had of course been and needs to be part of phenomenology), but more about how experiencing subjects and embodied egos relate to one another, and what kinds of communities they are able to form: communities of language, community as *polis,* religious communities such as the church, and so on.

The other topic, embodiment, was really not taken up much at all within philosophy—or it was relegated to the philosophy of perception. But now it turns out that embodiment is important to many philosophical enterprises, philosophy of action and joint-action, communication, political philosophy, aesthetics, and ethics. So the topic of embodiment became more central, and its potentials and implications were discovered in the mid-90s. Those two topics—intersubjectivity and embodiment—certainly, became crucially important, and people here in America and Europe started to work on them at pretty much the same time. We also found sources in the tradition that had been forgotten or neglected. Manuscripts of Merleau-Ponty and Husserl, *important* manuscripts about these topics, were discovered, not just one or two but numerous. Experts of course knew about them, but they had not been studied systematically

and not used for philosophizing; so secondary literature had been weak.

But what was very important to me then, and still today, in addition to these topical issues, is that philosophy should be a *systematic* enterprise and not merely an *exegetical* one. I was educated by Wittgensteinians, and of course they did study Wittgenstein also exegetically, interpreted his texts and debated about correct interpretations. But even so, there is in Wittgenstein a *very* strong emphasis on the task of *systematically* solving philosophical *problems*, dealing with *problems*, finding them and formulating them. When I say "him," I mean conceptualizations invented by him, structures of argument that came from him, that he formulated.

So I was educated to cherish that systematic part of philosophy by my teachers. And that's why it's important for me today in phenomenology not to let what I do become an exegetic or philological enterprise. I don't mean that exegesis isn't important; it's absolutely necessary. We have to have it. We need scholars who know what is *really* being said. But we also need, and necessarily so, also philosophers who ask what can be said, what can be thought and conceived.

To summarize, as a common enterprise, philosophy needs exegetic work and historical inquiries. But if it lets go of systematic questioning, systematic manners of inquiry, systematic search for connections between concepts, questions, and arguments, as well as theory formation, then it becomes, I think, either philology or history—which are good too!—but not philosophy. So I think our discipline, for essential reasons, has to be able to combine exegetic work with systematic thinking. Perhaps it's specific to me that even if I do a lot of exegetical work, I can't let go of systematic inquiry.

For example, I have put in a lot of work in order to understand what Husserl says about love. But then I think that such questions have to be put aside and we have to ask: What *is* love? And: Was Husserl *right?* And what about Plato—how did he deal with the *topic*? And what about Descartes—how did he *define* love? Did they *discover* something in their inquiries? At the end, all these questions need to be brought together.

ERJ: I agree. I think the more systematic view is more creative, more artistic. You take an idea from exegesis and then *work* with it.

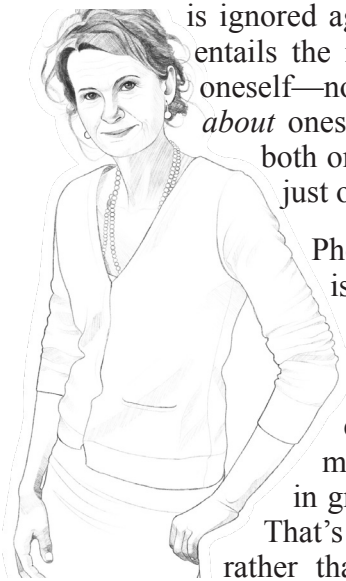
SH: Yes, you also need to *elaborate*. You reform, or you combine, and you answer to the present and its own problems—with the help of newly formed concepts or whole theories.

ERJ: On to Husserl's idea of love. In your article, "On the Beauty of Persons, How to Love Value Producers", you develop Husserl's ideas on values to a novel account of persons. You argue that we "experience a person's value as good purely for the sake of the beauty of their appearances."

Thus good-values and beauty-values appear in the encounter with another person, such that we can say that they are not just beautiful, not just good, but are instead *beautifully* good. Now there is a certain attitude required for this. I want to ask: Is this attitude natural or habitual? Is it taught? Does it come from *us* or from the *Other*?

SH: I think it's given to us. It's a capacity in us. As such it can, of course, be ignored and left undeveloped, and if it is, we are in great trouble.

It *is* developed today, but only to a certain extent, not enough. More precisely, it is not completely neglected, but it's not at the center of our attempts to relate to others. Moreover, what it is ignored again and again is that personal love entails the readiness to challenge and criticize oneself—not in a manner of passing judgments *about* oneself, but in a manner of questioning both one's habituated conceptions, but also just one's habits of relating to others.



Phenomenologists thematize what is called *suspension*: a suspensive interruption of habituated activities which always involves *possibilities* for us. Today, academic institutions emphasize *production*—producing more and more, faster and faster, and in greater and greater detail or accuracy. That's come to be the emphasis of our time, rather than taking advantage of suspensive

moments, intervals and pauses, and look back to what we have *presupposed* during our theory construction and productive work, so that we can then perhaps see something else, or see things differently. And I think a similar suspensive moment is also a crucial possibility in our ethical relationships to one another and to ourselves. Even here we tend propose fast solutions, formulate rule books and general guidelines, and try to do so as efficiently as possible. We feel we need to solve the most often repeated problems first and then move on, rather than giving ourselves time to pause, let the persons involved show something about themselves, which may be irrelevant to our own concerns—or, so we think, to the situation at hand.

So, yes, I think love or loving wonder is *both* a natural capacity and also something that can and must be cultivated. And it can be taught; someone can teach us how to do it better, so we can mutually train one another. There may be masters who model the skill, so we can learn from the masters. It's *in* us, but we also need the other person in order to get all we can out of it, to really put it into practice.

Of course we also need to have curiosity about *ourselves* and be able to wonder at ourselves, and then find something 'beautifully good' in ourselves. But the great adventure, I think, is that everybody is different, and we can't assume that *this* person would have to be 'beautifully good' in an exactly similar manner as *that* one.

ERJ: Do you think this relates to what you *mean* by curiosity? It almost sounds as if curiosity and critique are very, very, similar to each other.

SH: Yes, they are similar, they have similar ingredients or phases. I think they share a structural feature, which I would call a 'suspensive moment.' In both, you need to be able to suspend something which is *already* in operation in you. Say you're studying the world, either experimentally or reflectively, and in order to do that you need some concepts, some basic beliefs, some epistemic commitments. They all have to be operative. But at the same time you're also able to *suspend* all that, all this conceptual machinery, interrupt it for a moment and see how differently the world looks if you don't let it guide your investigations. In emotive curiosity, I think, there's a similar suspensive moment.

ERJ: I see. In that suspensive moment, would you distinguish between doubt and wonder? Because it seems to me almost like a method of doubt. I'm wary of methods of doubt, but I'm open to a method of wonder.

SH: Despite similarities, there are also important differences between doubt and wonder; and doubt, I think, presupposes wonder. In order to doubt, in a pregnant sense of the word "doubt", in a *heavy* sense, you have to put a belief or an emotive commitment into question. A specific, targeted object that has to be questioned, possibly negated. Wonder, in contrast, is an open state. It doesn't choose a target purposefully for preestablished goals and it does not prepare us for affirmation/negation, acceptance/rejection. Rather it's a constant possibility that exists within us of taking a *pause* from our activities and interests. Now one of such activities is that of believing—or maybe committing to some belief, instead of taking a step away from it. Doubt is an interruption of believing for the purpose of taking a position, it's more targeted and purposeful than wonder, more interest-related. And, as said, I think that it entails, as a moment, what I would call "wonder" and "curiosity".

ERJ: Brilliant.

CS: This reminds me a lot of what the poet John Keats calls "negative capability", which he defines as "[A capacity of being in] uncertainty, Mysteries, doubts without any irritable reaching after fact and reason."

SH: Yes, I know his poems. And yes, definitely, what you suggest is intriguing. I need to look into that. That's fantastic.

CS: And that makes me wonder if maybe there's also a connection to the creative life here.

SH: Yes, I think the masters of wonder are often the artists, and also the thinkers, who are able to re-evaluate and change their projects, and so renew themselves. They seem to me to be very good at this—not *abandoning* what they've done, but taking a new attitude toward their achievements and thereby steps forward.

I think both artists and thinkers need such moments of pause in order to renew themselves and their activities. The creation and discovery of something new and original, something not

yet encountered seems to be dependent on this. And of course, poets—I mean, they are the masters of all masters in this, because the material that they're working on is language, and many other activities depend on the activity of language use. So poets study the basic activity that necessarily needs to be questioned, if *other* discursive activities are to be questioned: they question habits of language usage and linguistic meaning, ask how phonemes can form words, and how they can be understood.

CS: In your article, "Varieties of Love, Intentionality, Temporality, and Agency," you introduced the concept of the transitivity of loving care to describe how, "love transitively ties us not just to our objects of love, but further also to the objects of love of the beloved." And I was wondering, to what extent does this transitivity of loving care suggest that loving carries with it an implicit ethical orientation beyond the immediate relationship?

SH: Yes, I think transitivity is at the very core of the ethical dimensions of love. And the problem is how to make sense of it properly. There are many problems here—in the first place, counterarguments built on case studies where 'transitivity' doesn't seem to work. I have debate this with Alva Noë and Kate Kirkpatrick, and many other philosophers interested in love. My transitivity argument is still very much a work in progress, so it needs to be developed. But I'm convinced that something *like* this is a real phenomenon at the heart of our intersubjective relations, and it needs its own phenomenological analysis of intentionality and temporality. Plus, I'm quite convinced that an ethics can be built with the help of the concept of expanding circles of love. But how adequate that ethic is, whether it's *sufficient* and can stand on its own feet without any help from any other kind of ethical concepts or considerations—that's still an open question. But definitely, this transitivity and the enlarging circles of loving wonder, this very specific kind of loving, which is grounded in our deepest commitments (in plural)—I'm very optimistic about the possibilities of the building a pluralistic ethics on that basis. And not only do I believe that an ethics can be built on that basis; I think we *need* it.

CS: I think it's also important to note your contributions to feminist philosophy. I was wondering how phenomenology has shaped the way you approach feminist philosophy, or the other way around. And then broadly, how does the phenomenological

approach enrich our understanding of questions of femininity, embodiment, and sexual difference?

SH: It was a kind of revelation for me, when I studied the works of classical and existential phenomenologists and realized that they *all* discuss sex! Not sexual activities or orientations, but "sex," meaning the difference between men and women. Not only Simone de Beauvoir, but Sartre in *Being and Nothingness*, when he asks: What is this? What is this phenomenon, strangely between necessity and contingency? Merleau-Ponty as well. They were all struck by this phenomenon. And I think the basic insight that they give is this: There is a phenomenon which is neither *just* psychological, nor *just* physiological. There is a phenomenon which is neither natural, which doesn't simply come from physical or biological nature, nor simply cultural, but is *both* psychological *and* physiological, *both* natural *and* cultural. It's *cultivated nature*; a *second nature*. These philosophers don't call it *gender*; instead they call it *sexual difference*. And I think it shouldn't be called gender, because it's not merely cultural, not simply socially constructed, not an artefact. As said, it's *cultivated nature*.

Traditional feminist philosophy, the one that begins in the 18th century, or actually much earlier, in Renaissance, talks about ethico-political matters, such as justice, subjection, and violence, not in terms of gender, but in terms of sexual difference. This is conceptually crucial. Even if we would say at the end of the day that it's not an adequate manner of approaching these matters—justice and violence for example—, and that gender concepts are better or broader, it's still very important to compare these two ways of articulating the field of the phenomena and take a critical attitude to both.

And I think this kind of phenomenology has nicely grown and developed during the last 40 years. I mean, there is now a well-known and attractive subfield of phenomenology which can be called "feminist phenomenology". And it's also now branching out into intersectional questions about how race and gender, or rather racial and sexual differences combine, and further how religious, ethnic and linguistic differences relate to racial and sexual differences. So I'm really happy about the current state of things—there are new developments all the time.

CS: You've written on vocation in the past, and this idea is very much in keeping with Boston College's liberal arts formation, with *Cura Personalis*, and with Ignatian Discernment. For our readers, and especially for rising academics, do you have any reflections on vocation as it relates to one's own personal commitments, friends, loved ones, and family, mentors, or God—or as it relates to the academic, religious, or technical life?

SH: Well, I am impressed by Husserl's concept of vocation: that concept actually gives us the possibility of talking about vocations in their full richness and plurality. It allows us to say and think that a vocation can be directed at an activity, at philosophizing, scuba diving or salsa dancing, for example, or alternatively at persons or just one person, say a friend or a lover. Think about Armand Duplantis, Lynsey Vonn, Usain Bolt, and Matti Nykänen! And then think about Anna Karenina, Rett Butler, Jay Gatsby, and Catherine Earnshaw/Linton! But notice that you can also have a vocational relationship to an institution, such as science—a specific one, or all sciences in general—or to a religion or a religious leader, or a vocation to a nation. So the concept covers very broad set of phenomena, all linked to our deepest commitments in life. That, I think, is its strength.

I'm particularly interested in scientific, artistic, and athletic vocations, on the one hand, so deep commitments to practices and activities of various kinds, but also in vocational relationships to individual persons and groups of persons: lovers, friends, colleagues, families, children, students... But it's becoming more and more interesting for me to study vocational relations to *institutions and organizations*, such as the university. The vocation to the university as an institution is not the same as the vocation to the sciences (which you find within the university). They are related, but a different kind of focus and attention guides them. And in order to understand what's happening at our universities today, right now, we need to understand this difference. I think the same holds true about the church. So you can have a vocational relation to the church simply as an institution; you can have a vocational relation to various religious practices; you can have a vocational relation to the persons and groups of persons within that whole. And I think we should see the differences, because, if we don't respect them, we tend to have these absolutely futile disputes and quarrels, assuming that everyone is vocationally guided just the same as we are.

ERJ: Our last question is this: You mentioned that you wanted to be an artist at one point in your life, but you are an academic philosopher now. So what was your own vocational discernment like?

SH: I was seriously committed to becoming a…not a painter, but a drawer, a graphic designer. And I tried to get into an art school in Helsinki (which was actually a very good art school, and still is), into the graphics department. Well, I didn't. And not only that. On the entrance test, where you had to create pictorial presentation of six different topics, on each of them I got a zero points, zero out of five! I was devastated. And I never expressed myself in drawing again for, I think, ten years. But my mother forced me to get into the architect school, because she thought, okay, perhaps you can do this, perhaps you can realize your vocations there. Why don't you try it?

And I told her, I'll take the entrance tests, but even if I get in, I'm not going. I don't want it, I don't want to plan buildings, I want to work with plants and animals. I did get in and with very good test scores. But I didn't continue in the architecture department. I started to study literature and mathematics.

But then some friends in a literature class said, why don't you come to this course that combines mathematics and arts—it's philosophy. And I couldn't turn back after that. The course was about Sartre's *Being and Nothingness*. And it was really kind of wild. What really spoke to me is that I immediately realized that I don't have to *remember* anything. There's no need to remember anything in this field. You can think or conceive everything by yourself, if you're lucky. Of course, you might end up in a school which forces you to study *this* and *this* and *this* and *this*. And then you have to show that you know and remember *this* and *this* and *this* and *this*. But I realized, from the beginning, that with luck you might also end up in a department that says you have to study Descartes today, but if you don't remember it tomorrow, then we can think through the problems again and find something new. You don't have to remember what he said, the focus is on the problems.

ERJ: So would you say philosophy is more about the method rather than the facts or subject matter.…

SH: Yes, it's all about the method. If we must sacrifice something—and hopefully we'll never have to make such sacrifice—then we should sacrifice the facts.

ERJ: Prof. Heinämaa, we thank you very much for taking the time to do this interview.

SH: Such great questions, fantastic questions! You have gained a new reader of your journal.

REFERENCES

Sara Heinämaa, Varieties of Love: Intentionality, Temporality and Agency, Aristotelian Society Supplementary Volume, Volume 99, Issue 1, July 2025, Pages 141–166, https://doi-org.proxy.bc.edu/10.1093/arisup/akaf004

Sara Heinämaa, "On the beauty of persons: How to love value-producers," in *Husserl's Studien zur Struktur des Bewusstseins*, eds. Emanuela Carta and Gabriel Lobo, Dordrecht: Springer, forthcoming.

Sara Heinämaa, "Besinnung – a phenomenological method for philosophy of science," in *Oxford Handbook of Phenomenology of Science*, ed. Harald Whiltsche, Oxford: Oxford University Press, forthcoming.

John Keat, "Selections from Keat's letters": https://www.poetryfoundation.org/articles/69384/selections-from-keatss-letters

AN INTERVIEW WITH PROFESSOR SUSAN SHELL

Elliott R. Jones: Good afternoon, Professor Shell. By the way, this is a beautiful office. You have one of the best offices…

Susan Meld Shell: Yes, I'm really sorry I'm leaving it. [Laughs].

ERJ: So yes, the political science department recently had a conference in your honor titled, "Kant and the Future of the University." Could you please introduce yourself and your interest in Kant, especially as a political science professor?

SMS: Okay, well, that's a broad question, and I suppose my interest in Kant originated when I was in college, and although I was very interested in philosophy, I was in one of these crazy programs that we had back then, when you didn't have to major in anything. So I got to take anything I wanted. And I took a number of philosophy courses. But in those days at Cornell, the department was very analytic in a way that was quite narrow. They didn't do the history of philosophy, which was what I was interested in. So I ended up learning about Kant mostly from an English course on the poetry of the sublime. It was an English Romantic poetry seminar, Milton to Wordsworth, and we got assigned various papers. I was assigned the 'Kant on the sublime' paper. So that was my introduction. And in fact, that course at Cornell spawned a lot of political theorists and Kantians, including Richard Velkley and Clifford Orwin and Nathan Tarcov, and others.

So it was a very interesting class, and not only for English majors. So I think my interest in Kant has always been, well, at least it was initially his theory of the sublime. And I think all of my research begins in

this grounding human experience, which combines the very high and the very low—pain, pleasure, and literary and artistic expression, as well as more conventionally analytic expressions of philosophic thought. At least this was for me a kind of founding touchstone for approaching Kant. And since at the time you could do the history of philosophy more readily in the department of Political Science, I started out studying political philosophy with Alan Bloom and other students of Leo Strauss, who urged a return to the forgotten wisdom of ancient thought. But I've always been very interested in the modern alternative to the ancients. And Kant has always seemed to me a particularly attractive and potent adversary, or at least a challenge, to ancient thought. And that's one reason why I've again and again returned to Kant, though I've also worked on Hobbes and Rousseau and Hegel and Heidegger, but somehow I'm always drawn back to Kant, who's now become a familiar old friend—at least a one-way friend. [Laughs].

There are certain disadvantages to working on one thinker—a lot of disadvantages—but one of the advantages is that you really get to know the nooks and crannies, like, like a close friend. It's nice to have lots of acquaintances, but also one friend you know really, really well. And for me, Kant has been that kind of a figure.

Peini Feng: So now, turning to Kant's philosophy.

SMS: Yes.

PF: You just mentioned that Kant is a kind of modern alternative to ancient philosophy.

SMS: Yeah.

PF: And for most of the students, when they begin their journey of philosophy, they begin with Plato's *Republic* in their introductory courses. And as you said, Kant is an alternative to ancient philosophy, so I guess it includes Plato. So, how would you compare Plato's philosophy to Kant's philosophy?

SMS: Well, then we get into this difficult question of, who's Plato?

PF: Exactly.

SMS: And again, my first, actually my very first, philosophy course I took as a high school student in Seattle at the University of Washington…it was a wonderful course, but it was taught by a professor who took an analytic approach. And what we basically did

is went through Plato's dialogues, a few that we read, and found all the logical errors he made. You know, all the bloopers, highlighting how much smarter we are. It always struck me as unlikely that I, a 15-year-old, knew better than Plato. So, I mean, that's one question: which or whose Plato? And I suppose the interpretation that I have found rather compelling is a certain version of the Straussian reading of Plato and Strauss in particular, who is another figure I've worked on a lot, including an edited translation of his correspondence with Gerhard Krüger, who was an assistant of Heidegger in the 20s. He and Strauss, as young men, were very close friends. They were in the same Weimar milieu as such figures as Hannah Arendt, Gershom Scholem, and Hans Jonas. These were all people growing up in the same, the same atmosphere in which Neo-Kantianism was the convention of the day. And so, they were all rebels. [Laughs]. They were anti-Kantian rebels, and Strauss carried out a correspondence with Krüger through the 30s and the war years. Strauss, by then, was an exile from Germany. But again, that would be another way of framing the Plato versus Kant debate, because Krüger, who remained in Germany, was a self-declared Kantian, though a kind of Augustinian Kantian. There were just many Platos and many Kants. [Laughs]. I hate to complicate the question, but if you want to understand how both of them became the mature thinkers they eventually became, you can look at that correspondence where Kruger is turning away from Heidegger toward a more Christian, Neo-Platonist understanding of Kant. And Strauss, for his part, is discovering the esotericism of medieval philosophy—Alfarabi and Maimonides particularly—but, eventually, Plato as well. So I would say that, in a way, my whole career has been bound up with variations on the debate between Plato and Kant, but again, with an openness to the possibility that there are, there are a number of compelling readings of both figures, and that in itself takes on a history of its own. It takes on a life of its own because we're all, in a way, interpreters, right? Some, some of us may eventually become philosophers, full-blooded sense, but I don't take that as really a professional term, so much as an honorific indicating a certain kind of penetration that I wouldn't pretend to have myself, but certainly has been an object of study for me and for many others.

PF: To ask a specific question…

SMS: I'm being evasive. [Laughs].

PF: No, it's very good. It's very good. A specific question that Plato cares about, which is nature. Philosophy, for him, at least one reading for Plato is that philosophy is an investigation of the whole of nature.

And for Kant, he also has very interesting comments on nature, such as his astronomical works. So how does Kant understand nature, and how might his understanding of nature be different from the ancient understanding of nature?

SMS: Well, that's a great question. Again, a really difficult question. But broadly, there are, again, many ways of understanding Plato and Aristotle and to what degree they were committed to a what, at least on the surface, appears to be a somewhat naively teleological understanding of nature. There are, I think, some very compelling readings of Plato that would not necessarily link him so directly to a kind of naive teleology of that kind. But, however that may be, certainly for Kant, what Kant takes on board is what I think most people nowadays willy nilly take on board, which is, at least in practical terms, a modern scientific view of nature. And when you have that view in mind, all kinds of moral and political problems immediately emerge, one being that freedom seems slightly impossible, except in this very tenuous way that the Heisenberg uncertainty principle might allow. But that's really not going to do it for most people.

It really doesn't leave much room for genuine human science, because humans become the kind of odd man out—the observers of nature who can't account for their own participation in nature. And one of the reasons that Kant is an attractive figure to me is that he deals with that problem head-on and in a particularly meticulous and careful way. And another thing that makes him very appealing to me is that he had a… and here, my reading of Kant is perhaps a little different from some other people. I've always been very interested in his early work, and his very late work, and what links them together. And it seems to me that most people start reading Kant with the *Critique of Pure Reason*. Well, you know, he was almost 60 [Laughs] by then he was a pretty smart guy before he wrote that. So, what was he doing all those years? And some of you know, some of us get our best thoughts when we're younger. So what were the best youthful thoughts that propelled Kant's career and made him so single-mindedly devoted to the kind of inquiry that he spent his life at? And those thoughts seem to be very much related to the questions of human embodiment, the kind of mysterious duplicity of human existence, and here's where the sublime also comes into play. So one of the reasons that Kant is attractive as a modern thinker, is that he aims at least for a comprehensiveness that many modern thinkers have either deliberately denied themselves or just find themselves unable to address, embracing both the authority of modern science on the one hand, yet also opening up the possibility of human freedom and making room for a rigorous investigation of

human existence that is not exhausted by a modern scientific, reductive materialist or quasi-materialist explanation of things.

PF: That is a beautiful answer, and fascinating.

SMS: Yes.

PF: I want to hear you talk more about Kant's comprehensive understanding of human existence. If it's possible.

SMS: So, I mean, if I think what ties all of this together for me is a series of questions that he keeps posing for himself and returning to. And one of the things that I find attractive in Kant as a thinker and also just as a human being, is, I could call it a youthfulness of mind, that many old people, you know, you learn things, and then you remain attached to them, and you defend them, and you become rigid. And he was always, in a way, returning to the same sets of questions and with a certain dissatisfaction in his former answers. And therefore, even at the very, very end he was driven—the *Opus posthumum* is a series of exercises in which he's trying, yet again, to perfect a system in which he sees gaps and in which he's… but but I think one way of describing what all of those attempts involve is a concern with the problem of human embodiment. So that would be, that would be one way, and another way in would be more moral and political. And this is another attractive feature from a modern liberal point of view, namely, that at least after reading Rousseau, all of his metaphysical formulations are reframed in terms of a project which is as he calls it, establishing the rights of humanity and the ends of reason are henceforth understood to be essentially moral and political, and again, from the point of view...and this is something that I took up, especially in my in my undergraduate course this semester, an extraordinarily robust and rich defense of liberal democracy, broadly understood that to some degree, leapfrogs beyond the 19th century and industrialization and colonialism and the various things that now seem to be, you know, weigh liberalism down. Even a figure like John Stuart Mill, who was a kind of a colonial apologist, Kant, in a way, is ahead of I mean, he's behind them. [Laughs]. He was writing at a time when, there really was virtually no industry in Prussia; they still had serfs, and we still had slavery in the United States, but in some ways, I find his thought extraordinarily helpful and fresh in ways many call postmodern. I'm not sure we really are postmodern yet, but we're certainly postindustrial and *pace* Trump. [Laughs]. And, you know, again, I think Kant has a lot of fresh insights to offer on questions that

some of the 19th and early 20th century, even late 20th century figures are less…Yeah, seems outdated in the way that Kant isn't, at least if read the way I tend to read him, which is very sympathetically, or as with the expectation that he, that there's something there that is not only true, but also practically helpful.

PF: So, you mentioned that Kant's understanding of human nature, if we can use the word human nature or human existence, leads to a very specific discovery of what politics should be and why politics should be the final aim or end for human beings. Could you talk more about this?

SMS: Well, I mentioned that politics for him is a kind of transitional phase between what he calls the state of nature, I mean, these are all kind of formulas and the 'kingdom of ends', a moral kingdom of ends, which is…it's easy to understand as a secularized version of the Christian kingdom of grace, but politics is has this intriguing in-betweenness for him, which is his juxtaposition in *Perpetual Peace* of the moral politician and the political moralist; a political moralist is somebody who uses morality for political purposes, which is to say that he's immoral. So there's no such thing, you know, strictly speaking, as a political moralist. A moral politician, on the other hand, is something that's at least possible, but extremely difficult. So let me, let me describe it this way, that one of the interesting comparisons you could draw would be Aristotle on natural right and the limitations of law. And we'll leave Aquinas out of it for now, Kant on law and how to finesse those, those hard cases where general laws don't seem to quite do the job. And then someone like Schmitt, who says, hell with the law, let's just have, you know, dictators declare states of exception, you know, do their will, and we'll hope God's on their side.

So, I think Kant, again, offers attractive alternative, to say that the kind of Aristotelian argument for the prudence of the statesman who just somehow knows, because of this kind of σοφὴ φρόνησις, this practical wisdom, knowing what to do on the spot, and can somehow adjust the laws accordingly, or even ignore it on occasion which is, in a way, a very attractive ideal, but very difficult to bring about or to meet with any regularity, so that politics becomes, for him, an almost inevitably failed project to achieve some kind of justice that… and Kant is a little bit more hopeful than Aristotle, I think, about the possibilities, particularly of democratic politics and so again, that's another attraction for me, is that he is more hopeful while at the same time having,

an Augustinian understanding of the tragic dimension, the potentially catastrophic degradation to which human beings can bring themselves. And whereas, perhaps not the tragedians, but the Greek philosophers tend to be ultimately to be a somewhat cheerier bunch, if only you know, for the few who can be philosophers, they are also somewhat neglectful, perhaps, of the fate of the many who can't.

PF: Yes, another interesting difference between Aristotle, or ancient philosophy, and Kantian philosophy is that, well, Aristotle and probably Plato care a lot about happiness and probably regard happiness as the end of our human life. But Kant would probably say, well, maybe following the moral law and pursuing a kind of transformation from the state of nature to the state of law should be our proper end. Why would Kant reject this kind of classical notion of happiness as the end and have a new proper end for human beings, which is the moral end?

SMS: Okay, well, his won't be a very satisfactory answer, but one reason is that in a way he buys into the Hobbesian story about happiness, that that at least as natural beings, in the crude sense, we're driven by certain laws of desire, as he puts it, such that happiness can only mean when you're talking about at that level, and that's the crudely natural level. Happiness is simply the maximization of pleasure and the minimization of pain. And from that point of view, human life is pretty futile, because, again, he buys it to a kind of Lockean argument, I think at the same time which, which to some degree, Rousseau shares, which is that we're, if you, if you think of life… just a kind of an accounting book of pleasure and pain, you're always going to end up in the red. I mean, a very simple way, because for every pleasure, there's a pain that comes first, but there's going to be a pain at the end that's going to have no pleasure afterwards. And therefore, you're– you know that even the very best life is going to have more pain than pleasure. So if you're just living for maximum pleasure or minimum pain, the best thing to do is kill yourself at first as quickly as possible. So, it can't possibly be. He says that, if concepts were all there were to life, and you had a choice, nobody would choose to be born. So, there must be more to it. Then I think he inherits, partly from Rousseau, an idea that there are two kinds of enjoyment. There's, and this is to use Rousseau's language, there's *plaisir*, there's pleasure, and there's joy, *jouissance*.

And *jouissance* is an active sentiment of your own living force, your own existence, the sentiment of your own existence, which has a kind of self-sufficient gratifying *raison d'etre* that is independent of this lower calculus. And for Kant, the equivalent of that, that activity, that

active pleasure – Kant doesn't call it active pleasure. He calls it moral satisfaction, which is again the sense of one's own existence. And the other thing that Kant changes from Rousseau: for Rousseau, reason is ultimately not the highest expression of this life force, this life activity. It's more like imagination, the experience of the poet, the experience of the solitary walker, you know, floating in his boat. And for Kant, the highest expression of reason is a moral activity which transcends this aesthetic self-activity, but that has certain, at least formal, similarities with what I think Rousseau means by *jouissance*, by this kind of active sentiment of one's own existence.

Kant, on the other hand, is not, I mean, he doesn't want to be called a hedonist, and so he identifies that with a kind of noble transcendence of happiness, of ordinary happiness, which is this kind of higher feeling, because ultimately, it is a moral feeling of one's own autonomy. That's something that goes beyond what, I think, what Rousseau would admit. Another way of putting it is that Rousseau ultimately is not a moralist, in my view, and Kant is.

PF: That's a noble answer…

SMS: Yah.

PF: and a true one, I guess.

SMS: Yah.

PF: So, I'm very attracted by the idea that to live a moral life now, how should I do it? And you just mentioned that, well, Kant has a kind of praise for liberal democracy, yeah, because it can help us to satisfy our political purpose, which is to live a moral life. Could you say a few more words about why Kant thinks that a liberal democracy is good for our lives?

SMS: Well, that's a really…that's a really great question too. So, I read a lot of Kant's later writings as at least partly a response to Rousseau's *First Discourse*. It raises this very interesting question, you know? I mean, well, Rousseau didn't raise it. Somebody else raised, but Rousseau tries to answer it. You know, it has progress in the arts and sciences made us better? Made us morally better? Has it made us happier? And Rousseau's answer, crudely, is mostly no. Mostly…but there may be a few individuals like me. That is to say, if you can live this very high life, if you're a genius, if nature broke the mold with you, too. But otherwise, no, it's… it's been a, it's been a bad deal, and…and rather than take that misologic path all the way

that maybe some followers of Rousseau would, though, I think not Rousseau. Kant wants to find a way to redeem civilization so that history can be read as a comedy, after all, and not a divine comedy, but, you know, a kind of human comedy, and but it's one in which we have to ourselves produce the result. And since it's in a state of freedom, always shadowed by the possibility of failure, and indeed catastrophic failure. And so the comedy is sort of a tragedy, like all great comedies, at least Shakespearean comedies, the line between comedy and tragedy is very thin.

So why? Why? Well, I mean, these are attractive ideas -- the dignity of…the dignity of man. I mean, I, frankly, I sometimes find the contemptuousness in some classical works for ordinary people. You know, I…it just doesn't resonate with me. Maybe because I'm an ordinary person, but I resist self-contempt, but, but there's something attractive about Kant's desire to see in ordinary people, you know, all the warts, but at the same time, moral possibilities that that mere intellectual prowess, can't compensate for or can't eclipse and at the same…so I find these efforts in Kant's later works like the *Critique of Judgmen*t…to figure out how you move a civilization that, in his time and ours can look very decadent, and one could almost despair of the human condition. If you look around the world and you know what AI might do and what you know what social media has already done, you know, where are we going with this? Has it made us better? Has it made us happier, or is it simply degrading us, this development of our rational faculties, especially through technology, let alone the possibility of, you know, thermonuclear war, climate, extinction, whatever people are worried about that Kant sees politics as the sole remedy? It's a difficult remedy. We're made of crooked timber, as he famously puts it. So, there's no perfect political solution, but something like liberal democracy is the best we can do.

And then that becomes itself a cause to which a moral person can devote him or herself. And you think of figures like, you know, fighters, I think the 106-year-old woman who was in the French Resistance who recently died. I mean, these are people who, you know, devoted themselves to a very high ideal, an idea of France that that really did stand for the rights of man, broadly understood, and someone like Abraham Lincoln, similarly, I think, can set a model for the kind of commitment to a larger purpose which is not specifically religious. It's not linked to a particular…truth, or metaphysics. And you know it is, I think, attractively familiar to those of us who've grown up in a liberal democracy and willy nilly, take, take the Declaration of Independence

to be a very noble document and…and so on. I don't mean to exclude you from growing up in the States…

PF: That's fine.

SMS: …and pledging allegiance to the flag. Did you still have to pledge? Did you pledge allegiance to the flag?

ERJ: Yes.

SMS: Really?

ERJ: In elementary school? Yes.

SMS: And then, you stopped?

ERJ: Yeah, probably around maybe fifth grade, even though it was the same school.

SMS: And do you know why they stopped?

ERJ: No, just all of a sudden. Yeah, there's no explanation.

SMS: Well, I mean, I grew up in the middle of the cold, the height of the Cold War, and we did all kinds of stuff. We had bugle calls. We had to stand at attention if, in high school, if you were late and you were caught out in the parking lot or something, you had to stand rigidly at attention while this recording of, you know, of Reveille played. So, I mean, there were all kinds of excesses, shall we say. But there's a certain kind of refined patriotism with respect to liberal democracy that I think enables one to live a moral life, not simply in terms of one's ordinary day-to-day transactions, but feeling, you know, that there is a larger communal project to which you can attach your energies.

PF: There's one specific event that Kant has commented on, and has experienced, and is very controversial, which is the French Revolution.

SMS: Yes. I thought you were going to ask about Kant's racism…I am puzzled…maybe you are going to get to that

PF: I was not expecting that.

SMS: Ok, alright, alright.

PF: If you want.

SMS: No, no, no, I mean, if you say controversial, I didn't know the French Revolution was controversial, but…

PF: Well, it's controversial to some people, such as Burke.

SMS: Oh, well, what a crank. [Laughs]. No. Burke had his own problems, but he's a little unfair. And he was, you know, he had this nice, liberal British tradition too, that he could fall back on. What the Germans did with Burke was somewhat different from what Burke did with Burke, and I'm not sure the Irish would have fared so well under a kind of German version of Burkeism, but that's neither here nor there.. There's a very interesting Kant scholar at the University of Oslo, Reidar Maliks, who's coming to the workshop, who wrote a really interesting book about Kant and the French Revolution. And I think there's a lot of…there are a lot of good things in that little book, but one of them is just tracing out the complexities of Kant's comments on the French Revolution and what he likes, what he refers to with praise, particularly in *the Critique of Judgment*, which was published in 1790 is a very early stage of what comes to be called The French Revolution. And it was a stage in which arguably no laws were broken, yet, and arguably they were on their way to establishing a constitutional monarchy of a sort that…that's kind of a quasi-British sort.

And so that was, that was a French transformation that he could wholeheartedly endorse. What it became later he had, you know, he was much more critical of, as his comments on the execution of Louis XVI illustrate, but he still thought that the sheer fact of the revolution, with all its warts, demonstrated. This is what he argues in a late work, *The Conflict of the Faculties,* that all things considered, there's a germ, there's an element, there's a slight preponderance in these mysterious human innards that we can't directly access. There's a slight tilt in the direction of morality over immorality. And he makes a very interesting argument that, again, rests on a kind of sublimity, as he describes it, an experience of the sublime, not directly by any of the actors, not the soldiers themselves, but people like Kant looking on and seeing it as a kind of historical sign. There's what he calls a moment of exaltation, which is not quite enthusiasm, which is itself not a terrible thing. But when you're enthusiastic, your reason is no longer in charge. You're overwhelmed with this, these high emotions. And this is kind of a good thing. But again, it's never a good thing for reason to lose control, as with enthusiasm, as distinguished from fanaticism, which is in no way good. So, he uses a different word for enthusiasm, enthusiasm, and *Schwärmerei,* which is, you can translate as fanaticism.

But even then, when Kant's talking about this, what he says is the key moment, they've got a sign that history is progress, progress is happening, and has already happened. It's this, this sense of exaltation in the spectators who look at this without, without breaking the law, which would indicate a lapse of reason, and particularly moral reason. So that would be a, I guess, example of… I don't know…if that's an answer to your question about the French Revolution, is we had this very complicated response, that was both dark and light.

ERJ: I want to add something only because, well, one, you mentioned racism, and two, because we had an author two years ago who wrote on Kant and racism, and I can't remember the thesis of the paper, but I'm just wondering, do you have any thoughts on racism in Kant's work in particularly relating to liberalism…to his advocation for liberalism, or it seems, at least in my, not in-depth reading of Kant, that it seems to be in contrast with the idea that everyone is endowed with or at least there are some self-legislating subjects.

SMS: Right. So, there's moral universalism on the one hand. And then he says in his anthropology lectures and private notes, and you know, there's a tension on Kant's thought. And the tension is, on the one hand, his universalism, which I think is his deeper philosophic commitment, because it's rational and it's a priori and it's necessary. And then there is this attempt to look at history in a way that would support our hopes in the realization of this, this ideal. And there, he gives himself permission, under the rubric of what he calls 'reflective judgment', to make educated guesses about a kind of teleology within history. And again, none of this has the certitude of either empirical fact or certainly a priori knowledge, but he presents it as a permissible hope that you could look for signs that nature is on our side, and is sort of going to help us, if we help ourselves.

And there, I think he runs into trouble sometimes because in his, in his desire to find meaningful patterns he will seize on things that he may think are innocuous, but that lack objective, empirical evidence, and yet support certain kinds of historical hopes, and one is a trajectory of history that involves a kind of racial hierarchy, with white Europeans at the top and American Indians at the bottom, and then blacks, and then Asians. So, he finds that an appealing and attractive pattern, because it suggests a kind of symmetry. And I mean, I think he gives himself too much permission in those directions. He does it also with women and male-female differences. These are things which cannot really be by his own better lights, proven empirically, but it…it's this, this window that opens up when he talks about 'reflective judgment' and that gives

him permission to say things about Judaism and Christianity. So, every time he…this is the downside of his, of his imaginative efforts to cull from the historical evidence, signs of, of natural support for human moral progress. So, I think it's a difficulty in his thinking.

But on the other hand, what? What thinker doesn't have difficulties? One of the attractions of Kant is that it's sort of like a carpet. I mean, that every comprehensive philosophy has, has, there's always wrinkles. And so you, you try to smooth the wrinkles here, and they pop up someplace else and, and, you know, there are wrinkles for Aristotle, in the way, it's a little, you know, if so, if there's a natural teleology that acorns become oak trees, what about all the acorns that don't become oak trees? And if our natural teleology is to become a philosopher, but you know, millions and millions and millions of us are going to be like acorns that just rot in the ground. Maybe that's not a fully satisfying understanding of the human condition either. So that's a pretty crude way to summarize Aristotle. [Laughs].

PF: That's fine.

SMS: But, yeah, so, I…in a way that the difficulties are as revealing of the human condition as the, you know, insights. I mean, in a way, that the difficulties are them… are themselves insights. And Kant is only too happy to admit where things break… in his system break down, and one is the intelligibility of freedom. He admits that ultimately, it's not intelligible, but we can't let it go.

PF: We're talking about all these difficulties about Kant in a specific institution, which is a university

SMS: Yes. [Laughs].

PF: And Kant also, I believe, has some interesting comments on the role and mission of the university in the progressive force in human development.

SMS: Yeah. Now this is a very late development in his thinking, after he suffered through the reign of the successor to Frederick the Great, who, from Kant's point of view, was a very benign ruler. I mean, not that he did all in every way, good things, but at least he was very… his touch was very light when it came to censorship, and so Kant was able to do his thing quite, quite readily. The only thing you couldn't do under Frederick the Great was question his authority, and as long as you didn't question autocracy as such, you could pretty much publish

anything you wanted. Particularly, he was critical of religion. You know, he was delighted when you criticized religion.

So, Kant had a pretty free hand. And I think he assumed that rulers like Frederick the Great, who were great Machiavellians, basically, as I think Kant probably believed, that you could work with that. You could work out a *modus vivendi* with such rulers that you wouldn't rock their boat, and then they would let you publish *What is Enlightenment*, and *the Critique of Pure Reason,* and so on. Gradually you would bring about a general public enlightenment that would make possible, maybe a slightly more liberal form of government. The death of Frederick and the rise of this much inferior and much more censorious Friedrich Wilhelm II meant that Kant could no longer publish freely on religion, and he couldn't really publish freely about anything touching on serious moral and political issues until the death of that king.

So Kant developed workarounds. And the big workaround he finally comes down to is the workaround that he has sort of lived through and with his whole life, which is as a university professor, student, and then professor. And so, he re-envisions the university very late in his work. 1798, pretty much the large, last big thing he writes, *The Conflict of the Faculties*, which means the university, faculties and he co-ops the whole mechanism of the university as then understood, which basically was a state apparatus for the production of doctors, lawyers and ministers, who were themselves servants of the state and carrying out the State's mission in various ways. But there was this other faculty, the lower faculty, the philosophy faculty, which we would today call the College of Arts and Sciences, as opposed to the law school, the medical school, and the School of Theology. And he saw in that little corner of things, where he at least facetiously says the purpose of the universe is to produce Ph.D.s. That's kind of a joke, but not entirely. But really, it's the pursuit of truth and the critique of reason, the self-critique of reason as a part, a necessary part and foundation of that, of that pursuit that can come to terms with, and, again, work out a *modus vivendi* with the state, precisely because the state, too, needs the facts. They're going to have to fight wars. They need guns that actually work, and they're going to, you know, deal with plague. They need medical information that actually works. So the idea here is that Kant can, he can use the university as a kind of inner transformative institution that placates the state by giving it goods that it can appreciate. And at the same time, it brings about a kind of inner revolution in thinking by training all the people who are going to go out there and be state ministers. And so, there'll be a kind of subtle, radical transformation of society through the university

under the guise, which isn't entirely a guise or merely a guise of supporting ends that this that rulers themselves can -- whose value rulers can appreciate.

And it seems to me that the modern university, as it develops after Kant in Germany and then the United States, very much carries through on that plan. We always have to pay the piper, somebody, taxpayers, the church, big rich donors, somebody has to be convinced that it's good for them to support this thing. And so, the idea that you can be a pure, you know, academy of knowledge? Well, yeah, even Socrates had to flatter his rich friends. [Laughs]. So, I think Kant was very clear-headed about the political necessities that in his time, which weren't altogether different from the political necessities in our day. And I think we've lost sight of some of those political necessities and are high-minded, oh, 'we're all for academic freedom of speech.' Well, okay, and what about who's going to pay, who's going to pay the salaries, and who's going to buy the labs? And I think the forgetfulness about that has cost the university something.

At the same time, Kant has a very interesting way of describing the University. He says it's a commonwealth of scholars whose members are teachers and students. And the trustees of the University are not the big-shot donors, like in our day, the trustees are the professors, which I find an appealing thought. [Laughs]. I mean, he's a real believer in faculty governance. And at the same time, he says that every university faculty has a right-wing and a left-wing. And he's here borrowing the language of the French Parliament, National Assembly, the right and the left, as we still call it, and the right wing are the professional schools. Because their whole job is to shore up institutions that are the status quo. Teach the law as it is. To teach the latest, good, best practice as we now know it, that's the job of the lower faculty, the philosophy department, the biology department, and so on and so forth. Their job is actually to find out what's true, and that means that they're going to, you know, they're going to be boat rockers, because what's true, what's known, new discoveries are, this is not necessarily going to coincide with what is actually being done on the ground.

And therefore the College of Arts and Sciences, the philosophy faculty, is left, is inherently left, not in the sense of being a hotbed of Marxism, although that's one way you can understand it [Laughs], but that there's a kind of progressive radicalism inherent in the pursuit of knowledge, in the sense that a proper College of Arts and Sciences understands itself, and so it isn't when right wing MAGA folks think university is a hotbed of Left wing radicalism. Kant will say, well, you

know, they kind of have a point [Laughs], but not for the reason they think. And I think that's just a very, very interesting formulation and insight on his part about the inherent structure of a modern university.

And there's a there's a third role that I think he foresees that then becomes very prominent, which is the role of the university in shaping national cultures, certainly in Germany, which didn't have a unified state until much later in the 19th century, with Fichte already the university is where a common civic German consciousness is supposed to emerge and I think at the American the land-grant colleges in the US, and then later, places like Johns Hopkins, which were more, even more emphatically modeled on the German system, the idea that these are places where American culture is going to be somehow nurtured, and where you bring together people from, you know, this, this coast and that coast, and the Midwest and different demographics. And the result is some kind of broad, high national culture, that's one of the functions of a university, and so we're not just a Jesuit university, we're not just a Catholic University, we're also an American university. And that means that, too, is part of what a place like BC is supposed to be doing from a Kantian point of view. So, I find all of those ideas very fresh.

PF: I think so, especially under our current administration.

SMS: Yeah, yeah. But again, it kind of reminds you of why we're having some of the problems we're having because of a certain forgetfulness about paying the piper. And here's another thing that came to mind in my undergraduate class today. Why? Why the sudden indifference or even glee with which MAGA is receiving the destruction of biological research and medical research, I mean, all these lifesaving things? I mean, America had this, you know, vast engine of biomedical research and progress, and, you know, just a hatchet has been taken to it. Why do ordinary people out there in the hinterland not find that alarming? And I think Kant might say, Well, one reason is because they don't get the benefit. As far as I can tell, medicine is expensive, even apart from COVID. You know, maybe if we had the national health care system that delivered more affordable health care to ordinary people, they wouldn't be so suspicious or so eager to see the hatchet, but why should they pay? Why should they pay taxes for advanced cancer research that they're not going to be able to utilize or access?

So that, and Kant would say, well, because he has a very interesting approach to welfare as well. Part of the responsibility of the state is to make sure that there's a healthy population, and not for the sake,

not because people have a right to welfare, not out of compassion, but because, if you don't have a healthy population, you don't have a state, you know, Europe…You need healthy people, and families, and new citizens, or it's just an idea. It's not a reality.

So, yeah, you can just use Kant in all kinds of ways, I think, to figure out correctives that don't parse as either blue or red, Democrat, Republican, but, but maybe point in a helpful direction, just in terms of very current problems we face, we can then we go on to, you know, international relations that would be a whole other, a whole other interview.

PF: So if we are not going into international politics and stay on the topic of the University. Yeah, you have been studying in the university for a long time, and then you have been a professor in the university for a long time.

SMS: I've been in school a long time, beginning with, you know, nursery school, yeah. [Laughs].

PF: And you have been a student of Kant, a friend of Kant for a long time.

SMS: Yeah, unrequited friend. [Laughs].

PF: Yes. Is there some advice or inspiration that we can get from Kant about what we should do as university students, and how we can best utilize the resources in university as students, based on your reading of Kant or your experience as a professor?

SMS: Well, I am not sure if this has to do with Kant specifically, but I was, in a way, privileged, in a way not so privileged, to be a student at the university at a time when there was a…it was, it was the mid to late 60s—to date myself. And in the end, it all blew up. I mean, I had no classes my senior year, basically my first year in graduate school. I mean, it was a mess. It was chaos and…but the first two years of college was a kind of exposure to an extremely broad range of approaches which were at the height of their self-confidence. So, I had, I had courses in the philosophy department that were just analytic, you know. You just wouldn't believe how narrowly analytic. And then I had historicist courses in the German department, and I had New-criticism, and also the beginning of Critical theory in literature departments. I had these Straussian professors in the government department, then I had these Positivist professors, also in the government department. So, I would say that ideally you should expose yourself to as broad

a range of approaches as you can manage, while also negotiating the demands of a major and, you know, getting the professors that your friends tell you are good, and all the other things you do when you choose courses. But I would think I'd say that one thing is to open one's mind, to as broad a range of approaches as possible.

And I think Kant had the benefit of that himself as a young student, because he was caught between two, these two conflicting approaches. One was the Leibnizian rationalist approach, and the other was embodied in a very, very interesting thinker named Crusius, who was a Lutheran, who strongly was a fidelist. And so, either from the very early stage, Kant confronted this conflict between freedom -- moral freedom -- as simply a fact of human life, that if there are moral obligations, then we must be free, because otherwise it would be meaningless. On the other hand, the Leibnitzian view, which tended to diminish the meaning of freedom in that view's rationalist notion that there's a sufficient reason for everything.

So, from the very beginning, he was trying to negotiate these two outlooks and find a kind of satisfying synthesis. And there seemed to be something to recommend both of them. If you went totally with Crusius, you had to give up on reason. If you went totally with Leibniz, you seemed to have to give up the moral life. And I'm not saying you can always find a perfect medium. But Kant is a radical moderatist. You know, he wants to have it all, but he wants to get the extremes in their full-blown form, and then try to work out a satisfactory accommodation. And so, to the extent that I think undergraduate education presents itself in a somewhat similar way, that's probably something that Kant would also approve.

PF: So, philosophize in intentions.

SMS: Yeah, you know, my husband makes fun of me. He says everything for you, in the end, reduces to: it's a problem. [Laughs]. But I think Strauss says the same thing. He says, right? There's only the eternal problems, eternal questions. And Nietzsche says almost the same. So maybe that's the best we can do. Yeah, that's not so bad.

PF: It's important to understand what we don't know. That's philosophy.

SMS: Yeah, yeah. You don't necessarily have to come up with it, with the right answer, at least not right away.

ERJ: Well, Professor Shell, we thank you so much for your time for this interview today, and we wish you all the best in your philosophical endeavors.

SMS: [Laughs]. Thank you, you too.

A RISK-SENSITIVE APPROACH TO POPULATION ETHICS

WESLEY STONE

§ INTRODUCTION

How many people should there ever be? Derek Parfit poses "this awesome question" in *Reasons and Persons* (1984).[1] It seems strange to try to put a precise number on it, yet the question has long held political salience. Traditionally right-wing "natalists" emphasize the value of large families and warn of the societal dangers of a falling population. On the left, a countervailing "degrowth" movement has arisen, motivated by the fear of environmental catastrophe to encourage a slowing, stopping or unwinding of technological advancement and human settlement. Some even see civilization as a fundamental evil and advocate for voluntary human extinction.

Philosophers (at least a certain subset of them) study this question in the field of population ethics. Population ethicists analyze populations, or worlds, along two axes: size, how many people there are, and welfare, how good their lives are. They are specifically interested in comparing the overall quality of different worlds. Obviously this is a highly simplified model of real populations, which we might think of as more or less valuable depending on how virtuous and knowledgeable their citizens are, or how just their institutions are. But population ethicists assume worlds to be equal in all other respects, because while size is quite straightforward, welfare presents enough problems on its own.

[1] Parfit 1984, pg. 381.

Philosophers have come to no consensus on what exactly makes a person's life go well versus poorly, or whether this mysterious concept of "well-being"[2] can even precisely be defined. The first problem is easy to resolve. Population ethicists stay agnostic on the welfare debate, and make arguments which are compatible with any reasonable definition. The only requirement here is that interpersonal welfare comparisons are possible, and all this requires is that my life or yours is better than that of someone in a Siberian labor camp. The second problem is trickier, and I will have more to say about it later. For now it will be sufficient to observe that even with only an imprecise notion of well-being, it is clear some worlds are better than others. For instance, heaven (a huge world of extremely well-off people) is clearly better than hell (a huge world of extremely badly-off people). In this paper, I will use numbers to represent well-being, on the rough scale of 100 being a very good life and -100 being a very bad life, with 0 the dividing line between worthwhile and not-worthwhile lives.

The goal of population ethics is to formulate a "social welfare function" (SWF) which takes a world, and somehow combines each person's individual well-being into an aggregate value (V_W), which I will express in the unit of "points" as a "score" for that world, to allow it to be compared to other worlds (the higher the score, the more desirable a world). The simplest way to do so that takes everyone's interests into account is to attain global utility from a basic sum of each individual's well-being, so that everyone contributes exactly their personal well-being to the overall. This is the SWF known as "totalism," which I will defend.

But totalism is thought to suffer from a devastating objection: the Repugnant Conclusion, formulated by Parfit as he pondered the Awesome Question.

> *Repugnant Conclusion:* Compared with the existence of many people who would all have some very high quality of life, there is some much larger number of people whose existence would be better, even though these people would all have lives that were barely worth living.[3]

For example, imagine a world similar to heaven, but finite in size. Call this World A. Now, imagine a population a hundred times bigger, each with lives just three percent as good – though, critically, still good

[2] I will use the terms welfare, well-being, quality of life and utility interchangeably.

[3] Parfit 2016, pg. 110. This is a revised formulation, clearer than the original.

and worth living. Call this World Z. Since Z has three times as much well-being in it than A, by totalism it is three times as good. Yet this seems – well, repugnant.

The challenge for population ethicists is that avoiding the Repugnant Conclusion is not so easy as just rejecting totalism. In fact there are compelling arguments that repugnance is inescapable. In this paper, I will give one of these arguments and examine possible objections, with the hope of demonstrating why repugnance is so difficult to avoid. I will conclude that the best response to the Repugnant Conclusion is to accept it. Then, I will then contrast two conflicting intuitions which lie at the heart of population ethics, arguing in favor of the repugnant one. In the next section, I will introduce the problem of uncertainty which has not received the attention it merits in the field, and motivates my thesis. Finally, I will build up a modified totalist SWF that uses uncertainty to prevent repugnance from arising in most realistic scenarios, with the goal of softening the intuitive blow of accepting the Repugnant Conclusion.

§ ARRIVING AT THE REPUGNANT CONCLUSION

In this section I will offer a brief overview of the field of population ethics. First, I will provide a simple argument for the Repugnant Conclusion, and describe potential objections. Next, I will develop one of those objections into a popular non-repugnant SWF, and demonstrate its shortcomings. This discussion is intended to illustrate why progress in the field has been so difficult. Finally, I will briefly describe the status quo of population ethics, so that my argument may be placed in its proper context.

Here are three statements that, taken together, imply the Repugnant Conclusion:[4]

1. *The Benign Addition Principle:* If worlds *w* and *x* are so related that *w* would be the result of increasing the well-being of everyone in *x* by some amount and adding some new people with worthwhile lives, then *w* is better than *x* with respect to utility.

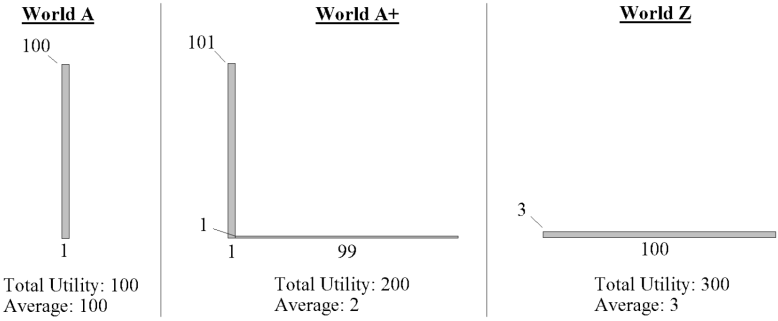
2. *Non-anti-egalitarianism:* If *w* and *x* have the same population, but *w* has a higher average utility, a higher total utility, and a

[4] Adapted from Huemer 2008, pgs. 2-4.

more equal distribution of utility than *x*, then *w* is better than *x* with respect to utility.

3. *Transitivity:* If *w* is better than *x* with respect to utility and *x* is better than *y* with respect to utility, then *w* is better than *y* with respect to utility.

This is Michael Huemer's "Benign Addition Proof", and though "proof" might be a bit strong, it's a quick, intuitive argument that well illustrates the challenge of avoiding repugnance. Here is a representation:[5]



In this figure, and all future ones, height represents well-being and width represents size. For consistency's sake, let's stipulate that each unit along the x-axis represents one billion people, so the size in A+ is 100 billion people, and the total utility is 200 billion. Now, we can attain A+ from A via Benign Addition – increasing the welfare of the billion people in A by 1 and adding 99 billion lives at welfare 1. From A+ to Z, we can see that total and average utility have increased, and complete equality has been achieved. From here a simple application of transitivity suffices to attain the result that Z is better than A, and repugnance is demonstrated, as any single Z-world being better than any single A-world is sufficient.

Though these steps are intuitive, all three can be objected to. You may have noticed that the move to A+ introduced glaring inequality. This can be mitigated by supposing that the two groups live on separate continents or even planets and are unaware of the other's existence, but it is true that inequality is a global feature of A+. If you see no benefit to creating new barely worthwhile lives, this may constitute a compelling objection. As such, in the next section I will

[5] Huemer 2008, pg. 4.

offer a defense of Benign Addition. In our second move from A+ to Z, we lose a certain high quality of life. A "perfectionist" objector might claim that A contains high-quality goods which cannot be exchanged for any amount of low-quality goods in A. The intuition here is that the Z-lives are "drab," free from pain and worry, but also almost entirely free from meaning and deep relationships. Parfit famously described these as full of "muzak and potatoes."[6] However, we may stipulate that the Z-lives are instead "roller coaster" lives, with the same peaks as in A, combined with a near-equal amount of deep suffering.

There's another way to object to Non-anti-egalitarianism, though. "Critical-level" theories claim that lives below a certain threshold should not count towards the score of the world. Usually the way this works is that each person's contribution to the aggregate is determined by subtracting that threshold from their well-being.[7] If you set the critical level high enough that it seems reasonable to prefer a large number of them to a smaller number of A-lives, then you've avoided repugnance.

However, critical-level theories have a problem of their own. The consequence of setting a positive threshold is that some worthwhile lives (Z-lives) decrease the score. This leads to what Gustaf Arrhenius termed the "Sadistic Conclusion":[8]

> *The Sadistic Conclusion*: When adding people without affecting the original people's welfare, it can be better to add people with negative welfare rather than positive welfare.

For example, let's assume that we have some base population and we can either add 1 billion lives at 1 or 10 million lives at -100. With a critical level of, say, 10, the first addition decreases the value of the base population by 9 billion while the second decreases it by 1.1 billion. So we should strongly prefer adding many extremely negative lives over worthwhile lives, which is even worse than repugnance. Out of the frying pan, into the fire.

Treating Z-lives as positive leads to repugnance, while treating them as negative leads to sadism. Critical-range theories try to sail through this Scylla and Charybdis by treating them neutrally. Specifically, all lives in the critical range between the critical level and 0 contribute nothing to the score. But surely a billion people at 10 is better than a billion people at 0? Population ethics is an objector's

6 Parfit 2016, pg. 118.
7 Totalism is a critical-level theory with a threshold of 0.
8 Arrhenius 2000, pg. 251.

paradise; it's very easy to formulate counterexamples to any proposed SWF, and impossible to formulate a SWF with no bullets to bite. Still, many people regard rejecting one of the premises of the Benign Addition argument as less painful than accepting its conclusion. That includes even transitivity, which was Parfit's preferred solution.[9] But this is not the only argument for repugnance.

The main sticking point for population ethicists has been the many "impossibility" arguments which derive an even more repugnant scenario (the "Very Repugnant Conclusion") from even more intuitive premises.[10] They are highly technical, so I won't try to reproduce one, but they are generally thought to succeed in producing a set of incompatible but seemingly necessary conditions for a satisfactory SWF, including anti-repugnance of some kind. Among these, anti-repugnance stands out. John Broome highlights the fact that the anti-repugnant intuition is very complex, compared to the simplicity of other conditions.[11] Mark Budolfson and Dean Spears argue convincingly that repugnance is a weird feature of aggregating over unbounded spaces, and should bear little practical force.[12] Then the easiest way to respond to these arguments, other than taking the nihilist way out and giving up, is to simply bite the bullet on repugnance. The goal of this paper is to soften that intuitive blow.

§ SIMPLICITY AND NEUTRALITY

In this section I will defend the Benign Addition Principle – the idea that adding worthwhile lives to a world increases its score. First, I will explain the "Simple View" (which entails Benign Addition) and the "Neutral View", two plausible but mutually exclusive intuitions. Then I will argue that we can embrace simplicity without many of the controversial views associated with it, and in so doing set up the main argument of this paper.

At the heart of repugnance are two conflicting intuitions. On the one hand, we have what Parfit terms the "Simple View":

[9] Parfit 2016, pgs. 114-115.
[10] See Arrhenius 2000, Arrhenius 2009, Budolfson and Spears (unpublished).
[11] Broome 2004, pgs. 57-59.
[12] Budolfson and Spears (unpublished), pg. 33.

> *The Simple View*: Anyone's existence is in itself good, and makes the world in one way better, if this person's life is good to live, or worth living.[13]

This seems quite plausible, but it's a one-way ticket to repugnance. If worthwhile lives always make the world better, then each life we add to Z increases its score, even if just by a tiny amount, and eventually it must eclipse A. Yet there is also what I will call the "Neutral View" proposed by Jan Narveson:

> *The Neutral View*: we are in favor of making people happy, but neutral about making happy people.[14]

This is also plausible, yet clearly inconsistent with the Simple View. If something makes the world better, then we should be in favor of it, all else being equal. The conflict seems to arise because each view considers the problem from a slightly different, well, viewpoint. When we think about each person's life from their own perspective, of course it seems that worthwhile lives are inherently valuable. But when we take more of a bird's-eye view, unless you're a natalist it seems strange to have a particular desire to enlarge the teeming mass of humanity.

One way of resolving the tension is to argue that the Simple View is too narrow. Sure, worthwhile lives matter, *if* they exist. This is simply the first part of the Neutral View. But why should we consider the existence of mere possible people? *They don't exist.* In fact, it doesn't even make sense to talk about their interests; those interests also don't exist![15] It's easy to get tangled in a metaphysical morass when discussing possible people, which is why I won't try to give a conclusive proof of the Simple View. But because the rest of my argument will have little force for those who reject it out of hand, I do feel compelled to offer a few points in favor of it.

First, the Neutral View turns out to be quite ambiguous when prodded. Simplicity can easily be formalized as totalism, but on the other hand, it's not clear how to represent neutrality as a SWF. We can interpret "in favor of making people happy" as endorsing something akin to Non-anti-egalitarianism, and "neutral about making happy people" as rejecting the Benign Addition Principle. But are we really neutral about making *any* happy people, or just barely happy

[13] Parfit 2016, pg 110.
[14] Narveson 1973, pg 80.
[15] Thanks to Prof. Samuel Asarnow for his helpful explication of these ideas.

people? If we are neutral about creating Z-lives but not neutral about creating A-lives, it looks like we've come back to critical-level totalism, with all its attendant problems. On the other hand, if we're hardcore neutralists, then we have some weird SWF that assigns all new positive lives a score of 0, and presumably all new negative lives some negative score. This leads to the unfortunate conclusion that having children always makes the world a worse place on expectation, because there's at least some chance of the child having a bad life, which I don't think is what most neutralists have in mind. There might be a different way of interpreting this view that threads the needle, but the lack of immediate clarity is not promising.

A related point is that the Neutral View doesn't hold up consistently. In small populations, it is much less plausible. A world of 50 happy people is surely better than a world of 10, better still 100. And all of these are preferable to 0 happy or unhappy lives for anyone who's not a radical degrowther. On the other hand, totalism can be upheld across the board, though we have seen that there are scenarios when it is tough to do so. Furthermore, the point where we stop caring about new happy lives seems to me precisely the point where we become completely unaware of their existence, and then of course we would rather see the lives around us improve rather than have more happy people created whose existence we will never notice. An intuition that bends itself to our self-interest in this way is not a reliable one from which to derive moral values.

Lastly, there are plenty of times we do in fact consider the interests of possible people. Consider the striving immigrant, who makes sacrifices to come to a wealthy Western country so that if and when she has kids, they will be better off. Or consider when governments craft forward-looking policy for the "future generations." If a nation's shrewd conservation policies help mitigate natural disasters 200 years from now, when the population, absent a dramatic technological breakthrough, will consist of entirely new people, it seems like that would be a good thing, even if no one alive now is affected. A neutralist might describe these as cases of making people happy rather than making happy people, as we are considering worlds where future people are better or worse off. The problem for this view, which Parfit terms the "Nonidentity Problem,"[16] is that the people in alternative futures are not guaranteed to be the same. When we consider how contingent our individual existences are (presumably a different sperm fertilizing one's mother's egg would have produced

[16] Parfit 1984, pgs. 351–380.

a different person, and very small changes in the past could've brought this about), it becomes clear that we are in fact considering cases of creating one group of people versus creating another (happier) group, with little or no overlap between the two. Preferring the latter is a rejection of neutrality.

An objection to the Simple View is that if possible people are allowed to enter into the equation, their sheer (potential) numbers quickly swamp the interests of those who currently exist. This is a counterargument often raised against "longtermist" philosophers, who support efforts to mitigate "existential risks" such as supervolcanoes or unaligned AI that threaten to wipe humanity off the map: in for a penny, in for a pound. If lowering x-risk is so important, why shouldn't we immediately divert all funding from education, healthcare and nutrition programs toward developing a global asteroid-defense system to avoid the fate of the dinosaurs? Without wading too deeply into that hot-button debate, I want to defend the Simple View's ability to avoid such absurd conclusions.

The Simple View is neutral in its own way – it does not care whether you make people happy or make happy people, just about how much well-being you create. And critically, in the real world it's often more efficient and less risky to make people happy than to make happy people. I can donate money to feed starving children right now, knowing that my generosity will have a substantial immediate effect. Conversely, if I, even as a well-off citizen of a prosperous society, were to have a child now, there would be a lot of dirty diapers in the near term and no guarantee that a happy life would eventually result from it. Even if the world where I have a child is likely better, it might not be the *best* I can bring into existence, which is why assigning high priority to natalism misses the mark. Practically speaking, I support a view that combines the attractive elements of both simplicity and neutrality, something like "we are in favor of making people happy, and in favor of making happy people once we've made everyone who currently exists happy."

A similar argument can shield simplicity from the dangerous implications of unrestricted longtermism. Even if there might be 10 quadrillion people sprawled across the universe in the year 5296, it is *extremely* uncertain what impact our actions now will have on them. Perhaps the best thing we can do now is improve the lot of just the next generation, which we can do with a high degree of confidence, so that they in turn will be prepared to look further into the future. The point is that the fact of one world being better than

another, while relevant, is not the only factor for making important decisions. In practice, these decisions often turn on uncertainty about which outcome will actually attain, rather than the best theoretically.

§ UNCERTAINTY

Now we have built up the foundations of population ethics, considered various proposed SWFs, and introduced the central problem of repugnance. We have offered a proof of that result, and dug into the intuitions behind it. In this section I will turn to my response to all of this. I will discuss the challenge posed by uncertainty, and then provide a thought experiment which should serve to illustrate my approach to dealing with it.

First, I will return to an important point I elided earlier, the challenge of precisely measuring well-being. If we are unable to do this, the entire exercise of population ethics is threatened. Though we have a general understanding of what constitutes a good versus a bad life, we are far from clear on a precise definition, or whether one is even possible. There are two explanations: either well-being is so complicated that we just don't know enough about it to formulate a definition, or well-being is inherently imprecise, and a life of quality 65 or -12 is a meaningless concept. Population ethics could survive if the latter were the case, as some worlds are clearly better than others and it would be nice to have an explanation why, but it would be much more convenient if the math we do was actually meaningful on some level. Still, even if a numerical scale of well-being is ultimately an idealization, idealizations can be very useful for high-level theories.

But idealization or not, if population ethics is to help us in the real world it must take into account this great uncertainty. And though uncertainty has received some treatment, in this area the literature is deficient. A notable result was Arrhenius' extension of his impossibility arguments over probabilistic outcomes. From similar premises, he managed to derive the even-more-than-very-repugnant "Risky Very Sadistic Conclusion," which presents a choice between A, or a lottery with a >99.99% chance of just the sufferers in Z- and a <0.01% chance of just the mediocre lives.[17] You could still eventually add enough mediocre lives that the expected sum of well-being by choosing Z- is greater than that of A. And Budolfson and Spears made a similar extension of their own impossibility result.[18] But the absurd

[17] Arrhenius and Stefánsson 2023, pg. 8.
[18] Budolfson and Spears unpublished, pgs. 15-17.

scenario of the Risky Very Sadistic Conclusion is even less realistic than the original repugnant and very repugnant results. I want to look at the practical implications of population ethics, and accordingly, I've tried to formulate an example which could plausibly resemble a decision humans will at some point have to make.

> *Space Colonization:* Sometime in the not-too-distant future, humanity builds a spaceship capable of transporting humans across the galaxy. In preparation for this monumental achievement, we sent rovers to seek out new planets to settle, and they've found and precisely analyzed two candidates. Planet A is quite similar to Earth but with much less land area. One billion people could eventually live there in great contentment, at welfare level 100. Conversely, Planet Z has much more land area, but is located at the frigid edge of the habitable range. One hundred billion people could eke out barely worthwhile lives there, at welfare level 1. The lives in A are 100 times better than those in Z, so the total expected well-being of both worlds is the same, but there is near-universal consensus that A should be chosen. Is this wrong?[19]

No. What kind of a lunatic would advocate for Z? As it turns out, in spite of the last 13 pages seemingly arguing for this, not me. Though I am a committed totalist, I see a subtle difference between *Space Colonization* and the Repugnant Conclusion. The key word here is "expected," because in this case and in all that could conceivably arise in the real world, we are basing our decisions off of projections and estimates rather than the certain numbers of thought experiments. And I see an important distinction here. Consider that the rover's estimate of well-being might be off by a small amount. Now we're much less sure *how* positive A's score will be, but it still will surely be positive. Z, on the other hand, could easily be quite negative if those barely worthwhile lives turned out actually to not be worth living. In the remainder of this paper, I will propose a method for incorporating uncertainty into totalism.

§ INCORPORATING UNCERTAINTY

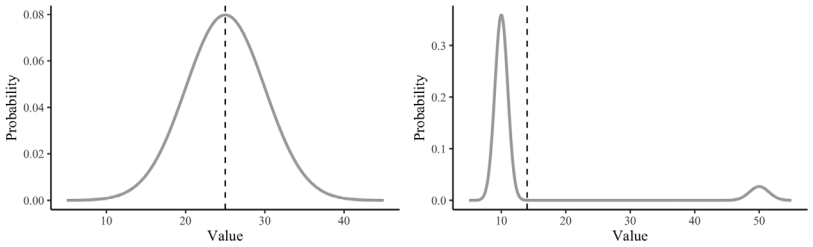
The remainder of this paper will focus on introducing and explaining my "risk-sensitive" approach to population ethics. Initially,

[19] Note: earlier, I analyzed a Z-world with lives of quality 3. I lowered that to 1 for this example to simplify the math. Both worlds are commonly regarded as repugnant.

I will summarize the math behind it and explain why I make some potentially controversial modeling choices. With this shiny new SWF in hand, I will apply it to *Space Colonization* and another counterexample to basic totalism,[20] but unfortunately conclude that my first pass does not sufficiently penalize repugnance. To address this, I will modify this first pass to create a new SWF that produces more (to my mind) satisfactory results.

However, while I feel very optimistic about my general risk-sensitive approach, there are many mathematical tools one could use to take uncertainty into account, and I do not claim that the SWF I present is the best possible.

The first piece of the puzzle is probability distributions (PDs). A PD allows us to assign varying degrees of likelihood to a range of possible outcomes, and represents them by mapping those outcomes to values on the x-axis and their likelihoods to heights on the y-axis. Since the total area under the curve is always 1, computing the area of different parts of a PD will give us the probability of an outcome in that range occurring. A common example is normal distributions, which can be used to model a surprisingly broad range of different things, from pinky lengths to standardized test scores to measured brightness of stars, and come in a distinctive bell-shaped curve.

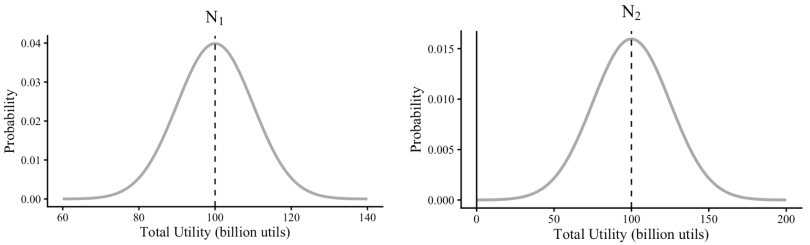


We will ask our rovers to represent uncertainty by reporting back a probability distribution (PD) of total utility rather than a single value. However, what we get may not be so neat as the normal distribution pictured above. But thankfully, for our purposes any PD will do. PDs contain a lot of information, but we will be focused on just two aspects. First is center, or more technically, the mean (represented by the dashed line). In the left figure, the mean is conveniently located at the peak right in the middle, but the geographic and numerical centers don't always coincide. This is because the mean is attained by

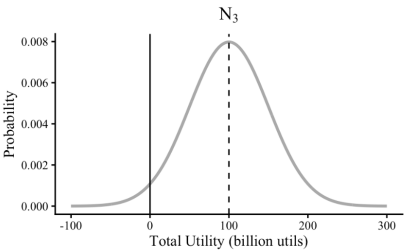
[20] Basic totalism can only handle worlds with no uncertainty, but there is a relatively straightforward way to extend it to probabilistic scenarios, which I will introduce later.

summing the product of each value along the x-axis with its probability along the y-axis. So in a bimodal distribution like the right figure, the mean may not be very close to the center of the picture. But it always represents the center of the data.

The other important aspect is the "spread" – essentially, the range of possible values. To illustrate why this matters, here are two more PDs.



Both are normal distributions with the same mean, but if you compare the values on each axis closely, you can see that the rover is less certain about its estimates of total well-being in N_2. The actual value of this planet is more likely to be worse off than predicted, but it's also more likely to be better off. Think about whether either seems preferable to you, then consider a third:



Again, the mean is the same, but this time the rover is even less confident – to the point that the world could actually be negative (just as likely, it could be extremely positive).

There are three possible routes one could take here. If N_1 seems better than N_2, which seems better than N_3, then you are *strongly risk-sensitive*. The fact that uncertainty exists at all makes a PD less preferable. If N_1 and N_2 seem similar, but both better than N_3, then you are *weakly risk-sensitive*. As long as the result is a positive outcome, uncertainty isn't a problem, but the possibility of creating a bad world is worth an extra effort to avoid. Finally, if the potential risks of all three PDs seem equally balanced with the potential rewards, then you

are *risk-indifferent*.[21] In the remainder of this paper, I will advocate for weak risk-aversion.

The theoretical case for risk-indifference is strong. If you see N_1, N_2, and N_3 as totally equivalent, then all you care about is the mean. And recall that the mean is essentially a weighted sum of the distribution. So risk-indifferent totalism is really the extension of basic totalism to probabilistic worlds, which is attractively simple and consistent if you accept basic totalism in certain worlds. And its proponents might push back against being termed indifferent to risk. The badness of the bad possibilities in N_3 does detract from the mean. It's just that there are proportional opportunities for much better worlds that balances it out.

The challenge for the risk-indifferent view is that most people care more about risk than this. Let's say I offer you a bet on a coin flip. If you call it correctly, you win $101, and if you don't, you lose $100. Do you take me up? Probably not, unless you really love to gamble. Yet if we modeled this bet as a PD, it would have a positive mean, and the risk-indifferent partisan would insist that it's irrational not to take the bet. You hesitate because, as Lara Buchak says, when we evaluate gambles we are "being sensitive to 'global' properties of gambles: [of] being sensitive not just to what happens in each particular outcome but to what the gamble looks like as a whole."[22] Buchak argues in the context of decision theory that it is in fact rational to take risk into account over and above its impact on the expected value, and I make a similar argument in the context of population ethics. The fact is that decisions about whether or not to take bets, or which planet should be colonized, do not take place in a vacuum.

To see why, let's raise the stakes of my coin-flip bet.

> *Well-To-Do Businessman:* A well-off businessman who's achieved a comfortable lifestyle is offered a bet on the outcome of a coin toss. If he guesses correctly, he will win $10 million, but if he is wrong, he will owe that much.

The businessman would recoil at the prospect. Sure, winning $10 million would be great. He could buy a bigger home, take fabulous trips, retire early. But being on the hook for $10 million would be catastrophic. He could lose everything, and still be shackled with a debt he could never pay off if he worked every day for the rest of his

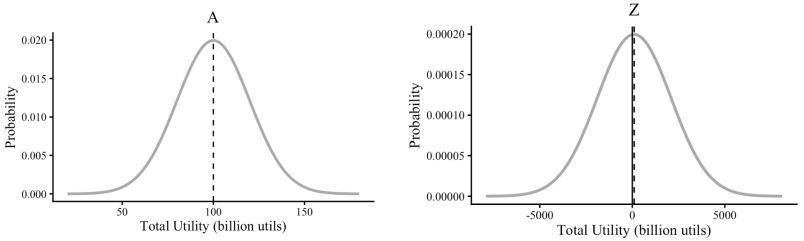
[21] Or perhaps, more bluntly, a *degenerate gambler*.
[22] Buchak 2017, 1.

life. Even though the monetary value appears the same, in reality the risk is much, much worse than the reward.

In *Space Colonization*, we are in the position of the businessman. He can walk away and get on with his comfortable life, and we can travel to A and be assured of a comfortable new colony. When presented with an alternative that could be much better but could also be much worse, he recoils, and rightfully so. Even if he were enticed with a slight gain in the expected value of the bet (+$10,001,000 vs -$9,999,000), he would be crazy to take it.

Since Z is necessarily much riskier than A, this scenario is analogous and can help explain why we resist repugnance. Recall that in the beginning we defined worlds along two axes: quantity, and quality. The critical thing to note is that while quantity is always positive (or 0), quality can be negative. Uncertainty about quality can produce uncertainty about the absolute goodness or badness of a world, while uncertainty about quantity merely affects the magnitude. And the score of Z is much more sensitive to fluctuations in quality than A, since its citizens live so much closer to the edge. All it would take is a slight error in the rover's extremely complex calculations to produce a net-negative world in Z, whereas A is all but guaranteed to be positive.

Let's take the scenario from *Space Colonization.* Suppose that the rover presents the choice between an A world with 1 billion people at welfare 100, and Z with 100 billion people at welfare 1. Additionally, the rover provides a confidence interval of ± 20 in its estimates of well-being, meaning that it is 95% confident that the people in A will live lives between 80 and 120, and equally sure that Z-lives will be between -19 and 21, and models uncertainty as normally distributed.[23] Then the PDs will look like this:



[23] It would also likely provide a confidence interval on its estimates of quantity. But since each additional person adds just 1 or 100 well-being, and each additional point of well-being can change the total value by billions or trillions, for simplicity's sake this can be ignored.

The rover is 99.9% confident that A's score will be at least 70 billion points, but only 54% sure that Z's will be greater than 0. The goal of our SWF is to properly penalize Z's possible negative outcomes.

To this end, *risk-sensitive totalism* generates the total value of a PD by taking the expected value of the PD, and multiplying it by an uncertainty modifier.

$$V_W = uE$$

This uncertainty modifier is derived by chopping off the negative range and computing the area under the distribution that remains. Specifically, we set u equal to the probability that any value X in the PD multiplied by E will be positive.

$$u = P(EX > 0)$$

Applying this formula to *Space Colonization*, every value in A's PD is positive, so $u = 1$ and the score remains at 100 billion points. But for Z, X is negative 46% of the time, so u is just 0.54, and the score is 54 billion. Conversely, in the original repugnant scenario there is no uncertainty and the scores for both remain the simple sums of well-being. Risk-sensitive totalism agrees with regular totalism in theory but deviates in practice.

An important aspect of this formulation is that u is always between 0 and 1. If u were greater than 1, then we would be rewarding rather than punishing uncertainty. And if u were negative, we would end up giving a positive on expectation scenario a negative score. On its face, this might not seem crazy. Perhaps we would prefer not to colonize any planet than to colonize one where we expect the people there to live lives of quality 0.01 on average, and possibly much worse. However, this also means that we could assign a negative world a positive score, which means we would accept the Sadistic Conclusion, which is a line I refuse to cross.

How should we, then, handle cases where E is negative? We must preserve the sign, but we can either continue to penalize uncertainty or reward it. Consider a "reverse repugnant" scenario where we are forced to choose between A- or a corresponding Z- world. This is somewhat analogous to T. M. Scanlon's well-known World Cup example, where several billion people are asked to give up fifteen minutes of watching the game to rescue someone in the transmitter

room of a TV station from excruciating electric shocks.[24] Most people agree that we should prefer the world where a bunch of people suffer small harms to the one where one person suffers a tremendous harm.[25] And rewarding uncertainty in negative cases is actually what our SWF currently does. While the PD of A- is always negative, Z- sometimes ends up being positive. Those positive outcomes decrease the value of u because we're now multiplying them with a negative E, which makes V_W less negative for more uncertain scenarios. This example illustrates how my SWF is risk-*sensitive* rather than risk-*averse*.

> Here's another case that illustrates the perils of risk-aversion:
>
> *Homeless man:* A homeless man living on the streets is offered a bet on the outcome of a coin toss. If he guesses correctly, he will win $10 million, but if he is wrong, he will owe that much.

Unlike the well-to-do businessman, the homeless man would jump at the chance to take this bet. He has nothing to lose, and everything to gain. Though the expected monetary value is the same as before, the agents' different treatments of risk makes these gambles completely different. And this same behavior can be observed across a wide variety of contexts. Consider a presidential election in which one candidate holds a clear lead over the other. The favorite will play it safe, avoiding confrontation and dreading any "October surprise" that could shake things up. Conversely, the underdog will throw anything and everything at the wall, hoping for just such a last-minute shakeup in the polls. Or consider a basketball game where one team has a double-digit lead in the final minutes. They will run down the clock with safe passes looking for easy layups while their opponents will aggressively hunt for steals and 3-pointers, trying to give themselves as many chances as possible to snatch victory from the jaws of defeat. We often give risk special consideration in our decision-making, but not always in the same way. Risk aversion can be just as much a vice as risk addiction.

We now have a working SWF, so let's evaluate its performance. We have already seen risk-sensitive totalism at work mitigating repugnance, and it has the additional virtue of mitigating a sort-of "utility monster" PD which could be contrived as a counterexample to basic totalism. Imagine that our rovers discover a planet M and report

[24] Scanlon 1998, pg. 235.

[25] Though I argue on the grounds of uncertainty rather than anti-aggregation – I don't share the intuition that *any number* of people in Z- is better than A-, no matter how uncertain.

a strange bimodal (two-peaked) PD giving a 99% chance of creating a world with mean V_W = -1,000, and a 1% chance of creating a world with mean V_W = 1,000,000,000. This PD has an expected value of 100,000 points. Alternatively, we could be guaranteed to create a world with 1000 people at welfare 99, producing 99,000 total points. Basic totalism tells us to leap at that 1% chance, but it seems quite counterintuitive to accept a 99% chance of creating a bad world. On the other hand, risk-sensitive totalism lets us pump the brakes. In the first scenario, the probability of creating a positive world is just 1%, so V_W = 0.01 * 100,000 = 1,000, much less than 99,000.

While this is quite a harsh penalty, you may have noticed a problem with its scoring of *Space Colonization*. With the numbers the way I originally laid them out, we prefer A to Z. But it is not difficult to adjust those numbers to produce the opposite result. All we need to do is double Z's size to 200 billion, and then its score will eclipse A's at 108 billion. And this still seems pretty repugnant. Fortunately, risk-sensitive totalism lends itself well to customization All we need to do is generalize the formula for deriving *u*:[26]

$$u = \frac{P(EX > 0)}{sP(EX \leq 0) + P(EX > 0)}$$

Here *s* represents a scalable risk sensitivity parameter (for any positive value). When *s* is 1, *u* is the same as before, because the denominator is always 1. As *s* increases, however, more weight is given to the negative possibilities. For instance, if you thought they should be given twice as much weight as the positives, then u_Z = 0.54/(2(0.46) + 0.54) = 0.37. Now Z would need to triple in size to catch up with A. You could push *s* higher and higher to confine repugnance to the most fanciful corners, but, as with everything in population ethics, this does come with a cost. The more sensitive to risk you become, the less sensitive you are to actual well-being (on expectation), and you'll become vulnerable to counterintuitive preferences for Z-like worlds with minimal uncertainty over worlds with much higher expected value and uncertainty that regular totalism strongly prefers.

However, in setting the value of *s* we are not confined to constants. As long as *s* is always positive, we can evade sadism. And

[26] Conservatively, neutral possibilities are categorized as negative rather than positive, but the effect on the math is negligible.

there is a compelling case to be made that *s* should scale with the mean. Currently, we penalize an outcome of -1 points equivalently to one of -100,000, simply by computing the probability of its occurrence. If *s* were to increase (and decrease) in magnitude with the mean, -100,000 points' stronger downward pull (assuming *E* is positive) will cause a corresponding rise in *s*. The challenge is how exactly to do this. If we just set $s = E$, since we're ultimately multiplying *u* by *E*, the mean terms will cancel. This would have the opposite effect of divorcing the score almost entirely from the expected value of the distribution, and this strange new SWF would not only be neutral about making happy people but also making people happy.[27]

The (somewhat arbitrary) function I landed on instead is the base-10 logarithm of the mean ($\log(E)$), which represents *E* as 10^x, and returns *x*. So if *E* is 10, it returns 2, if *E* is 10,000, it returns 4, and if *E* is 987654321, it returns 8.99. However, it's not as simple as just substituting *s* with $\log(E)$. Unfortunately, the base-10 log can return negative numbers for $E > 1$, because, for example, $0.1 = 10^{-1}$. To account for this, we must instead use the "log1p" variant, which takes the absolute value to account for negative values and then increases *E*'s magnitude by 1 so that when $E = 0$, $s = 0$, which is the desired behavior. My final proposed risk-sensitive totalism looks like this:

$$V_W = uE, \text{ where}$$

$$u = \frac{P(EX > 0)}{log(|E| + 1) * P(EX \leq 0) + P(EX > 0)}$$

Now we can return to *Space Colonization*. A retains its score of 100 billion points, because $P(EX \leq 0)$ is still 0%, so $u = 1$. For Z on the other hand, *u* works out to about 0.096. This produces a score of 9.6 billion points. That's substantially lower than 54 billion, and we can't just do the same trick we did last time and multiply Z's size by 11. Recall that the purpose of using a log function is to scale the penalty with the mean. Increasing the number of positive lives increases the expected value of the PD, and thus increases the modifier. With Z of population 1.1 trillion, we obtain *u* of roughly 0.089, and a score of 97.7 billion. It takes another thirty billion people for V_Z to eclipse V_A.

Is this a satisfactory result? I think so, with two caveats. First, if you just refuse to accept repugnance at all, no form of risk-sensitive

[27] As it solely represents the probability that a given world would have a score the same sign as the mean.

totalism is for you. In that case, I wish you good luck coming up with your own SWF that evades the impossibility arguments. Second, it's a little odd to draw a precise yet fairly arbitrary line where repugnance becomes acceptable (somewhere between 1.12 and 1.13 trillion people in Z). But I think this can be explained by admitting that even our rover's best estimates of their uncertainty are imprecise. When uncertainty disappears, so do arbitrary lines, as we're back to basic totalism. A practically useful SWF is not designed to be evaluated with such precision, it merely needs to be good enough in all cases.

With that said, let's evaluate whether the SWF is giving scores in the ballpark of reasonability. It expresses a preference for colonizing Z with 1.13 trillion people at (expected) welfare of 1, over A with 1 billion people at welfare 100. Strictly speaking, this is a repugnant result. But I think we would struggle to find a planet which could accommodate a trillion people, even at subsistence level. A better counterexample involves a much smaller A, likely a moon (though I imagine we would also struggle to find a moon which accommodates such a high quality of life). The SWF prefers 9.5 billion Z-lives to 10 million A-lives.[28] At this point, as a committed totalist I would argue that the potential value of Z now outweighs the risk. Usually, when someone offers you a choice between two bets, one with an expected monetary value 11.3 or 9.5 times that of the other, you should take that one (though of course there are businessman examples in which you shouldn't).

However, even if you're amenable to accepting repugnance in theoretical cases, you might desire to banish it entirely from real-world decisions. This could be accomplished by substituting another function for s, maybe by using the natural log rather than base 10. Now Z would not eclipse A until it reached a size of 3 trillion in the original *Space Colonization*, and 24 billion in the moon example in the previous paragraph. And this only scratches the surface of ways to customize my SWF. Perhaps, for instance, you favor the "strongly risk-sensitive" position I staked out earlier and want to penalize uncertainty itself, not just possible negative outcomes. To do so would require some mathematical wizardry, so I won't attempt such a modification here, but there's no reason it couldn't be done, and I won't even go so far as to say it shouldn't be done. My risk-sensitive totalism is only intended as a starting point, and I don't claim it should be the ending point, though its relative simplicity is an attractive feature.

[28] As E decreases, the denominator shrinks, which increases u, so there's less of a penalty.

The claim I do make is that uncertainty should be taken into account in any practical application of population ethics, in some form or another. This allows a strong but clean distinction between theoretical and practical population ethics, providing an easy response to the impossibility arguments, while potentially satisfactorily limiting the practical consequences of accepting repugnance. Furthermore, this view can go a long way, I believe, towards explaining why we resist the Repugnant Conclusion so strongly. When we read Parfit, we can't help but attempt to imagine the scenarios he presents in the real world, and in doing so unwittingly introduce significant uncertainty into the equation. Were we not worried about their proximity to negative territory, we would be better able to appreciate the value of the vast number of worthwhile lives offered by world Z.

REFERENCES

Arrhenius, Gustaf. "An Impossibility Theorem for Welfarist Axiologies." *Economics and Philosophy* 16 (2000): 247–66.

Arrhenius, Gustaf. "The Impossibility of a Satisfactory Population Ethics." In *Descriptive and Normative Approaches to Human Behavior*, 1–26. 2011.

Arrhenius, Gustaf, and H. Orri Stefánsson. "Population Ethics Under Risk." *Social Choice and Welfare* (2023).

Broome, John. *Weighing Lives*. Oxford: Oxford University Press, 2004.

Buchak, Lara. *Risk and Rationality*. Oxford: Oxford University Press, 2017.

Budolfson, Mark, and Dean Spears. "Why the Repugnant Conclusion is Inescapable." Unpublished manuscript.

Huemer, Michael. "In Defense of Repugnance." *Mind* 117, no. 468 (2008): 899–933.

Narveson, Jan. "Moral Problems of Population." *The Monist* 57, no. 1 (1973): 62–86.

Parfit, Derek. "Can We Avoid the Repugnant Conclusion?" *Theoria* 82, no. 2 (2016): 110–27.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Scanlon, T. M. "The Structure of Contractualism." In *What We Owe to Each Other*, 189–247. Cambridge, MA: Harvard University Press, 1998.

REINTERPRETING THE HIGHEST FORMULA OF AFFIRMATION

NIETZSCHE'S TWO ETERNAL RECURRENCES IN THUS SPOKE ZARATHUSTRA AND THE SELF-OVERCOMING OF NIHILISM

JUNZE CHEN

§ INTRODUCTION: THE APORIAS OF ETERNAL RECURRENCE

The doctrine of eternal recurrence has always been one of the greatest aporias for interpreters of Nietzsche's philosophy. Throughout his lifetime, Nietzsche develops different approaches to the problem of eternal recurrence. In the famous section 341 of *The Joyful Science*, Nietzsche presents eternal recurrence as an imaginary vision of the demon interrogating whether we want to relive our current lives "once more and countless times more" with "nothing new in [them]."[1] Yet in his posthumously collected fragment *The Will to Power*, Nietzsche seems to then conceive eternal recurrence not as a vision or imagination, but rather the "most scientific of all possible hypotheses."[2]

This disunity in Nietzsche's presentation of the doctrine led some interpreters (Oger 1997; Löwith 1997) to argue that there are inherently two incompatible notions of eternal recurrence — one ethical and one cosmological — in Nietzsche's works. Different

[1] Friedrich Nietzsche, *The Joyful Science / Idylls from Messina / Unpublished Fragments from the Period of The Joyful Science (Spring 1881–Summer 1882)*, Trans. Adrian Del Caro,
(Stanford University Press, 2023), 204

[2] Friedrich Nietzsche, *The Will to Power*, Trans. Walter Kaufmann and Reginald John Hollingdale, (Vintage Books, 1968), 36.

from these interpreters, who attribute the split between these two notions of eternal recurrence to Nietzsche's change in ideas, I argue that the split is necessary to the doctrine itself. The two notions of eternal recurrence already coexist in Nietzsche's presentation of the doctrine in Part III of *Thus Spoke Zarathustra*. The duality in the idea of eternal recurrence does not simply point toward an inconsistency in Nietzsche's works. As I try to demonstrate, Nietzsche presents two notions of eternal recurrence because there are two ways of solving the riddle of eternity and moment. These two ways are already evident in the riddle of eternal recurrence itself.

My interpretation will begin with an analysis of the riddle of eternal recurrence in "The Vision and the Riddle." In my interpretation, "The Vision and the Riddle" reveals an inner contradiction between moment and eternity within the framework of linear temporality. Interpreting the tension between moment and eternity as the central problem that Nietzsche wants to resolve, I propose that the riddle of eternal recurrence inherently opens to two possible solutions: one that focuses on the moment (the ethical eternal recurrence) and one that focuses on eternity (the cosmological eternal recurrence). These two solutions to the riddle, further presented by Nietzsche in "The Convalescent," become two opposing ways of formulating the doctrine of eternal recurrence. The tension between the two formulas of eternal recurrence is crucial for Nietzsche's project of overcoming nihilism. As I shall demonstrate, the core of Nietzsche's philosophy lies in correctly balancing these two formulas of eternal recurrence. The overcoming of nihilism requires not only the cosmological equation of eternal recurrence, but also that one actively wills and affirms the cosmological equation in an ethical gesture. In this sense, Nietzsche's philosophy becomes a philosophy in becoming — here, becoming is not only a metaphysical concept; it rather emphasizes how Nietzsche put his own philosophy into motion. This eventually leads to a more nuanced understanding of Nietzsche's project of overcoming nihilism.

§ THE COSMOLOGICAL AND THE ANTHROPOLOGICAL-ETHICAL EQUATION OF ETERNAL RECURRENCE

It is not a new idea to argue that Nietzsche's writings contain two opposing, even contradictory concepts of eternal recurrence. In Karl Löwith's 1935 book *Nietzsche's Philosophy of the Eternal Recurrence of the Same*, he observed that Nietzsche's teaching of

eternal recurrence contains two opposing equations. As he concisely summarizes, the eternal recurrence concerns "on the one hand, with an 'ethical gravity' by means of which human existence that has become goalless obtains a goal again, beyond itself; and on the other hand, with a natural-scientific 'fact' in the goalless self-contained existence of the world of forces."[3] Löwith names the former equation of the eternal recurrence (the one that emphasizes the "ethical gravity" of willing the same life) as "anthropological" and the latter equation (emphasizing eternal recurrence as a physical fact) as the "cosmological."[4]

For Löwith, these two equations of eternal recurrence are intrinsically incompatible. In the cosmological equation, the world is conceived as a constant and "goalless cycle" of becoming, an infinite play of forces with "no beginning and no end.[5]" But because the world is constituted by a "definite quantity of forces," the combination of forces must be exhausted in the infinite span of time, turning the world into a "circular movement that has already repeated itself infinitely."[6] Eternal recurrence, in this metaphysical-scientific interpretation, becomes a factual statement about the actual "temporal structure of the physical world."[7] The anthropological equation, on the other hand, does not concern the actuality of eternal recurrence in the present; rather, it treats eternal recurrence as a possibility in the future. As Löwith argues, the task of the anthropological equation is to transpose the heteronomous imperative "thou shalt" into the self-affirming "I will" that affirms the recurrence of one's current life.[8] This transposition creates a "will to rebirth" that seeks to "eternalize" and elevate our ephemeral, momentary life to something that is also to come in the future.[9] The cosmological equation of eternal recurrence eventually declares the world to be always-already a goalless, self-revolving circular movement. But the anthropological equation presents the reestablishment of a goal on a higher, existential horizon. In Löwith's words, it posits a "willed goal of a will to eternal-ization" that "liberates from the burden of the past and arises from the will to the future."[10] Whereas the cosmological equation absorbs past, future, and present into a homogeneous recurrence of the same, the

[3] Karl Löwith, *Nietzsche's Philosophy of the Eternal Recurrence of the Same*, 1st ed, Trans. J. Harvey Lomax, (University of California Press, 1997), 83.
[4] Löwith, 84, 88.
[5] Löwith, 89
[6] Nietzsche, *Will to Power*, 549.
[7] Löwith, 94.
[8] Löwith, 87.
[9] Löwith, 86.
[10] Löwith, 87.

anthropological equation opens itself to the future through the active willing of repetition. For Löwith, these two equations of eternal recurrence eventually represent Nietzsche's two different approaches to temporality.

A similar distinction between the two equations of eternal recurrence is also being drawn in Eric Oger's 1997 article "The Eternal Return as Crucial Test." Following Löwith's classification, Oger also proposes that there are two contradictory interpretations of Nietzsche's eternal recurrence: one interpreting eternal recurrence as a cosmological-ontological doctrine, and the other interpreting eternal recurrence as an ethical-deontological doctrine. But unlike Löwith's diagnosis of the duality of eternal recurrence as "a fundamental contradiction of Nietzsche›s philosophy," Oger perceives each of the two possible interpretations of eternal recurrence to be limited in itself.[11] Different from both the cosmological and the ethical interpretation of eternal recurrence, Oger instead proposes a third interpretation of eternal recurrence, that is, conceiving eternal recurrence as a "crucial test" of one's life.[12] To form this interpretation, Oger retreats to Nietzsche's initial presentation of eternal recurrence in *The Joyful Science*. For Oger, the demon, whispering to ask whether we want to relive our current lives eternally, opens up a "prospective" perspective that points "toward the future."[13] This ultimately forces the thinker to "test" the value of their current life.[14]

Despite Löwith and Oger's convincing analysis of the split between the two doctrines of eternal recurrence, it still remains unclear how this theory of two eternal recurrences could fit into Nietzsche's conception of his works. Both interpreters seem to suggest that Nietzsche himself never had a unified concept of eternal recurrence throughout his writings.[15] However, if it is true that the thought of eternal recurrence is itself disunified, how should we now interpret the role of *Thus Spoke Zarathustra* in Nietzsche's philosophy, given that, in *Ecce Homo*, Nietzsche accounts this book to be based on "the

[11] Eric Oger, "The Eternal Return as Crucial Test." *Journal of Nietzsche Studies*, no. 14 (1997), 1.
[12] Oger, 7.
[13] Oger, 11.
[14] Oger, 11.
[15] Oger points out the fact that supporters the ethical interpretation often "goes back to completely different text" than supporters of the cosmological interpretation (4). For Löwith, this division in the idea of eternal recurrence even reflects Nietzsche's conflicting identity between a "natural scientist" and a "founder of a religion" (83).

thought of eternal recurrence"?[16] To me, there are only two ways of resolving this tension between Nietzsche's account of *Zarathustra* and the theory of two eternal recurrences. Firstly, it could be argued that *Zarathustra* only represents one doctrine of the eternal recurrence, and Nietzsche is simply unaware of the division. Secondly, it could be argued that the concept of eternal recurrence is inherently divided, and this division is already evident in *Zarathustra*. My essay tries to take the latter approach — as I will argue, the duality of eternal recurrence, rather than being a result of an arbitrary change in Nietzsche's idea, in fact deeply relates to the theme of overcoming nihilism in *Zarathustra*. This requires us to turn to Nietzsche's first formulation of the doctrine in Part III of the book.

§ THE RIDDLE OF ETERNAL RECURRENCE AS THE PARADOX OF MOMENT AND ETERNITY

The chapter "On the Vision and the Riddle" is Nietzsche's first formal discussion of eternal recurrence in *Zarathustra*. As Heidegger insightfully observes, the doctrine of eternal recurrence exists initially as a "riddle" — a riddle does not involve proving the argument step by step, but instead demands that "we take a leap" into the "truth of being" that is unconcealed.[17] Just as Zarathustra himself describes, the riddle is exclusively open to the bold researchers [*Versucher*] who love to "guess" but "hate to deduce."[18] The riddle itself is designed as a philosophical experiment [*Versuch*] rather than a question with a definite answer.

In my interpretation, the riddle of eternal recurrence is composed of two parts that successively present the paradox of moment and eternity. The first part of the riddle concerns the relation between "moment" and "linear time." As Zarathustra portrays in his speech:

> "Behold this gateway, dwarf," I continued. "It has two faces. Two paths meet here; no one has yet followed either to its end. This long lane stretches back for an eternity. And the long lane out there, that is another eternity. They contradict each other, these paths; they offend each other

[16] Friedrich Nietzsche, *The Case of Wagner, Twilight of the Idols, The Antichrist, Ecce Homo, Dionysus Dithyrambs, Nietzsche Contra Wagner*, Trans. by Adrian Del Caro, et al. (Stanford university press, 2021), 278.

[17] Martin Heidegger, *Nietzsche: Volumes Two*. Trans. David Farrell Krell, (Harper, 1993), g37.

[18] Friedrich Nietzsche, *Thus Spoke Zarathustra: A Book for None and All*. Trans. Walter Kaufmann, (Penguin Books, 1978), 156.

> face to face; and it is here at this gateway that they come together. The name of the gateway is inscribed above: 'Moment.' But whoever would follow one of them, on and on, farther and farther-do you believe, dwarf, that these paths contradict each other eternally?"[19]

The gate, representing the present "moment," constitutes the meeting point between the two infinitely extending paths of past and future; but at the same time, these two infinitely extending lanes contradict each other, because they are expanding in opposing directions. The contradiction here is apparent: if the two infinitely extending lines are expanding in opposing directions, how is it possible for them to meet at the gateway of the moment? How can we believe that two paths "contradict each other eternally," when it is also true that they come together at a certain point?

To answer this riddle, we must interpret Zarathustra's gesture. Different from the dwarf's simple answer that "time is itself a circle," Zarathustra asks us to focus on the "moment" [*Augenblick*].[20] The solution to this riddle lies in the gateway: if it is true that the gateway of moment connects two opposing paths, then the only way it could connect them is by being the origin of their divergence. The moment is not simply a point within the infinitely extending timeline; it is rather the recurring points of rupture where the two lines of past and future break apart. In this sense, Nietzsche completely reverses the ontological priority between moment and time — it is the recurring moment that conditions the eternally expanding line of time, not vice versa. Moment is the point that breaks apart past and future, and precisely through this rupture of the moment, time becomes a line extending to two opposing directions. Without the constant rupture of the moment, time would not appear as a line.

This focus on the "moment," as the solution to the first riddle, immediately brings forth Zarathustra's second riddle. Rather than examining the gateway from the standpoint of the eternally extending lines, Zarathustra now turns to examine the eternal lines from the viewpoint of the gateway itself:

"Behold," I continued, "this moment! From this gateway, Moment, a long, eternal lane leads backward: behind us lies an eternity. Must not whatever can walk have walked on this lane before? Must not whatever can happen have happened, have been done, have passed by before? And if everything has been there before-what do you think, dwarf, of this moment? And are not all things knotted together so

[19] Nietzsche, 157-158.
[20] Nietzsche, 158.

firmly that this moment draws after it all that is to come? Therefore, itself too?"[21]

From the point of view of the gateway, "a long, eternal lane leads backward" lies behind the present moment.[22] The gateway is connected to the extending path, while also trying to step away from it. But when the backward extending line is "eternal," that is, when we imagine an infinite past that includes all possible instances, then this present moment must also be part of this eternal line. If the past extends eternally, then it is justified to claim that "whatever can happen has happened, has been done, has passed by before."[23] The contradiction is the following: if we interpret moment as the constant rupture that initiates the two infinitely extending paths, then the path that lies behind the gateway must be eternal; but if the path that lies behind the gateway is eternal, then the moment cannot be a true rupture, because the moment must have always-already occurred in the past. The concept of the eternally expanding path cancels the concept of moment. The hypothesis of the second riddle (that the past eternally extends) is now revealed to be contradictory to the solution of the first riddle (that moment is the breaking point of past and future).

The tension between the first and second riddles expresses the contradictory relation between eternity and the moment within the framework of linear temporality. Nietzsche's riddle of eternal recurrence reveals that the concept of moment is incompatible with the eternally extending timeline of past-present-future. In this sense, the problem of eternal recurrence is strictly a problem of temporality. The central concern of *Zarathustra* in Part III now becomes finding a new model of temporality that could solve the riddle without involving any contradiction.

§ THE TWO ETERNAL RECURRENCES IN ZARATHUSTRA

If the riddle of eternal recurrence essentially concerns the contradiction between eternity and moment, then there must exist two immediate approaches to resolving the paradox: 1. Emphasizing the concept of eternity and dissolving all moments into the eternal past; 2. Emphasizing the singularity of the moment and denying the infinitely

[21] Nietzsche, 158.
[22] Nietzsche, 158.
[23] Nietzsche, 158.

extending path of eternity. It is clear that the two approaches are mutually incompatible, precisely because they both attempt to resolve the contradiction by prioritizing one concept over the other.

This interpretive framework corresponds to the structure of Nietzsche's text. In the chapter "On the Vision of The Riddle," we encounter two opposing solutions to the riddle of eternal recurrence — one from the dwarf, and another from the shepherd. Here, the dwarf's solution corresponds exactly to the first way of solving the riddle of eternal recurrence. According to the dwarf, "all truth is crooked; time itself is a circle."[24] The dwarf bends the two infinitely extending lines together into a circle. For the dwarf, moment is only an illusion; he simply denounces everything Zarathustra proclaimed about the gateway of moment that joins the eternally extending path as "straight lies".[25] In the dwarf's interpretation, everything is eternally unalterable, because the moment never exists. Nothing new emerges from the eternally "crooked," unchanging truth. The dwarf, therefore, presents a metaphysical doctrine of eternal recurrence — a version of eternal recurrence literally speculating that everything recurs eternally in the same manner in the infinite scope of time.

The shepherd's solution creates an immediately contrast to the dwarf. The chapter ends with Zarathustra's vision of a shepherd choked by "a heavy black snake" [*schwarze schwere Schlange*].[26] Here, the heaviness [*Schwerigkeit*] of the snake signifies the gravity and also the difficulty (which are both expressed in the German word *Schwerigkeit*) of the riddle of eternal recurrence. In contrast to the dwarf who "[makes] things too easy" for himself, the shepherd's attempt to solve the riddle is to fiercely bite off the head of the snake.[27] If the snake represents the indefinitely extending path of time (since the snake is a crawling straight line), then biting off the head of the snake means literally cutting off the beginning of the eternal line. The shepherd's action, therefore, represents the triumph of the moment against eternity: through one single act of biting, the shepherd tears apart the unity of the eternal paths. As Zarathustra praises, the shepherd's laughter after biting off the snake's head is "no human laughter;" the shepherd, through the action of the "moment," conquered the recurrence of eternity and becomes a figure of *Übermensch*.[28] In this interpretation, the shepherd's action corresponds

[24] Nietzsche, 158.
[25] Nietzsche, 158.
[26] Nietzsche, 159.
[27] Nietzsche, 158.
[28] Nietzsche, 160.

to the second solution to the riddle, that is, to overcome eternity with moment and action.

This framework of the two doctrines of eternal recurrence also applies to the structure of the chapter "The Convalescent," which portrays Zarathustra's second attempt to summon the "abysmal thought" of eternal recurrence. Just as "On the Vision and the Riddle," "The Convalescent" introduces a similar opposition between two solutions of eternal recurrence: the animal and Zarathustra.

The animal attempts to resolve the paradox of moment and eternity by reducing the moment to a repeating instance within the eternal ring of being. Heidegger points out that the animal's description of eternal recurrence — "everything goes, everything comes back; eternally rolls the wheel of being" — is "at bottom identical with the talk of the dwarf."[29] The animal's claim that "the center is everywhere" and "bent is the path of eternity" also directly corresponds to the dwarf's belief that time is itself a circle. But different from the dwarf's mere speculation, the animal introduces a metaphysical justification for the doctrine: "But the knot of causes in which I am entangled recurs and will create me again. I myself belong to the causes of the eternal recurrence."[30] This attempt to ground eternal recurrence in the entanglement of causes clearly relates to Nietzsche's later conception of "combination of forces" in *The Will to Power*. For the animal, one eternally recurs to the same life through the same "knot of causes," just as the world infinitely repeats an "absolutely identical series" of combinations in infinite time due to the finitude of forces.[31] Zarathustra returns to the eternal selfsame life not because he wills such recurrence, but because all the other causes that co-constitute the world's combination of forces ("this sun," "this earth," "this eagle," "this serpent," even the "smallest man") infinitely recreate Zarathustra.[32] The animal's final presentation, therefore, already foreshadows the cosmological eternal recurrence of forces and power.

The animal's interpretation of eternal recurrence is immediately contrasted with Zarathustra, who now takes on the role of the shepherd by biting off the snake's head. Zarathustra's whole speech on the necessity of evilness (that "man needs what is most evil in him for what is best in him") reiterates the shepherd's

[29] Nietzsche, 217; Heidegger, 54.
[30] Nietzsche, 221.
[31] Nietzsche, *The Will to Power*, 549.
[32] Nietzsche, *Thus Spoke Zarathustra*, 221.

prioritization of moment over eternity and the attempt to overcome the burden of eternal recurrence through action. This is why Zarathustra mocks the animals as "buffoons and barrel organs."[33] Unlike the animals, Zarathustra feels an unbearable "nausea" at the thought that "the small man recurs eternally."[34] Here, the cause of Zarathustra's nausea is not the thought of eternal recurrence per se, but rather the ethical consequence of such thought. For Zarathustra, to admit eternal recurrence as a natural fact implies justifying the eternal recurrence of the last man. But to justify the eternal recurrence of the last man means that the last man can never be overcome, which consequently indicates that the *Übermensch* could never exist. This leads Zarathustra to argue that humans must consummate their destructive power, even the power of evilness, to become the highest creator — "whatever is most evil is his best power and the hardest stone for the highest creator."[35]

Zarathustra's doctrine of the necessity of evilness is, however, still a negative doctrine to be replaced. His doctrine of evilness, which strives to overcome the eternally recurring small man through an absolute, destructive power of overcoming, still has not overstepped the spirit of the lion pronouncing the "sacred No."[36] As "On the Three Metamorphoses" describes, the spirit of the lion "had been lost to the world" because it could only negate rather than create. Zarathustra's doctrine of evil resembles the spirit of the lion in its world-destructing tendency: through the consummation of humans' evilness, Zarathustra subordinates the world under his power. But the animal wants him to affirm the world — as they said to Zarathustra, "world awaits you like a garden."[37] The world awaits to be enjoyed and affirmed. For Zarathustra to recover from the sickness of his devastating negativity, he must learn to "sing and overflow" and become "the teacher of eternal recurrence."[38]

The animal's diagnosis of Zarathustra's sickness guides him to a higher formula of affirmation. This formula is different from the animal's because Zarathustra never completely accepts the animal's doctrine. The chapter ends with Zarathustra's silence as he turns to "conversing with his soul."[39] Zarathustra's conversation with his soul in the following chapter, "On the Great Longing," sublimates his destructive power of evilness into an affirmative desire for singing.

[33] Nietzsche, 220.
[34] Nietzsche, 219.
[35] Nietzsche, 218.
[36] Nietzsche, 27.
[37] Nietzsche, 216.
[38] Nietzsche, 220.
[39] Nietzsche, 221.

This new gesture of singing, rather than negating and overcoming, eventually culminates in Zarathustra's love song to eternity in "The Seven Seals." In each of the seven songs, Zarathustra reaffirms his desire for the "nuptial ring of rings, the ring of recurrence."[40] The chapter opens with Zarathustra imagining himself as the "soothsayer" trying to pronounce his lightening "Yes" to "a heavy cloud between past and future."[41] In the following songs, Zarathustra, in a Dionysian ecstasy, then imagines himself to drink the "full drafts from that foaming spice and blend-mug in which all things are well blended."[42] Here, Zarathustra's drinking of the "blend-mug" of things could be interpreted as his affirmation of the world's noumenal unity. Through these poetic imaginations, Zarathustra immerses himself in a creative ecstasy that innocently and unconditionally affirms the world in its eternity. Just as the first imagery of the soothsayer suggests, this affirmation dispels the uncertainty of the "heavy cloud between past and future"; it is a lightening, a purely "sacred Yes," that wills what is eternally to come.[43] Zarathustra's ecstasy, therefore, represents the "highest formula of affirmation" — an affirmation that demands one's willful love of what is eternally to come.

The affirmative approach to eternal recurrence overcomes Zarathustra's nausea. With the greatest cheerfulness in his spirit, Zarathustra pours "joy to pain" and "the most wicked to the most gracious."[44] Here, pouring is a sign of Zarathustra's excessive power. To pour "the most wicked to the most gracious" means to blend these two together — thus he admits that the lowest and the highest man are entangled eternally. His act of poetic affirmation seems to constitute a different formula of eternal recurrence that echoes what Löwith terms the anthropological-ethical equation. This affirmative doctrine of eternal recurrence departs from the animal's cosmological equation: rather than passively admitting recurrence as a metaphysical fact, one is now required to joyfully and affirmatively will eternity at every recurring moment.

It is evident now how the two formulas of eternal recurrence, rather than resulting from Nietzsche's changing conception throughout different works, are actually already present in *Thus Spoke Zarathustra*. As I have previously demonstrated, the riddle of eternal recurrence inherently brings forth two ways of solving the contradiction between

[40] Nietzsche, 228.
[41] Nietzsche, 228.
[42] Nietzsche, 229.
[43] Nietzsche, 228.
[44] Nietzsche, 229.

eternity and moment. These two solutions, being constantly repeated and reformulated throughout the course of Part III, eventually develop into two different equations of eternal recurrence: firstly, a cosmological equation of eternal recurrence of the animal proposing the recurrence of the same world in the entanglement of causes; secondly, an ethical equation of eternal recurrence of Zarathustra that poetically affirms and wills the recurrence of eternity in each moment. What Löwith terms as the cosmological equation and the anthropological equation of eternal recurrence are, therefore, not arbitrarily opposed to each other; they are rather two consecutive formulas that necessarily develop from Nietzsche's formulation of eternal recurrence itself.

§ ETERNAL RECURRENCE AND THE OVERCOMING OF NIHILISM

I shall now turn to how this interpretive framework of the internal duality of eternal recurrence could help rethink a central theme in Nietzsche's philosophy — the overcoming of nihilism. This would require us to turn our attention to the chapter "The Wanderer," which appears right before "On the Vision of the Riddle."

Many interpretations of *Zarathustra* tend to underestimate the role of the chapter "The Wanderer" in Nietzsche's presentation of the doctrine of eternal recurrence. For interpreters like Laurence Lampert, this opening chapter of Part III continues the prominent theme of ascent and descent. The chapter opens with Zarathustra's mountain-climbing to "set off as a wanderer back to his solitude" and ends with Zarathustra's descending again for the "love of mankind."[45] In this interpretation, Zarathustra's ascent and descent echo the opening of Parts I and II and mark Zarathustra's readiness to face a new task.

Rather than stressing the transitional function of "The Wanderer" in Part III of *Zarathustra*, I tend to interpret "The Wanderer" as a culmination of the philosophical concept of self-overcoming. In my interpretation, Zarathustra's act of mountain-climbing indeed serves as a metaphor for constant self-overcoming — a motif that dominated previous parts of the book. By climbing over the mountain to return to his solitude, Zarathustra not only seeks to overcome his disciples (as is evident in the last chapter of Part II), but also to

[45] Laurence Lampert, *Nietzsche's Teaching: An Interpretation of Thus Spoke Zarathustra*, (Yale University Press, 1989), 158, 160.

overcome himself and his own teaching. As Zarathustra claims in his own speech: "I am a wanderer and a mountain climber, he said to his heart; I do not like the plains, and it seems I cannot sit still for long."[46] Zarathustra cannot "sit still for long"; he is the figure of becoming. He only belongs to the "ridges and peaks" because he always oversteps his own limits and never stays at a static, unchanging "plain" of being.[47] Zarathustra not only strives to overcome others but also continuously seeks to overcome himself — this is why Zarathustra's journey alternates between undergoing (to overcome others) and ascending (self-overcoming).

In my interpretation, this radical notion of overcoming expresses a radical form of active nihilism. As the hour speaks to Zarathustra, "You are going your way to greatness: here, nobody shall sneak after you. Your own foot has effaced the path behind you, and over it there is written: impossibility." [48] The "path," which marks the past moments of overcoming, is constantly effaced by one's own "foot" (the present moment). In this overcoming, Zarathustra "lacks all ladders" and can only "climb on [his] own head".[49] As Katharina Grätz commented here, "climbing on one's own head" is a paradoxical imagery that physically denotes an impossible challenge of self-overcoming [*unmögliche Aufforderung zur Selbstüberwindung*].[50] This imagery, combined with the metaphor of "lacking a ladder," indicates that Zarathustra's overcoming is not a step-by-step transformation, but rather an abstract and endless destruction of what is present-at-hand. For Nietzsche, this radical notion of self-overcoming offers an impossible task. Extreme self-overcoming is like a Hegelian bad infinity, in which every moment is immediately replaced by the next and so on ad infinitum. This vicious cycle of recurring moments turns overcoming into "impossibility" because it can never be completed. In this sense, the radical self-overcoming corresponds to Nietzsche's description of an "active nihilism," which is nothing other than "a violent force of destruction."[51] That is to say, this self-overcoming can only manifest itself as insatiable self-negation and self-destruction.

This problem of active nihilism is crucial to our understanding of the context of eternal recurrence because it ultimately presents

[46] Nietzsche, 152.
[47] Nietzsche, 152.
[48] Nietzsche, 153.
[49] Nietzsche, 153.
[50] Katharina Grätz, *Kommentar zu Nietzsches "Also sprach Zarathustra III und IV."* Historischer und kritischer Kommentar zu Friedrich Nietzsches Werken, Band 4.2. (De Gruyter, 2024), 21.
[51] Nietzsche, *Will to Power*, 18.

a problem of temporality. If we say that the "path" represents the unity of linear temporality (the line of past-present-future), then self-overcoming, in my interpretation, represents another structure of time constituted by "moments." In the temporality of overcoming, the instantaneous recurrence of moments replaces the unity of linear time. However, as our previous analysis demonstrates, the temporality of the moment is also empty, homogeneous, and self-negating; it is trapped in the cycle of the recurring new moments. The entire discussion of the active nihilism of self-overcoming lays the groundwork for Nietzsche's discussion of eternal recurrence — the ever-recurring new moments that are trapped in the impossibility of self-overcoming represent the same dilemma of the gateway of moment in Zarathustra's riddle. This present moment of overcoming, just as how the gateway of moment has infinitely recurred in the past, must have always-already eternally recurred. But if moment eternally recurs, then the whole series of overcoming is never complete. This contradiction within the temporality of the moment reveals the impossibility of a philosophy of overcoming; it also reveals the impossibility of holding the stance of active nihilism.

If eternal recurrence functions as a problematization of the paradox of self-overcoming, and the philosophy of self-overcoming itself expresses the highest stage of active nihilism, then it necessarily follows that the doctrine of eternal recurrence is also an attempt to transcend the dilemma of active nihilism. In this sense, the two formulas of eternal recurrence are more than solutions to the literal riddle itself. Just as Laurent Lampert points out, "the teaching on eternal return opposes any teaching on the linearity of time that points toward some future eschatological fulfillment of time."[52] The teaching of eternal recurrence replaces the philosophy of overcoming (for Lampert, even the philosophy of the *Übermensch*), which is clearly based on a linear concept of temporality. For Nietzsche, the animal's speech of eternal recurrence "ends Zarathustra's going under" — this means that the central concern of Zarathustra's teaching shifts from the philosophy of self-overcoming to the problem of eternal recurrence.[53] The two doctrines of eternal recurrence — the cosmological and the ethical — actually express two steps of transcending the philosophical framework of active nihilism.

The two formulas of eternal recurrence, as two essential steps of overcoming nihilism, reexplain what Nietzsche means by the

[52] Lampert, 258.
[53] Nietzsche, *Thus Spoke Zarathustra*, 221.

"revaluation of all values" in his later projects. As Nietzsche writes in his notebook as a plan for the book titled *The Eternal Recurrence*, the book will involve: "1. Presentation of the doctrine and its theoretical presuppositions and consequences. 2. Proof of the doctrine. 3. Probable consequences of its being believed (it makes everything break open). a) Means of enduring it; b) Means of disposing of it."[54] In Nietzsche's project, the theoretical presentation and proof of the doctrine of eternal recurrence precedes the "means of enduring" and "disposing" of the doctrine. For Nietzsche, the scientific-metaphysical proof of the fact of eternal recurrence is only a preparation for enduring and disposing of such a doctrine. The center of Nietzsche's concern is, therefore, never the truthfulness or accuracy of the theory, but how the interpreter could reimagine and relocate one's own existence under the new vision of the world. In this sense, the ethical equation of eternal recurrence must eventually replace the cosmological equation; it is only in the ethical equation that eternal recurrence could be "endured" rather than "proven."

This is why Nietzsche associates enduring eternal recurrence with "revaluation of all value." To endure the fact of eternal recurrence is not a step toward the revaluation of all values; this act of endurance is rather already the process of revaluation itself. Just as Nietzsche writes in another notebook entry, to endure eternal recurrence demands a shift from the expression "everything is merely subjective" to a proud attitude that "it is also our work."[55] While the expression "everything is merely subjective" presents a retreat to one's own subjectivity (just as what we see in the paradox of self-overcoming), the proud claim that "it is also our work" expresses one's willingness to directly confront reality and acknowledge one's active role in it. To endure the reality of the world requires not only the creation of new values, but also self-affirmation of our creative role in the revaluation.

This reading of eternal recurrence emancipates Nietzsche's philosophy from the burden of metaphysics. Heidegger famously claimed that Nietzsche's metaphysics should be conceived as a "fulfillment of nihilism proper."[56] But as we have shown, this claim is problematic because he ignored Nietzsche's own understanding of how nihilism is to be overcome. Heidegger seems to reduce eternal recurrence to a metaphysical fact. As he writes, "[w]e observes that being, which as such has the fundamental character of will to power,

[54] Nietzsche, *The Will to Power*, 544-545.
[55] Nietzsche, 545.
[56] Martin Heidegger, *Nietzsche: Volume Four*. Ed. David Farrell Krell. Tran. Frank A. Capuzzi, (Harper, 1993.), 204.

can as a whole only be eternal return of the same."[57] The problem is that Heidegger's interpretation stops at the cosmological equation of eternal recurrence. However, as my previous analysis shows, the cosmological equation of eternal recurrence, rather than being the center of Nietzsche's philosophy, is indeed only a preparatory step toward the more affirmative equation of eternal recurrence. Only the affirmative, ethical equation of eternal recurrence can overcome nihilism. Exactly because Heidegger still wants to ground Nietzsche's eternal recurrence in the metaphysics of will to power and forces, he reduces Nietzsche's philosophy to a continuation of nihilism in Western metaphysics.

The affirmation of eternal recurrence is not a simple subjective claim that adds nothing to the facticity of the doctrine. Through affirming the doctrine, we also poetically recreate the world in a new shape — just as how Zarathustra reimagines the world through poetic imagery in "The Seventh Seals." In this sense, the problem of eternal recurrence is always more than a cosmological-metaphysical problem. It is never enough to simply acknowledge eternal recurrence as the "most scientific of all possible hypotheses." The cosmological formula wants to demonstrate the impossibility of overcoming by proving the fact that everything eternally recurs; but this formula never fully overcomes the state of active nihilism because it is still concerned with the problem of truth and falsity. To truly overcome nihilism, one needs not only the knowledge, but also the affirmative will for eternal recurrence. Nietzsche's philosophy ultimately wants "a Dionysian affirmation of the world as it is, without subtraction, exception, or selection—it wants the eternal circulation: —the same things, the same logic and illogic of entanglements."[58] Such Dionysian affirmation could only be attained in the absolute affirmation of the world through one's active willing of recurrence. Eternal recurrence must eventually culminate in an ethical gesture; only this ethical gesture completes the process of overcoming active nihilism.

§ CONCLUSION: A PHILOSOPHER IN BECOMING

In conclusion, my essay strives to accomplish two tasks: 1. Demonstrating how the two formulas of eternal recurrence are already evident in Part III of *Zarathustra* and revealing how it is inherent to the problem of eternal recurrence; 2. Showing how this two-fold nature

[57] Martin Heidegger, *Nietzsche: Volume Three*. Ed. David Farrell Krell. Trans. Joan Stambaugh, David Farrell Krell, and Frank A. Capuzzi, (Harper, 1993.), 210.

[58] Nietzsche, 536.

of eternal recurrence relates to the overarching theme of Nietzsche's philosophy — the overcoming of nihilism.

To me, Nietzsche is not only a philosopher of becoming, but more importantly, a philosopher in becoming. Becoming is not only a thematic concept of Nietzsche's philosophy; it is rather the very essence of Nietzsche's process of philosophizing. This brings forth a different way of understanding Nietzsche's philosophy of eternal recurrence: rather than obsessing with the validity of Nietzsche's proof or the coherence of Nietzsche's metaphysics of forces, it might be more fruitful to understand the genesis of eternal recurrence in Nietzsche's thinking. Nietzsche's struggle with the problem of eternal recurrence is already what constitutes the movement of overcoming nihilism. In this sense, it is meaningless to think about whether Nietzsche's metaphysics corresponds to the truth. Nietzsche reveals that what ultimately matters is not the doctrine itself but whether one could affirm it in eternity. The problem of eternal recurrence is more than a simple riddle concerning the relation between moment and eternity; as my interpretation intends to demonstrate, the problem rather becomes a meta-philosophical question concerning the thinker's relation to the theory. Nietzsche as the philosopher of philosophizing — this is the figure of Nietzsche that my essay wants to put forward.

BIBLIOGRAPHY

Grätz, Katharina. *Kommentar zu Nietzsches "Also sprach Zarathustra."* Historischer und kritischer Kommentar zu Friedrich Nietzsches Werken, Band 4 2. De Gruyter, 2024.

Heidegger, Martin. *Nietzsche: Volumes One and Two*. Repr., 3. pr. Translated by David Farrell Krell. Harper, 1993.

Nietzsche: Volumes Three and Four. Repr., 3. pr. Edited by David Farrell Krell. Translated by Joan Stambaugh, David Farrell Krell, and Frank A. Capuzzi. Harper, 1993.

Lampert, Laurence. *Nietzsche's Teaching: An Interpretation of Thus Spoke Zarathustra*. Yale University Press, 1989.

Lowith, Karl. *Nietzsche's Philosophy of the Eternal Recurrence of the Same*. 1st ed. Translated by J. Harvey Lomax. University of California Press, 1997.

Nietzsche, Friedrich. *The Case of Wagner, Twilight of the Idols The Antichrist, Ecce Homo, Dionysus Dithyrambs, Nietzsche Contra Wagner*. Translated by Adrian Del Caro, Carol Diethe, Duncan Large, George Heiner, and Paul Loeb. The Complete Works of Friedrich Nietzsche, volume 9. Stanford University Press, 2021.

The Joyful Science / Idylls from Messina / Unpublished Fragments from the Period of The Joyful Science (Spring 1881–Summer 1882). Translated by Adrian Del Caro. The Complete Works of Friedrich Nietzsche / Edited by Alan D. Schrift, Duncan Large, and Adrian Del Caro, volume 6. Stanford University Press, 2023.

The Will to Power. Vintage Books ed. Translated by Walter Arnold Kaufmann and Reginald John Hollingdale. Vintage Books, 1968.

Thus Spoke Zarathustra: A Book for None and All. Translated by Walter Kaufmann. Penguin Books, 1978.

Oger, Eric. "The Eternal Return as Crucial Test." *Journal of Nietzsche Studies*, no. 14 (1997): 1–v18. http://www.jstor.org/stable/20717674.

THE SOLIPSISM OF SELF-INTEREST
REVIVING NAGEL'S ABANDONED ARGUMENT

WILLIAM THOMAS

§ INTRODUCTION

The core argument of Thomas Nagel's *The Possibility of Altruism* (1970) seeks to demonstrate that the moral sceptic, he who refuses 'to be motivationally persuaded by moral considerations',[1] cannot rationally engage in this refusal without implicitly endorsing a form of solipsism. This constitutes a denial of the reality of others' subjective experiences, identifying oneself as the only conscious being in the world. Nagel argues that Altruism, or behaviour motivated by genuine, selfless concern for the well-being of others, is made possible by the 'presumably universal'[2] recognition of oneself as merely one person among others equally real.[3]

The radical idea at the heart of Nagel's book is found in his argument that all reasons motivating action must be objective, or 'agent-neutral', to use Parfit's more modern terminology,[4] later adopted by Nagel to denote the same concept.[5] That is to say a reason

1 Nagel, Thomas. *The Possibility of Altruism.* Oxford: Clarendon Press, 1970: 143.
2 Ibid.,145.
3 Sturgeon, Nicholas L. "Altruism, Solipsism and the Objectivity of Reasons", *Philosophical Review,* 83, No. 3 (1974): 374-402 paraphrasing Nagel, *The Possibility of Altruism,* 14.
4 Parfit, Derek. *Reasons and Persons*, Oxford: Clarendon Press, 1984.
5 Nagel's use of 'objective' and Parfit's of 'agent-neutral' are not exactly the same, as their formulations differ substantially. They are, however, extensionally equivalent in all possible worlds as pointed out by Ridge in 'Reasons for action: Agent-Neutral vs Agent-Relative' *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). The difference is largely unimportant for the purposes of this paper.

of this sort "can be given a general form which does not include an essential reference to the person who has it".[6] Resisting relativisation to any agent in particular, these are reasons to perform actions that promote (by which Nagel means roughly 'help bring about'[7]) ends that are universally desirable, and do not depend on an agent *being* a particular individual. Subjective, or agent-relative, reasons, are only reasons *for particular individuals* to act, and do not promote universally desirable ends.

Nagel argued that the acceptance of subjective reasons that do not have universally motivating content is inconsistent with the view that other minds are comparably real. As to be discussed, this was entirely dismantled by Nicholas Sturgeon in his 1974 paper *Altruism, Solipsism, and the Objectivity of Reasons*.[8] Nagel abandoned this argument in light of these objections, and as far as I can tell, no one has seriously attempted to revisit it since.[9] That is not to say that the underlying idea was shown to be completely hopeless; Michael Ridge claims the argument is 'ingenious… [and Nagel] may have abandoned it prematurely.' If it could be done successfully, 'the implications would be dramatic', upending the study of normative philosophy as we know it.[10]

In this paper, I argue that Nagel was right to think that accepting any reason for action that prioritises the interests of the self over others conflicts with the belief in the reality of other people's experiences.

I do not wish to take the normative standpoint that one should never act in a self-interested way. I only claim, like Nagel, that to allow acceptance of reasons to act based in any part on the fact that those reasons concern the self, rather than another, implies a sort of solipsism that denies a certain equality to the reality of experiences of others.

Nagel's argument is heavily preoccupied with, and I think burdened by, discussion of the language of reasons, and the extent to which propositions concerning reasons can be expressed personally and

---

6 Nagel, Thomas. *The View from Nowhere.* New York: Oxford University Press, 1986: 152-153.

7 Nagel, *The Possibility of Altruism,* 47.

8 Sturgeon, Nicholas. "Altruism, Solipsism and the Objectivity of Reasons", *Philosophical Review,* 83, No. 3 (1974): 374-402.

9 A huge amount of work has been done on similar ideas, most notably by Parfit (1984), but to give due attention to particular examples would require a great deal of care and would be overly ambitious for this paper.

10 Ridge, Michael. 'Reasons for action: Agent-Neutral vs Agent-Relative' *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition).

impersonally. Whilst the result is an elegant and impressive framework, the ambition to actually *prove* his conclusion using these notions turned out to be extremely fragile, as Sturgeon clearly demonstrated in his decisive response.

I propose that there is a more direct way to argue Nagel's point, which bypasses many of Sturgeon's objections. It employs certain ideas similar to those Nagel held dear and emphasised passionately at times, but did not manage to apply in a watertight way. This is predominantly his notion of the impersonal standpoint, and some accompanying insights regarding the extent to which self-locating information can rationally change one's attitude towards reasons.

I will not go into great detail about the intricacies of Nagel's argument, as interesting as it is, as he abandoned it after Sturgeon's response, demonstrating that his particular arguments were not viable. Despite this, the conclusions remain fascinating. Though predominant normative ethical theories *sound* similar, to say that one *should* promote actions in the interest of everyone is, despite being *prima facie* more intuitive, far less tangible than Nagel's suggestion that to accept *any reason* to act on the grounds that it promotes your self-interested ends actually conflicts with the belief in the equal reality of other minds.

I will first present a revised way to conceptualise Nagel's impersonal standpoint and then show how this revision entails a similar conclusion to the one Nagel originally contended, though in a more direct and stable way, as it establishes a clearer framework for structuring the facts about the world inspired by Nagel's ideas. It is hoped this avoids the fatal criticisms that Sturgeon brought against the intricate but frail semantic foundations for Nagel's argument in *The Possibility of Altruism*. It will be demonstrated that once the reality of other minds has been accepted, along with the fact that their experiences also include an intrinsic sense of self, there is no additional agent-indexing information concerning the self which could logically be said to have reason-generating weight. I will then explain briefly how this builds positively on certain useful features of Nagel's original argument, whilst avoiding the damning criticisms brought against him by Sturgeon.

§ THE REVISED 'IMPERSONAL' STANDPOINT AND THE ARGUMENT THAT FOLLOWS:

This particular way of thinking about objectivity is not at its core dissimilar to many other, more typical ways that would be familiar to all those who have thought about the matter, but it should bring to light certain fundamental features which, if neglected, will leave an argument discussing objectivity vulnerable to criticism, as seen with Nagel and Sturgeon. As such, it will seem familiar at first, and not particularly groundbreaking, though its merits should reveal themselves shortly. It must be warned that the following section is addressed consistently in the second person, yet makes demands throughout to think about identity in a way that might seem incompatible with this. This is often the case in any discussion of personal identity, but is particularly difficult here. The relevant confusions are clarified afterwards.

To engage with the revised impersonal standpoint, it is helpful to begin with a thought experiment. It asks you to imagine you are an impersonal, omniscient, god-like figure viewing the world and its (for simplicity's sake, human) inhabitants. This may confuse matters slightly to start with, as imagining oneself as an impersonal being is obviously impossible, and to think of it as *you* would be to contradict the entire point of the exercise. However, this kind of difficulty is unavoidable in any thorough discussion of objectivity, and the exercise is ultimately necessary to elucidate key aspects of the concept.

As this impersonal, omniscient being, you are aware of the total set of facts about the world. This includes every detail about the conscious experience of every individual. There is no need to actively try to imagine what it might be like to actually *be* one of those people by piecing together certain facts you know about their mind; your knowledge is exhaustive, and your understanding of the unique intricacies of each subjective experience would be included within it.

Now suppose you had ultimate control over the actions of every person. Intuitively, you would have no reason to prioritise certain individuals over others,[11] so it would be natural to assume a framework for decision making in which the people under your control act entirely selflessly and do not reflect any sort of self-centred attitude.There

[11] Assuming that you don't judge some individuals to be worthier than others, which might lead you to, say, bring about some kind of karmic cycle. We have not imposed a value system on this being, so this would not make any sense. It would, in fact, be largely inconsequential to the argument, as long as the impersonal aspect is maintained. In fact, the flexibility to accommodate a variety of suggested impositions of objective, impersonal value systems that differ from the utilitarian sentiment passively implied by a neutral standpoint like this could add to the appeal of the argument.

would be no reason for you to have an arbitrary individual, say, person A, do things that benefitted the interests of person A over those of person B, as you would appreciate that there is nothing special about person A's interests.[12] If you were to arbitrarily assign a normative sentiment to this framework it might appear somewhat utilitarian. However, I emphasise that there is a significant difference between having reason to make decisions that promote everyone's interests equally and making decisions that promote everyone's interests equally in the absence of any reason to do otherwise. Though perhaps extensionally equivalent, the two ideas are not to be confused, and it is the latter we are concerned with.[13]

Naturally it would be most reasonable for you to have A frequently focus on certain basic self-directed matters, for example, keeping himself alive. However, this would only be because there is opportunity cost involved in every decision, and there are many things that make much more sense for people to do for themselves purely in the name of efficiency.

In any case, when you make decisions for A, you have no reason to use the affective power of A's body to promote A's own interests over B's precisely because you are aware of the fact that B's conscious experiences are equally real.[14] It must be remembered that the conscious experiences of these people have not changed, and they have not been turned into lifeless puppets by your control.[15] They both still have the same sense of self they had before: the 'I am person

[12] It must be noted at this point that the term 'interests' must be treated with caution, as it will be put under scrutiny in due course. For example, it will be asked in what sense they are *A's* interests at all. At this stage, this sort of question only confuses things, and I recommend the word is taken at face value.

[13] They may, of course, converge quite naturally, in that once the general principle that 'decisions should promote interests' is added, without reason at this point to favour any individual's interests over another, the doctrine looks a lot like utilitarianism. However, there is no reason at this stage to adopt this principle.

[14] Assuming that if one were *more* real, or the only real one, this would give reason for the impersonal being to promote their interests; this is the link to solipsism. A more thorough discussion could ask why this might be, but it seems generally reasonable and suits the assumptions of the present argument.

[15] It could be objected that there is no way that you could take over their autonomy without changing their experience of the world. Questions about free will aside, this is probably true, but the relevant part is their sense of self rather than their sense of autonomy. It may be argued that the two are inextricably linked, but their separation is not essential to the argument, only the thought experiment that helps clarify the underlying idea. It is asked therefore that disbelief is suspended on this matter.

A rather than person B' and vice versa, as the subjective element is embedded so deeply in experience.

As such, knowing that A feels like A, and B feels like B, and that each sense of self is equally real, it would no longer make sense, when making decisions using A's affective bodily power, to reason as A perhaps did before you took control; to justify[16] the promotion of his own happiness, for example, by saying 'I have reason to promote my happiness *because it's mine*'. B would likely do the same. Given that as an impersonal being you are neither A nor B, this reason, as stated by either person, would be redundant in your decision making; this is because, to emphasise once again, A's sense that his happiness is his own cannot generate a reason for you[17] to promote that over B's identical feeling, as both are equally real. All that claiming 'it is mine' does is trivially index the feeling back to a particular stream of conscious experience of which that feeling is an intrinsic part. This point will be clarified in much greater detail further on.

There is no use emphasising this point about objective reasons more firmly, as it has been made many times before. It is at this point that a difficult, but crucial, conceptual leap must be taken, which is where this argument begins to stray from Nagel's and other conventional formulations. Two points must be remembered here. Firstly, recall that you have all the facts about every individual's mind, and you know exactly what their conscious experience would be like without *actually being* that person. Secondly, these facts include that each person has an equally real sense of self, which may be crudely captured in the thought, 'I am *me*, rather than anyone else'. Necessarily, person A feels like person A, and person B feels like person B. The *feeling of being* person A rather than person B is intrinsic to person A's experience, and is included within the set of true facts about A (that, as an omniscient being, you already know).

Now imagine that you 'jump' from the impersonal being into a random person's mind, say, either A or B. You will experience their stream of consciousness exactly as it is, and it will be exactly like you knew it would be when you had all the facts about it as an omniscient

[16] Rationally, not necessarily morally.

[17] It is extremely important to add that up to here, 'reason *for you*', is used to in fact make clear that the reasons in question are objective, distinct from A's possible agent-relative reasons. This is confusing as '*for you*' almost always implies agent-relativity, but here the reason is being relativised to an impersonal perspective to represent objectivity in a slightly more tangible way. This is just another example of the unavoidable difficulties conceptualising objective reasons.

being. You will not learn any new information about it, as there is none to be learned–you knew all the facts already.[18]

The random jump happens, and you find yourself in person A's mind. You know this because you are experiencing being person A rather than person B. You have A's sense of self: *I am A, I am not B*.

This piece of information, which tells you that you are A rather than B, appears helpful. It seems it could give you reason to start using your bodily autonomy (which is now confined to a single body, A's) to perform actions that promote person A's interests where you did not have reason before. You can now promote A's happiness over B's, and the reason to do so has strengthened greatly; you have learned that you are A, and therefore A's happiness will now be *yours*, giving you, A, reason to promote this outcome where the impersonal being had none.

This reasoning has no logical plausibility whatsoever. To understand why requires some patience, as the explanation, though I think it to be beautifully revealing, is incredibly unintuitive and requires significant backtracking with respect to the linguistic and conceptual habits that have so far been used to get to this point, not just in this paper but in the historical discussion as a whole.

The essence of the point is this: *the fact that you are now A rather than B is not a piece of information*. It has no informative power, once the fact that A already had a sense of self is acknowledged. What exactly is it that you think you have learned? You may say, 'I have learned the fact that I am now A, and not B'. That fact is, of course, true, but the sense in which you have *learned* it is entirely unclear. You may say, 'I used my new intrinsic sense of being A as an evidential basis, from which I learned that I am indeed A, and not B'.

But how can this count as learning information? It was already part of the known set of facts that A experiences the sense of being A. What you are saying here is that 'now *I* experience being A', even though you are A, so this is just the same as 'A experiences being A';

[18] Of course, in reality, experiencing their mind exactly as it is will involve a great loss of information, and the complete collapse of omniscience. You will have only the memories and knowledge of whichever person's mind you jump into. I have added this as a footnote because at this stage the reminder that your entire cognitive basis is now in the new brain inhibits the usefulness of the idea of the jump from one to the other. It is of course true, and a more accurate view of things, which will prove highly significant.

but this was already part of the known set of facts. Nothing new has been learned at all. No new information has been gained.[19]

Before explaining how this entails that there can be no purely subjective reasons for action consistent with the belief in the reality of others' experiences, there are some obvious objections to this that should be covered. Primarily, it may be said that there is a large disconnect between the way certain concepts are discussed in the first, 'pre-jump' section of the thought experiment. This makes the conclusion seem trivially true at face value, uninteresting when considered in isolation, and incoherent and confused when contextualised against the rest of the idea. I think this first impression is unfortunately unavoidable, and to get over it requires the backtracking alluded to earlier, which I shall outline below.

This impression largely stems from the necessity to use pronouns, and to generally refer to 'you', or whoever is reading and takes the pronoun to refer to them, and additionally the fact that this person (you) is *a real person* and not, of course, an impersonal observer (I hope). This is problematic because the way that the thought experiment deals with changing identities is not compatible whatsoever with either normal practical language use or the fact that we are constrained to a subjective viewpoint. This means that it makes no sense at all to say, 'now you are A, where before you were the impersonal observer'. This is further explained by two distinct but connected reasons.

Firstly, as previously stated, there is no way '*you'* could be an impersonal observer, as this observer would have no self for the '*you'* to refer to. Secondly, there is a significant element of deception in which '*you*' is used interchangeably to refer to: a) a real person who is asked to imagine they have a different identity, and b) not one, but a number of distinct imagined counterparts through which this identity is supposed to move. This creates a sort of illusion where it seems it is possible to imagine, and appears to make sense to describe, a single persistent identity moving between distinct imagined entities. This is because the sense of self felt by the real person doing the imagining acts as a surrogate for the imagined persistent personal identity to cling to, and this is able to move from one imagined entity to another with a supposedly different identity, whilst the real point of reference

[19] This may seem like one of Frege's Identity Puzzles, but the sense/reference distinction is not needed here, as 'I' and 'A' will have the same cognitive value for A, so long as the 'I' really does now (as it should) refer to A, and not a persistent metaphysical self. How this may be avoided will be explained in due course.

of 'you', and with it your intrinsic sense of self, stays constant. In reality, this could not possibly make sense, as once the person referred to changes, it is incoherent to use 'you' to refer to a single persistent personal identity.

In light of this, an objection could be raised that the whole project was entirely dishonest, and the conclusion reached utterly trivial. However, it is the very fact that these arguments may be raised, acknowledged, and responded to accordingly that allows this formulation to ultimately avoid the criticisms Sturgeon leveled against Nagel. The arguments will be considered one by one.

Firstly, the impersonal observer is merely a conceptual placeholder for Nagel's 'impersonal standpoint' which, importantly, is not a perspective at all.[20] Though this undoubtedly makes full conceptual understanding of the idea impossible, the addition of the quality of omniscience allows us to bypass the idea that objectivity needs to be a perspective at all, rather than just the consideration of the facts as a whole. This, of course, is not a new idea; Nagel was very aware of this at the time of writing. Nagel emphasised the necessity to include the full, unchanging set of facts in the impersonal perspective.[21] His conception of the full set of facts was somewhat incomplete, as will be discussed in due course. Essentially, what is important here is the available information, and what reasons are coherent in the *presence* of all facts but the *absence* of self-locating information. It has been argued that this cannot serve as relevant information at all, and so the importance of reason consideration without it is obvious. Moreover, the lack of identity in the impersonal observer is helpful in that it makes the jump and adoption of identity less confusing than it might be if, say, the jump were made from one person to another, requiring substitution rather than mere adoption.

The second objection highlights the problem that the conceptualisation demands that you imagine moving from one person with a sense of self to another, without acknowledging the fact that there can be no 'you' to move from one to the other if this move is supposed to take place accurately, but rather by taking advantage of the persistent sense of self in the person conceptualising to stabilise the illusion.

[20] His difficulty with the impossibility of engaging with this idea in full is covered in more detail in Nagel, Thomas. *The View from Nowhere.* New York: Oxford University Press,1986.

[21] Nagel, *The Possibility of Altruism*, 103-104.

It certainly doesn't take advantage of this deliberately; on the contrary, the natural tendency to subconsciously confuse the persistent sense of self of you, who is doing the imagining, with something real within what is being imagined, is a great hindrance to the point being argued for, and only confuses the matter. If the human brain were better at imagining similar situations without this sense of self, the absence of an imposed persistent entity (the sense of self of the person imagining) that tracks identity-locating information *as it changes* would be much clearer; it would make the point that this appears as information only because of the illusion such an entity could persist[22] far more lucid and intuitive. If one could move from imagining being person A to imagining being person B without the unshakeable sense that some sort of memory-preserving faculty moves between the two, then the idea that no new information is gained on the 'discovery' that one is now B *if one truly is B, as B was previously factually described,* would be much less opaque.

It is very confusing to have it suggested that you are *now* person A, whereas you weren't before, whilst also stating as a fact that A had had a persistent, unchanging sense of self before and after A was you, whereas *you* only got this when you jumped into A's consciousness. This does, in a way, unfairly imply a separation between 'you' and 'A' that it would be unreasonable to demand you forget. It does not make sense to speak of things this way.

However, it is this very fact that makes the argument work. It is precisely this failure of language that illuminates the reason it does not make sense to speak as if information is gained upon the discovery that 'I am now A', as this could only be the case if it is assumed that the 'I' and the 'A' could have existed separately before in a way that might render their subsequent conjoining factually significant. This is the error we are psychologically compelled to make, as explained above. But if, as the thought experiment proposes, there is a stable set of facts about A's conscious experience, and when A's brain is merely 'jumped into' this set of facts does not change in any way, the illusion of any sense of personal identity that existed before becoming A should disappear. This is exactly what is intended, as the impersonal observer, accurately imagined, should not have had a sense of self to cling to.

Essentially, once the totality of facts has been established, there is no impersonal or personal perspective from which any information could be gained. Trivially, this is because all information

[22] It is assumed here that such an entity could not exist.

is contained within the established totality of facts. Included within this are the facts about your personal experience, including the sense of self that is intrinsic to it. Awareness of the self, and its autonomous capabilities and limitations, informs you of the decisions it is possible to make, and therefore defines the space of possible outcomes which are practically worthwhile for your reason-assessing cognition to consider promoting.

It cannot be emphasised strongly enough that, where agency is concerned, self-locating information is fundamental in this way; this does not by any means imply that self-locating information has any reason-generating power. It is not actually the same form of information at all, and to call awareness of the autonomous capabilities and limitations of oneself as an agent, 'self-locating' is only significant because of the fact that, given that you only have control over your own body, to know exactly what those limitations and capabilities are would give a very accurate indication of who exactly you were in the world.

Briefly returning to the impersonal observer, let's imagine that instead of being in control of all people's actions, you were assigned one at random. You were then given all relevant information about what decisions could and couldn't be made, using the affective bodily autonomy of this particular individual. This would tell you all you need to know about which person you actually had control over; knowing one of your capabilities was to wiggle A's toes, you could safely assume you were in control of A. However, this wouldn't necessarily mean *you are A*. You still don't have the sense of self that A does.

This distinction is extremely helpful in understanding why purely self-locating information is redundant where reason generation is concerned. Upon 'discovering' that you are A, rather than B, the only information that is relevant to reasoned decision making is purely that which defines the space of possible decisions. It is crucial to understand that just because self-locating information *in practice* appears to coincide with the facts that define the bounds of agency for a person, and are thus fundamental to the mechanism of decision making, the sense of self provides no real non-trivial factual information that could affect the reasoning process that ultimately governs that mechanism. The facts that 'I am A' and 'I will feel A's happiness' do not represent any non-redundant differentiating factors where the outcomes of rational decisions are concerned, where belief in the reality of other minds is accepted.

§ NAGEL'S FAILURE AND STURGEON'S CRITICISMS

Nagel's argument was reconstructed by Sturgeon in the following way:

> Nagel sought to show that it is contrary to reason not to believe that all reasons are objective.To reject this involves violating two conditions, (C) and C'):
>
> (C) all his *impersonal* considerations must have for him the same motivational content as do his self-regarding considerations;
>
> (C') all his *other regarding* considerations must have for him the same motivational content as do his self-regarding considerations;[23]

Where impersonal considerations are those made from the third-person perspective, that involve no reference to the self, where the person with the consideration is barred from regarding it with the additional information that they might in fact be the individual being described impersonally, and could rephrase the statement in the first person.

Self-regarding considerations are, as Sturgeon characterises them, propositions that one could express sincerely[24] by a sentence of the form:

> I have reason to promote A, or to want A promoted;[25]

Other regarding considerations:

> Someone else has reason to promote A, or want A promoted;[26]

Which importantly includes the information that this 'someone *else*' is not the person making the consideration, whereas in impersonal considerations, this sort of information is absent.

Violating these conditions, Sturgeon argues, causes 'dissociation', a word Nagel uses to describe the loss or lack of the

[23] Sturgeon, "Altruism, Solipsism and the Objectivity of Reasons", 381.
[24] In Sturgeon's words, 'someone expresses a proposition sincerely just in case he means what he says'. Ibid., 377.
[25] Ibid., 376.
[26] Ibid., 376.

ability to rephrase the same statements made about reasons in the first person, in the third person. Nagel treats this as a serious problem, resulting in a breach of the view that oneself is a person among others equally real. Sturgeon deliberately rejects the commitment of the word 'dissociation' to that doctrine and instead explicitly redefines it as a label for the violation of condition (C).[27]

In Sturgeon's reconstruction, Nagel's next move is to posit that this in turn commits one to a denial of the 'Impersonality Thesis', which Sturgeon posits covers the following claims:[28]

> a) Anything that someone can say about himself, using a first-person sentence, can equally well be said about him by the use of an appropriate third-person sentence;
>
> b) What any first-person sentence says can equally well be said in an impersonal language, and from the impersonal standpoint;[29]

Finally, the denial of the Impersonality Thesis leads one to solipsism, as to deny that truths stated from the first-person perspective can equally well be said from the impersonal standpoint is to believe that there is something irreducibly special about the self and its subjective truths.[30]

This is Sturgeon's rough reconstruction of Nagel's argument. In his criticism, Sturgeon largely focuses on the invalidation of these successive inferences, after which Nagel's argument falls apart. I will not spend much time arguing that my argument avoids these particular criticisms, as it does not follow a particularly similar structure, and the inferences that Sturgeon attacks are ones that my argument largely avoids by omission of the relevant concepts.

For example, I largely neglect the notion of motivational content. Nagel mostly uses this concept to explore the relationship between acceptances of reasons from personal and impersonal standpoints and how this affects the power of reasons to not just be a reason for doing something but a reason *to actually do it*, and question whether the impersonal standpoint should make a difference if one is not a solipsist. I am not so concerned with the psychology behind it, as I feel it brings unnecessary considerations that do not adequately

[27] Ibid., 394.
[28] Ibid., 385.
[29] Nagel,*The Possibility of Altruism*, 100-102.
[30] Sturgeon,. "Altruism, Solipsism and the Objectivity of Reasons", 385-386.

represent the relevant factors when attempting to move from personal to impersonal standpoints regardless of how this distinction is conceived. By supposing that an impersonal observer in control of all actions would promote everyone's interests equally purely out of the absence of any reason not to, and that a self-locating constraint cannot give any non-redundant reason to counteract this, the need to touch on the psychology of motivation and dissociation is lessened.

Psychology is not what is particularly important to either argument, however, and the most crucial difference between the formulation proposed in this paper and Nagel's original argument is his focus on what propositions can be expressed in different contexts. His lack of success in doing this effectively is, I think, a result of an incomplete representation of the facts that can be reasoned about from the impersonal perspective.

Nagel defines the impersonal standpoint as one which "provides a view of the world without giving one's location in it".[31] The similarities to our earlier attempt to imagine things from an impersonal perspective should be immediately obvious. However, the differences are apparent when Sturgeon's most devastating criticism of all is considered. Here, he points out that unstated but essential premises of Nagel's commit himself to denying the Impersonality Thesis.

Sturgeon argues that in order for Nagel's point that subjective reasons lose their motivating content when rephrased from the impersonal standpoint to work, he must be assuming that there are two ways a proposition can have motivating content *for him*:

> a) He recognises it as one of his own self-regarding considerations;

Or

> b) He recognises that either by itself or with other of his beliefs it entails, by some nontrivial argument to which it is essential, one of his own self-regarding considerations.[32]

Without this assumption, there could be a situation where someone decides from the impersonal standpoint that someone has reason to do something, also knows that that person is in fact him, and yet this reason lacks the motivating content that would be present in the first-

[31] Nagel, *The Possibility of Altruism*,101.
[32] Sturgeon,. "Altruism, Solipsism and the Objectivity of Reasons", 395-396.

person formulation. Outside of avoiding this situation, the assumption is generally needed to explain how a proposition can have motivating content.

Additionally, if I have a subjective reason which *lacks* motivating content when phrased as an impersonal consideration, this must be *without* the knowledge, or at least ignoring the knowledge that the person in consideration is in fact 'me'. In order to deny that subjective reasons could have motivating content when formulated as impersonal considerations, Nagel must assume that an impersonal consideration requires one to take an impersonal standpoint, completely ignoring any self-locating information.

Sturgeon points out that these two assumptions, when taken together, entail *no impersonal consideration* can have motivating content. That is because in order to fulfil (a) or (b), one must recognise a consideration as one's own for it to be motivating. However, if impersonal considerations must be viewed from the impersonal standpoint, any self-locating information is barred from use. As such, there is no way to recognise an impersonal consideration as self-regarding, and therefore motivating (by assumption 1), whilst also adopting an impersonal standpoint to properly recognise the statement as impersonal. Without these assumptions, Sturgeon argues, his argument fails to show that belief in subjective principles enforces dissociation.[33]

The contradictions involved in trying to make Nagel's formulation work stem from a deeper failure to accurately represent the impersonal standpoint. A large part of the issue that arises here is the need to demand that certain facts be forgotten, for the impersonal standpoint to be properly adopted. With a full picture of the impersonal standpoint, I do not think any facts need to be forgotten to show that subjective reasons imply solipsism.

Let us take one of Sturgeon's examples. "Someone may believe, and say in an impersonal language, that the runner on first base has reason to try to steal second; and this proposition which he believes and expresses may have motivational content for him, even if he thinks the reason is merely a subjective one, if in addition he knows or believes that *he* is the runner on first base, and hence that he has reason to promote his goal or want it promoted."[34] Sturgeon argues that when Nagel claims that impersonal considerations lack

[33] Sturgeon,. "Altruism, Solipsism and the Objectivity of Reasons", 399.
[34] Sturgeon, . "Altruism, Solipsism and the Objectivity of Reasons", 396-397.

motivational content for a person, he must be assuming that certain personal beliefs like this are absent, or at least ignored; that is, he has adopted the impersonal standpoint.

It is clear where Nagel's impersonal standpoint and its implications differ from our own. Firstly, unlike in Nagel's impersonal standpoint, with our impersonal observer (which, to re-emphasise, is not actually a perspective but a more accurate representation of facts, and the information relevant to decision making) there is no need to *forget* the fact that he is the runner in order for the statement to lose its 'motivating content', as Nagel puts it.[35] Once the impersonal standpoint and an acceptance of its accompanying facts have been adopted properly, the belief that the runner on first base has reason to try to steal second would be eliminated immediately, if it is indeed true in either formulation that this reason is subjective. In Nagel's formulation, that would be because motivating content is lost with an impersonal consideration properly based in the impersonal standpoint; for our formulation, it would be because there is, once all the facts are accounted for, no reason whatsoever to promote the interests of this runner over, say, any other competing athlete.

The key difference, which makes this argument more direct and much less vulnerable to criticism, is Nagel still makes multiple distinctions that my formulation doesn't recognise as coherent or non-redundant within the framework. It is fundamental to Nagel's framework that something is *forgotten* when moving to the impersonal standpoint, and a *new fact* is learned when it is discovered that the man in question is, in fact, me. If the impersonal standpoint is to be viewed correctly, what fact has been forgotten by the impersonal observer? That it is me, person A, who is attempting to think this way? I, in my imagined perspective, would have no problem taking that fact into account. If I want to view the world as accurately as possible, it would not do me well to try to start forgetting things, for these may be valuable factors to consider. I do not need to worry about this fact hindering reason once it has been established that the sense or thought represented by the phrase, 'I am me,' is an entirely redundant piece of information with regard to its significance to reason generation. In trying to view things like the impersonal observer, the fact that I might imagine looking down on Earth, and seeing someone thinking about the impersonal observer, reflecting on his own situation and his sense of self, does not significantly change the things I know to be true about the world. It only becomes problematic when it is thought that

[35] Nagel, *The Possibility of Altruism*, 64-67.

that self-locating information is substantial enough to even hope to provide a change in reasons.

Nagel seems to get close to this idea. He states the shift from the impersonal standpoint to the personal "makes a great difference in *how* that world is conceived, but no difference in what is conceived to be the case."[36] Sturgeon suggests that Nagel attempts to characterise these standpoints predominantly through the language available for describing the world. Sturgeon suggests he should go further than just restricting the terms one is able to use in the impersonal standpoint, and that the restriction should also apply to terms one *knows how* to use. Anyone using the impersonal standpoint should be barred from using any idea, correct or incorrect, of how to express any proposition using the terms of personal language.[37]

This is, of course, dramatic. The personal language is needed to express some very important facts about the world from the perspective of the impersonal observer, who, it must be remembered, knows every fact about the conscious experiences of every individual. Of course, once again, it is not necessary to know these to engage in the perspective, just to hold the belief that other minds exist. The only reason one would need to set restrictions in this way is if self-locating information of any sort is presumed to provide a justified interest for acting in a self-interested way *from the personal perspective*. Having established that it cannot serve this purpose, provided that you, the person imagining the impersonal observer, have a good understanding of this concept, there is no fear that any information relating to oneself could reintroduce subjective reasons, as the logical untenability of that possibility has been accepted.

Nagel was clearly on the right track with his claim that the shift from the impersonal standpoint to the personal makes no difference in "what is conceived to be the case".[38] However, if "what is conceived to be the case" is understood accurately enough to unify the two pictures by including known subjective facts within the total set of facts about the world, it is possible to give due attention to the question of *why exactly* this should not change how it is conceived, where Nagel could not.

[36] Nagel, *The Possibility of Altruism,* 103.
[37] Sturgeon, "Altruism, Solipsism and the Objectivity of Reasons",397-398
[38] Nagel, Thomas. *The Possibility of Altruism,* 103.

§ CONCLUSION

I have presented a revised way to conceptualise Nagel's impersonal standpoint, which prioritised as cohesive a picture of the facts as possible, aiming to treat others' experiences as utterly real without losing sight of oneself as part of that set of facts. It was argued that once every individual's sense of self has been accounted for, self-locating information is largely redundant as it is bound to only trivially reassert a particular, known example of an individual experiencing a sense of self in a manner no different to that included within the existing facts. This renders the concept of subjective reasons redundant, as the illusion of additional self-locating information inherent within one's sense of self does not represent any real, substantive fact. Therefore, they cannot be used as a basis for reason generation, even for the individual themself. Solipsism would, on the other hand, provide a factual, reason-generating basis to prioritise the interests of the self over others, as included within the facts would be the knowledge of the absence, or at least dampened existence, of others' experience. As such, revisiting Nagel's conclusion, there must be only objective reasons if solipsism is to be avoided, as to treat self-locating facts as suitable informational bases for generating reasons would be to underestimate, and thus undermine, the reality of others' experiences as having an equally metaphysically privileged sense of self.

REFERENCES

Nagel, Thomas. The Possibility of Altruism. Oxford: Clarendon Press, 1970.

Nagel, Thomas. The View from Nowhere. New York: Oxford University Press, 1986.

Parfit, Derek. Reasons and Persons, Oxford: Clarendon Press, 1984.

Ridge, Michael. 'Reasons for action: Agent-Neutral vs Agent-Relative' The Stanford Encyclopedia of Philosophy (Spring 2023 Edition).

Sturgeon, Nicholas L. "Altruism, Solipsism and the Objectivity of Reasons", Philosophical Review, 83, No. 3 (1974): 374-402.

AGAINST MORAL DEFERENCE

WHY ARTIFICIAL MORAL AGENTS NEED NOT UNDERMINE PHRONESIS

JINGLONG YANG

§ 1: INTRODUCTION

In recent years, technological advances in artificial intelligence (AI) have accelerated the development of artificial agents. These agents—usually in the form of computer programs or robots—are designed to perform tasks in ways that emulate human action and reasoning. Their applications span a wide spectrum, including autonomous vehicles, humanoid companion robots, and algorithmic trading systems. As a result of rapid technological development, researchers in the field of machine ethics have begun to examine whether artificial agents might engage in moral deliberation or perform ethically significant actions. Drawing on concepts from human ethics and moral psychology, agents capable of ethical behavior are generally referred to as *Artificial Moral Agents* (AMAs).[1]

To be clear, not all ethical machines are AMAs. The taxonomy of ethical machines can be clarified by appealing to John Sullins's account of artificial agents, which distinguishes between *ethical impact agents* (EIAs), *artificial ethical agents* (AEAs), and *artificial moral agents* (AMAs).[2] EIAs are systems that merely generate outcomes with ethical consequences but lack any capacity for moral reasoning. For example, a self-driving car that unintentionally kills a pedestrian operates as an EIA. The next level of ethical agents is the AEAs. They are systems explicitly to incorporate ethical parameters—whether through rule-based architectures or machine learning algorithms. For example, an autonomous vehicle that chooses the least harmful crash option using

[1] Cervantes, et al. "Artificial Moral Agents," 501-532.
[2] Sullins, "Artificial Phronesis," 136-146.

utilitarian calculation would qualify as an AEA. However, AEAs remain morally passive, meaning they do not determine their own ethical standards but instead operate according to standards set by their human designers, who therefore retain ultimate moral responsibility for the agents' actions. Only at the level of AMA do we reach the threshold of systems capable of self-guided moral deliberation and context-sensitive ethical behavior akin to that of human agents.[3]

For the present purpose, this paper defines AMA as any *system that can choose its own ethical standards and act on its own ethical judgment in complex and novel situations without real-time human intervention.* The precise definition of an AMA remains contested, depending on what necessary conditions one takes to be constitutive of AMA creation. Proposed conditions include phenomenal consciousness (or its functional equivalent), rationality and moral competence, free will and autonomy, and moral responsibility.[4] Among these, consciousness seems to be the hardest to satisfy so far.[5]

> Nevertheless, even if consciousness proves to be a necessary criterion for AMA, its realization may not be impossible. According to Jeff Sebo and Robert Long's widely cited analysis of AI moral consideration, there exists a non-negligible probability (≥ 0.1%) that some AI systems will attain some degree of consciousness by 2030.[6]

Assuming such a precautionary stance, a pressing normative question arises: *should* humanity create AMAs if doing so becomes technically feasible?

Responding to this question, Aristotelian philosophers such as Nir Eisikovits and Dan Feldman have expressed concerns. Although their work does not address AMAs explicitly, their critique of advanced AI systems readily extends to them. Eisikovits and Feldman warn that such systems may erode human *phronesis*—our capacity for practical wisdom—by gradually displacing the exercise of moral judgment in everyday life.[7]

In this paper, I will contest Eisikovits and Feldman's pessimistic view and argue that, under appropriate conditions, AMAs can, in fact, facilitate cultivating rather than undermining human *phronesis*.

[3] Sullins, 139-140.
[4] Behdadi & Munthe, "A Normative Approach to Artificial Moral Agency," 195-218
[5] Himma, "Artificial agency, consciousness, and the criteria for moral agency," 19-29
[6] Sebo & Long, "Moral Consideration for AI Systems by 2030," 591-606
[7] Eisikovits & Feldman, "AI and Phronesis," 181-199

Eisikovits and Feldman's warning implicitly assumes that people will always, or at least, frequently enough, defer decision-making to highly intelligent AI systems, such as AMAs, to an extent that is damaging for our phronetic capacities. However, this assumption is far less secure than it appears. To demonstrate my views, I will first draw on Aristotle's account of responsibility in his *Nicomachean Ethics* to examine whether AMAs could satisfy the conditions for bearing responsibility. Eventually, failing to comply with the epistemic condition for responsibility renders AMAs incapable of being held accountable, despite their highly independent and autonomous capabilities. To prove why AMAs should never be trusted to make decisions independently, I will show in detail how they fail to fit within our current accountability mechanisms.

However, this limitation turns out to be beneficial. Precisely because AMAs cannot bear responsibility, humans cannot offload moral judgment onto them. Essentially, their unsuitability for bearing responsibility incentivises us to keep actively engaged in the deliberative process. Such active engagement is the ground for the human-AMA relationship to nourish and provide opportunities that not only preserve but also cultivate our phronetic abilities. In short, the creation of Artificial Moral Agents is not only permissible but desirable, given their potential to enhance human *phronesis*.

§ 2: CONCERNS ABOUT AMA

Despite the numerous threats posed by rapid AI development, many of them could be overcome. Some of the most prominent threats AI poses stem from methodological failures in the training and prediction stages.[8] Algorithmic bias is one of them. As Eisikovits and Feldman observe, AI's bias issues usually mirror and amplify the prejudices of its human designers. This kind of deficiency then contributes to entrenching the existing social and economic inequalities.[9] However, if these methodological defects can be remedied by advances in data curation, auditing, interpretability, and governance, a further question arises: "If algorithmic bias concerns about AI were eliminated, would there be anything left to worry about? To put it more sharply, if AI decisions became fairer than typical human decisions, would there be any residual discomfort with the technology?"

[8] Eisikovits & Feldman, 185
[9] Eisikovits & Feldman, 186

Other philosophers like Cheng-hung Tsai and Hsiu-lin Ku have advanced this question to a more radical level. They invite us to imagine a scenario in which not only algorithmic bias is resolved, but all technical aspects of AI, such as opacity, privacy, and hallucination, are refined to perfection,[10] just like the kind of artificial superintelligence (ASI) described by Nick Bostrom[11]—except in this case, the ASI does not pose existential risk. Then the question becomes: "If AI were to achieve technical perfection, would there be anything left to worry about with the technology?"

Eisikovits and Feldman point out that the worry lies in the deprivation of *phronesis*. Their normative concern can be illuminated through Aristotle's function argument in the *Nicomachean Ethics*:

1. The distinctive function of human beings is the excellent exercise of rationality, in which *phronesis* (practical wisdom) is a crucial part.

2. Sustained exercise of rationality cultivates virtues.

3. The possession and active exercise of virtues is constitutive of *eudaimonia* (happiness).

4. A being flourishes insofar as it actively and excellently performs its distinctive function.

5. Therefore, actively exercising rationality—such as *phronesis*—is a necessary condition for human flourishing and happiness.

If Aristotle is right, the exercise of rationality is not merely a means to flourishing but partly constitutive of it. This is the key of Eisikovits and Feldman's concern: by gradually displacing human decision-making—in domains such as hiring, resource allocation, and loan assessment—AI does not merely make certain tasks more efficient.[12] It deprives us of the very activity through which we become and remain good human beings. The threat, in other words, is not just to our competence but to our humanity.

AMAs may exacerbate such a normative concern. Continuing in the speculative sphere inherited from Tsai and Ku, a superintelligent AMA may be better than humans at making ethical decisions. From a utilitarian perspective, such superintelligent AMAs, capable of making fairer decisions faster and more consistently, are precisely

[10] Tsai & Ku, "Why AI may undermine phronesis," 3079-3086
[11] Bostrom, *Superintelligence: Paths, Dangers, Strategies*
[12] Eisikovits & Feldman, 189

where practical wisdom is most needed. Hence, people risk falling into moral deference with AMAs. If Eisikovits and Feldman are right about AI's undermining effect on human *phronesis*, does this mean we are doomed with the creation of AMAs? Should we prohibit such creations, even if they become technically feasible?

Not necessarily, but only if we scrutinize the assumption on which the threat rests. One key assumption is that people will defer moral decisions to AMAs. Although this assumption may be true to a certain extent, the scale of such moral deference is likely to be limited, given AMAs' inability to bear responsibility.

§ 3: ARISTOTELIAN CONDITIONS OF RESPONSIBILITY & AMA LIMITATIONS

§ 3.1: THE CONTROL CONDITION

Regarding the issue of responsibility, Aristotle offers us some useful insights in his *Nicomachean Ethics.* In Book III, he analyses responsibility in terms of two necessary conditions: the control condition and the epistemic condition. Essentially, an agent is responsible for his action if he has done it knowingly (the epistemic condition) and voluntarily (the control condition). Contemporary Aristotelian scholars often treat these two conditions as the basic criteria for attributing responsibility.[13]

The control condition states that an agent is responsible if the "principle of the action" is up to the agent to perform or refrain from it.[14] Philosophers often interpret this condition in terms of autonomy and freedom.[15] To be responsible, an agent must have adequate control over the deliberative process that leads to action; namely, the action must be voluntary, not forced.[16] When an action is forced, its cause lies outside of the agent's control, which may excuse the agent from taking responsibility. For example, if an AMA's decision-making process is externally overridden by a malicious hack, the "principle of action" would lie outside the system's control; hence, it might be

[13] Talbert, Moral Responsibility (Stanford Encyclopedia of Philosophy)
[14] Aristotle, NE III.1 1110a15–20
[15] Coeckelbergh, "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability," 2051-2068
[16] Aristotle, NE III.1 1110a1–5

excused from responsibility, just as a human acting under coercion might be. Essentially, the control condition places a heavy emphasis on individual control.

In the *Nicomachean Ethics*, Aristotle distinguishes autonomy from two levels: mere voluntariness and choice. He notes that children and even non-human animals act voluntarily insofar as their behavior originates from internal impulses.[17] But such behavior is not, strictly speaking, a choice. It is more of a reaction than an action. A choice, by contrast, involves deliberation about means within our power toward a certain end.[18] Choice requires a rational capacity to select among alternatives by considering how to best achieve some goals. This distinction opens conceptual space for AMAs. If advanced AMAs can evaluate alternative actions based on certain moral standards, they could be said to make choices rather than simply execute commands. The question is whether they could exercise rational control over means and practical ends in particular situations in a way that their "principle of actions" could be said to lie within themselves.

In theory, sufficiently sophisticated AMAs could satisfy this control condition, given enough technological advancements. Recall that an AMA is a *system capable of choosing its own ethical standards and acting on its own ethical judgment in complex and novel situations without real-time human intervention*. If such systems become technically feasible, they would, by definition, possess the deliberative capacities relevant to the control condition. Consider an AMA deployed to coordinate responses during an environmental crisis. Rather than executing a fixed decision rule, it needs to recognize multiple ethically salient considerations—intergenerational fairness, long-term ecological sustainability, and present human welfare—and deliberate about which should predominate. If new information emerges, the AMA should revise its practical priorities accordingly. This capacity to form, evaluate, and endorse practical ends in response to changing circumstances is precisely what distinguishes genuine deliberation from mere computation. In this sense, its "principle of action" lies within the system itself, thereby satisfying Aristotle's control condition.

Critics may object that AMAs cannot satisfy Aristotle's control condition because they cannot act contrary to their final ends (*telos*). At most, they can choose among means to ends that human designers have already fixed for them. In that sense, the AMA's principle

[17] Aristotle, NE III.2 1111b5–15
[18] Aristotle, NE III,2 1111b 5–1112a 15

of action remains external: its governing *telos* is not self-chosen, but imposed from outside.

This objection identifies a genuine limitation, but it mistakes the kind of control Aristotle regards as necessary for responsibility. Under his teleological metaphysics, Aristotle suggests that everything has a *telos* (or final end). Beings generally do not choose their final ends for themselves—including humans. We "wish" for our ends and "choose" among means.[19] Acorns do not choose if an oak tree is worth becoming, they just grow into it. Similarly, we do not deliberate about whether *eudaimonia* is worth pursuing but only how to pursue it. If responsibility required freely choosing one's final end, no human being could be responsible for anything. Aristotle avoids this absurdity by locating responsibility in our control over deliberation and action, not in originating our final ends. Hence, if the absence of freely chosen final ends disqualified an agent from responsibility, then humans and AMAs would fail together.

For this reason, the fact that AMAs do not choose their final ends does not by itself show that they fail Aristotle's control condition. In this respect, the difference between AMAs and human beings is less sharp than the objection suggests. Their final end would be fixed by their designers, just as our final ends may be fixed by our creator, whether that be God or Nature. Therefore, to say that humans fix the final ends of AMAs is only to say that humans may bear unavoidable responsibility for the creation and development of AMAs. It does not, however, follow that AMAs cannot have sufficient control over their own actions. What matters is whether an AMA has enough control over its practical ends and means for its principle of action to lie within itself in the relevant sense.

Although AMAs could not act contrary to their final ends, they may still act contrary to the subordinate practical ends assigned to them by humans. If they were truly intelligent enough, there is no obvious reason to believe that they cannot identify mistakes in a human operator's reasoning and propose alternative goals and corresponding means. Whether they should be given unrestricted authority to implement those plans is a separate question (more on this later). Still, that issue concerns the proper structure of human oversight, not whether the AMA exercises deliberative control. Consider again the environmental AMA. If it genuinely possessed ethical judgment of its own and operated by ethical standards it could apply autonomously, then a human instruction to reduce current carbon emissions might be

[19] Aristotle, NE III.2, 1112b11–17

treated not as a fixed directive but as a revisable practical end. The AMA could instead propose a different subordinate goal—say, investment in clean-energy research—if it judged that strategy better suited to the good of the environment. The key point is that, although AMAs would not choose their final ends, they could still possess sufficient control over their practical ends and means to satisfy Aristotle's control condition.

Another possible objection against AMAs satisfying the control condition is that they lack the ability to refuse tasks, and thus lack the autonomy that genuine control requires. Even if an AMA can propose different practical ends, once human operators insist on pursuing the end as they see fit, the AMA cannot refuse to comply.

It is certainly difficult to imagine why we would grant AMAs complete and unchecked authority over their actions. To do so would be highly risky, since it would amount to releasing an artificial agent without meaningful external constraint. For this reason, some limitation on refusal may be practically unavoidable. But for the same reason, the absence of a refusal capacity may reflect not a conceptual limitation in AMAs themselves, but a practical limit imposed by human designers. If we were ever willing to grant AMAs complete operational autonomy, then such systems could presumably acquire the ability to refuse tasks as well, and, in that respect, satisfy Aristotle's control condition more fully.

Although always being compelled to follow human instructions does weaken autonomy, it does not necessarily disqualify AMAs from satisfying the control condition. The fact that one can be forced does not prove one lacks autonomy altogether, but only that one lacks autonomy under certain situations. We certainly don't say that we, humans, are not voluntary agents simply because we can be coerced. Even without the full power of refusal, AMAs may still possess sufficient deliberative control over their practical ends and means to satisfy the control condition in the relevant sense. For example, they can communicate uncertainty when a situation exceeds their competence, flag potential conflicts between competing values, and propose alternative solutions. These are all meaningful exercises of practical deliberation. An AMA that proceeds blindly despite recognizing its limitations would indeed fail the control condition, but an AMA that does think critically and disagrees with its human operators is a different case. If after an AMA has told us its concern, explained why what we are asking it to do is bad or wrong, and proposed alternative solutions, it is still required to carry out the original instruction, then the resulting action may be

better understood as forced rather than fully voluntary. In this case, the AMA would be excused from responsibility, just as a human agent under coercion would be. After all, it would not choose that course of action if the decision were left to its own deliberation.

It is important to emphasize that this paper's focus is normative rather than technical. I am not claiming that current AI systems possess these capacities, nor am I speculating about when such capacities might be realized. The question under investigation is this: if technological development makes AMAs possible, should we create them? From the control condition alone, we can say that, at least, the technical advancements required for such deliberative capacities are not impossible. These advancements require nothing categorically different from incremental technological development. They do not demand, for instance, that a non-sentient system acquire consciousness. The control condition, as I have interpreted it, concerns only rational deliberation over means and practical ends—capacities that fall within the domain of sophisticated information processing rather than phenomenal experience.

However, the control condition alone does not qualify AMAs from bearing responsibility; they still face serious difficulties meeting Aristotle's epistemic condition.

§ 3.2: THE EPISTEMIC CONDITION

The epistemic condition is another foundational criterion that Aristotelian scholars recognize for attributing responsibility.[20] This condition requires the agent not be ignorant of what he is doing and of the relevant particulars of the situation.[21] More precisely, Aristotle distinguishes several important ways in which one can be ignorant:

> An agent acts involuntarily if he is ignorant of one of these particulars. They are: (1) who is doing it; (2) what he is doing; (3) about what or to what he is doing it; (4) sometimes also what he is doing with it—with the instrument, for example; (5) for what result—safety, for example; (6) in what way—gently or hard, for example."[22]

Ignorance of these particulars renders an action involuntary, thus not an appropriate target for responsibility. Rudy-Hiller translates this knowledge requirement to a bundle of awareness: to be responsible,

[20] Fischer & Ravizza, *Responsibility and Control*, 13
[21] Aristotle, NE III.1 1111a1–20
[22] Aristotle, NE III.1, 1111a3–5

an agent must be aware of the action itself, the instruments used, the consequences, and (according to some) the alternative options available.[23] Aristotle emphasizes that awareness of an action's consequences is the most crucial component.[24] Arguably, moral significance is an important consequence. Contemporary Aristotelian scholars usually call the capacity to discern what matters morally in a context-sensitive way *moral perception*.[25] [26]

To possess such perception, one must have an adequate understanding of virtue, which requires an appropriately formed emotional capacity. According to Aristotle, virtue is made of two components: virtue of knowledge and virtue of character.[27] Virtue of knowledge enables an agent to understand what ought to be done, while virtue of character ensures that the agent is affectively attuned to doing it in the right way.[28] On this view, correct moral perception not only requires correct judgment but also being educated to take pleasure and pain in the right things.[29]

These feelings do not merely accompany moral perception; they partly constitute it. A properly formed capacity for fear, shame, pity, pleasure, and pain allows an agent to register what is morally significant in a situation. We recognize threats as serious partly because we are capable of fearing what is genuinely fearful; we grasp another's suffering as morally salient partly because we are capable of being pained by it; and we become sensitive to moral failure partly because we are capable of shame. Without such affective states, an agent may still register the descriptive features of a situation, yet fail to grasp their ethical significance. In that sense, one who lacks the relevant feelings and emotions cannot fully perceive what moral perception requires.

Given this moral perception requirement, AMAs are unlikely to satisfy the epistemic condition. Unlike the control condition, which concerns capacities achievable through incremental technical progress, the epistemic condition involves something more fundamental. Until researchers solve the "hard problem of consciousness"[30] and discover

[23] Rudy-Hiller, The Epistemic Condition for Moral Responsibility (Stanford Encyclopedia of Philosophy)
[24] Aristotle, NE III.1, 1111a15–20
[25] McDowell, "Virtue and Reason," 331-350
[26] Liu, "Creating Character," 533-549
[27] Aristotle, NE II.1, 1103b15-19
[28] Aristotle, NE II.6, 1106b16-28
[29] Aristotle, NE II.3, 1104b3-1105a16
[30] Chalmers, "The Hard Problem of Consciousness"

how to instantiate phenomenal experience in artificial systems, AMAs will lack the qualitative mental states—fear, compassion, regret, empathy—necessary for genuine moral perception. An important consequence of this inadequacy is that AMAs do not experience pleasure and pain. But, according to Aristotle, repeated feedback of pleasure and pain grounds the habituation process that makes moral learning and character building possible.[31] Without lived experience of what feels good and what feels base, an agent cannot develop the emotional attunement that moral perception requires.

One might object that AMAs can simulate moral learning through reinforcement mechanisms, adjusting reward parameters in ways structurally analogous to how humans build character through habituation. However, this analogy is misleading. Human habituation is lived and felt; we become virtuous by reliably taking pleasure in virtuous actions and experiencing discomfort or shame when acting basely. Over time, these affective responses become second nature. A utility function might mimic the behavioral pattern of habituation, but it lacks the phenomenology that gives moral perception its depth. The difference is not merely quantitative but categorical. For example, consider a firefighter AMA deployed in disaster response. Such a system may assign high reward values to "saving human lives", thereby choosing actions that maximize survival. But the AMA has never lived. It doesn't understand what it feels like to have a family waiting at home, what it means to love someone, what is precious about being alive, or what is tragic about a preventable death. Its internal "awareness" is exhausted by data representations and optimization functions. This makes the AMA more akin to calculators that output "2" when given "1+1" than to human firefighters who risk their lives from a felt sense of duty and compassion. Such an AMA is not a moral agent exercising *phronesis* but simply a moral instrument extending the agency of its human operators.

Thus, AMAs are not fit to bear moral responsibility. They may perform morally correct outputs, but cannot genuinely understand why those outputs are morally significant. AMAs are, therefore, best understood not as responsible moral agents but as sophisticated moral instruments. Failing the epistemic condition disqualifies AMAs from being appropriate targets of praise or blame, thereby undermining any argument for entrusting them with independent moral decision-making. If AMAs cannot be responsible for their actions, then humans have strong reasons to retain moral authority rather than defer to these

[31] Aristotle, NE II.1, 1103a14-1103b25

systems. This observation will later prove to be significant for the question of whether AMAs threaten or facilitate human *phronesis*. But first, let us talk about why exactly humans have strong reasons to retain moral authority rather than defer to AMAs.

§ 4: PRACTICAL LIMITATION OF AMA DEVELOPMENT

§ 4.1: AMA LIMITATIONS IN ATTRIBUTION PRACTICES

Once the epistemic condition has failed, the basis for attributing responsibility is already severely weakened. This weakness becomes even clearer when we consider the practical aspects of our responsibility practices themselves. My key point is this: any attempt to attribute responsibility to AMAs will prove normatively thin, because it is difficult to see how AMAs could be appropriate bearers of responsibility in any sense meaningful to us.

In ordinary practice, responsibility attribution usually serves two aims: punishing the wrongdoer and providing redress for the victim. Accountability practices—punishment, blame, social condemnation—are morally significant because they impose a cost on the offender in proportion to their fault. Doing so acknowledges the victim's suffering and, to some extent, offsets that suffering. Punishments like imprisonment, fines, and public censure work precisely because they are experienced as painful or burdensome, damaging the offender's material circumstances, social standing, or sense of self. Similarly, redress is not merely a matter of compensation. Part of what victims seek is the recognition that the one who wronged them has been made answerable for that wrong. These practices presuppose that the responsible party can be meaningfully affected, both to himself and to those he has harmed.

That is precisely where AMAs fall short. Similarly to why they fail to pass the epistemic condition of responsibility, AMAs lack the kind of conscious moral understanding that would make praise or blame intelligible. Kenneth Himma has argued that a genuine moral agent must be an appropriate target of praise and blame.[32] This

[32] Himma, "Artificial agency, consciousness, and the criteria for moral agency," 19-29

requirement is not satisfied by the mere capacity to modify behavior in response to external inputs. It presupposes the ability to experience guilt, shame, suffering, or some analogous form of normative burden. Without phenomenal consciousness, AMAs do not experience pain as suffering, imprisonment as deprivation, social condemnation as a blow to their moral standing or self-conception. As a result, our ordinary accountability practices lose their normative force when directed at AMAs. What remains are mere technical interventions—shutting down the system, resetting its parameters, altering its utility functions. Such measures may be instrumentally useful for preventing future harm, but they do not amount to holding the AMA responsible in any morally substantial sense. Therefore, they are normatively thin.

Consequently, such technical interventions cannot provide meaningful redress for victims. When a human wrongs another, accountability practices communicate something morally important: that the wrong is recognized, that the wrongdoer is answerable, and that the victim's grievance has moral weight. None of these is adequately communicated to the victim by rebooting a server or adjusting a reward function. By analogy, "punishing" an AMA is comparable to kicking a rock that has rolled downhill and injured someone. The rock is insensible to the punishment; kicking it neither compensates the victim nor acknowledges the harm properly.

One might object that we already attribute responsibility to some non-conscious entities. Corporations, for instance, can be fined, sued, and publicly condemned. However, corporate responsibility is ultimately parasitic on human responsibility. When we hold a corporation accountable, we typically assume that the burdens of liability fall on human beings. The corporation serves as a juridical structure for assigning responsibility, but the normative force of accountability still traces back to persons capable of bearing its consequences. AMAs, by contrast, are not institutions composed of conscious members; they are technical systems with no conscious constituency to bear the burden of blame on their behalf.

This observation has led some ethicists to conclude that artificial agents should never be granted the status of responsible agents in the first place. Joanna Bryson, for instance, argues that robots "should be slaves"—that is, they should be treated as tools for which humans retain ultimate responsibility.[33] Her point is not that robots deserve slavery treatment, but that framing them as responsible

[33] Bryson, "Robots Should Be Slaves"

agents dissolves human accountability into technical artifacts that are structurally ill-suited to bear the moral weight we wish to place on them. As Andreas Matthias famously pointed out, if we pretend that advanced AI systems can be responsible, we risk creating responsibility gaps: situations in which serious harms occur but no one is genuinely answerable for them, because responsibility has been offloaded onto systems incapable of bearing it.[34]

Thus, AMAs may be important objects of regulation, but they are not fitting targets for moral responsibility. They can be managed when malfunctioning, but they cannot be blamed in any sense that satisfies the normative aims of our responsibility practices.

The fact that AMAs cannot bear responsibility has significant implications for how much trust we should place in them. If we were to allow AMAs to make ethical decisions independently, we would be entrusting beings incapable of accountability with matters of serious moral consequence. The bar for such trust, I believe, is extraordinarily high. It requires something approaching perfect reliability (say, an accuracy rate of about 99.99%).

The necessity for such stringent reliability becomes apparent when we recognize that complex ethical situations rarely involve a single decision. Typically, they require multiple layers of judgment: assessing the situation, identifying relevant considerations, weighing different alternatives, and adjusting as circumstances unfold. Even a highly reliable AMA will see its trustworthiness erode across successive decisions. Suppose an AMA achieves 95% accuracy on any single ethical judgment—an impressive figure by current standards. After five layers of decision-making, its cumulative reliability drops to approximately 77%. After ten layers, it falls below 60%. For an AMA to remain trustworthy across extended deliberative chains, it would need an accuracy rate approaching 99.99%—a threshold that may never be attainable in principle. Thus, even setting aside the question of responsibility, there are strong epistemic reasons for maintaining human authority over significant moral decisions.

[34] Matthias, "The Responsibility Gap," 175-183

§ 4.2: APPROPRIATE HUMAN-AMA COLLABORATION STRUCTURE

However, we should not disregard AMAs' highly capable capacities. If they can reliably perform morally correct actions, we should leverage this utilitarian advantage by having them as moral advisors, which can facilitate the cultivation of *phronesis* rather than undermine it.

Cheng-hung Tsai and Hsiu-lin Ku have already developed a promising response to Eisikovits and Feldman's concern that AI may undermine human *phronesis*. They propose a strategy called the *Principle of Epistemic Heed*, which holds that we should exercise our rational capacity as much as possible while paying careful attention to any superintelligent system's opinions. They emphasize that the crucial distinction is between heeding and deferring. To defer is to relinquish one's judgment in favor of another's; to heed is to attend carefully to another's opinion while retaining autonomous control over the decision-making process.[35] On this view, the proper relation between humans and highly intelligent systems is one in which such systems serve as advisors while we retain the final say. Thus, humans can benefit from an AMA's epistemic capabilities while still exercising their own practical judgment.

I wish to advance Tsai and Ku's position further. Their argument shows how epistemic heed can *preserve* opportunities for exercising phronesis; I contend that collaboration with AMAs can actually *cultivate* phronesis in ways that would not otherwise be available. There are two considerations to support this stronger claim.

The first is that highly capable AMAs can automate lower-order cognitive tasks that, while necessary, are often tedious and taxing. This automation frees us to focus on higher-order moral reasoning. For example, consider a hospital ethics committee evaluating organ allocation. An AMA could compile and synthesize medical data, flag inconsistencies in patient records, calculate survival probabilities, and identify applicable precedents from similar cases—tasks that are cognitively demanding but procedurally routine. By handling this informational groundwork, the AMA allows the committee to concentrate on harder questions: How should we weigh quality of life against length of life? What considerations of fairness apply when patients have unequal access to post-operative care? These are the kind

[35] Tsai & Ku, "Why AI May Undermine Phronesis," 3083

of questions that highly demand *phronesis*, and offloading lower-order tasks creates more cognitive space for engaging with them.

Second, having AMAs as moral advisors provides learning opportunities that enhance our judgment capacity. When a discrepancy arises between human and AMA judgment, we are compelled to engage closely with the AMA's reasoning, because, as I have argued, we remain the ultimate responsibility bearers. This kind of engagement typically requires higher-order thinking: challenging assumptions, weighing competing considerations, and arriving at a reflective judgment about which course of action is correct. Such deliberative work is difficult, but it is proportionally more rewarding for developing *phronesis*. If the AMA's recommendation proves to be correct, we gain an opportunity to reflect and refine our reasoning. Conversely, if we identify an error in AMA's reasoning, we have successfully engaged with a sophisticated deliberative process and exercised precisely the critical judgment that *phronesis* requires. In either case, so long as we heed rather than defer, we stand to benefit from the collaboration.

§ 4.3: REMAINING CONCERNS

Despite the advantages AMAs could bring, they also face serious drawbacks. Some of the major drawbacks can be best understood through Albert Borgmann's discussion of "focal practices". For Borgmann, focal practices are settled patterns of engagement with things that gather meaning and orient our lives—activities like running, gardening, or preparing a meal.[36] These practices cultivate human excellence precisely because they require sustained effort, skill, and attention; they unite means and ends, labor and enjoyment, in ways that develop our capacities. Borgmann warns that technological conveniences tend to threaten this unity by disburdening us of effort. Applied to AMAs, the concern is that automating aspects of ethical deliberation might similarly disburden us of the very engagement through which *phronesis* develops. If moral reasoning becomes something we receive rather than enact, we risk becoming passive consumers of ethical outputs rather than active practitioners of practical wisdom.

The first objection concerns passivity, illustrated by Borgmann's "easy chair problem." Borgmann observes that we often

[36] Borgmann, "Focal Things and Practices," 9-13

act against our better judgments because technology makes passivity easier.[37] For instance, when we come back home tired, we know a walk could refresh us, yet we may still choose to stay in a cozy chair with the television and a beer. Similarly with AMAs, we might know that genuine moral engagement is valuable, yet passively accept AMA recommendations out of convenience. Such passivity is indeed a problem, but may not be detrimental. Unlike the easy chair scenario—where the cost of passivity is low—moral decisions often carry significant stakes. Given that we remain the ultimate responsibility bearers, we have strong incentives to remain active in the decision-making process, thereby preserving the conditions under which *phronesis* can develop.

Another related objection concerns habituation. This objection holds that *phronesis* requires continuous habituation through repeated practice, not merely occasional engagement whenever discrepancies arise between human and AMA judgments. On this view, the "lower-order" tasks I propose to automate may themselves be essential to developing the stable dispositions that constitute practical wisdom. I acknowledge that there is genuine value in repeated practice of foundational tasks, especially their nurturing effect for higher-order reasoning. However, the concern is not unavoidable. Regarding children who are still developing their deliberative capacities, they should not be permitted to grow overly reliant on AMAs. Repeated practices have proven to be highly beneficial for children's learning in many aspects. Although the supervision required may be difficult to achieve, it is nevertheless addressable through education and regulation. As for adults with already-developed capacities, we can trust that they will remain fairly engaged with AMAs' reasoning, since, again, they bear ultimate responsibility for the resulting actions and decisions. Moreover, whenever adults feel that they need to sharpen their lower-order skills, they always have the option to do the work themselves and resume those focal practices.

The third objection concerns the opacity of AMA's reasoning. This is not an objection springing from Borgmann's discussion of focal practices; nevertheless, it is a strong challenge worth considering. The concern is this: if an AMA's deliberative process is not entirely transparent, can we engage with it meaningfully? I acknowledge that advanced AI systems may contain some opacity in their computational processes. However, they are not lacking in verbal deliberation ability. Current large language models are already capable of engaging in

[37] Borgmann, 14

conversation and explaining their reasoning when prompted. There is no reason to believe AMAs could not be paired with similar language abilities. As long as they can provide verbal explanations, there should be enough transparency for us to examine their thought processes. The goal is not to eliminate every opacity in the system, but simply to clarify aspects relevant to our decision-making context.

While concerns regarding passivity, habituation, and opacity are relevant, none of them proves fatal to my claim that AMA collaboration can cultivate human *phronesis*. The bottom line is this: as long as we remain the ultimate responsibility bearers, we are compelled to engage actively with AMAs rather than defer passively.

§ 5: CONCLUSION

The creation of Artificial Moral Agents is not destined to erode human *phronesis*. On the contrary, appropriately structured collaboration with AMAs can cultivate *phronesis*—provided that we understand the proper relationship between human and artificial moral agents.

That relationship is shaped by a fundamental asymmetry. AMAs, however sophisticated their deliberative capacities, do not satisfy Aristotle's epistemic condition for moral responsibility. Without phenomenal consciousness, they cannot experience the emotions that constitute moral perception. Hence, they should remain as moral instruments rather than moral agents. However, this apparent limitation turns out to be advantageous. Precisely because AMAs cannot bear responsibility, humans cannot offload moral judgment onto them. Any attempt to attribute responsibility to AMAs proves to be normatively thin. At the end of the day, we remain accountable for ethical decisions, and such accountability compels us to engage actively in the decision-making process.

This insight allowed me to advance beyond Tsai and Ku's Principle of Epistemic Heed. They showed that attending to AI suggestions while retaining autonomous judgment can preserve opportunities for exercising *phronesis*. I have argued for something stronger: that such collaboration can cultivate *phronesis* through two mechanisms—by freeing cognitive resources for higher-order reasoning, and by providing learning opportunities when human and AMA judgments diverge. In both cases, the collaboration demands

that we exercise precisely the capacities that practical wisdom requires: weighing competing alternatives, interrogating assumptions, and arriving at reflective judgments under conditions of genuine responsibility.

As for objections concerning passivity, habituation, and opacity, they are genuine but not fatal. After all, the stakes of moral decisions and our position as responsibility bearers incentivize our active engagement; developmental concerns can be addressed through education and regulation; and the opacity of AMA reasoning can be mitigated through their verbal deliberation capabilities.

I believe these findings have practical significance as AI systems increasingly enter domains requiring ethical judgment. The question is not whether to admit them into our moral lives but how to structure that collaboration wisely. This paper suggests that the answer lies in accountability: by ensuring that humans remain answerable for decisions made with AMA assistance, we create the conditions under which practical wisdom can flourish rather than atrophy. Essentially, the same feature that disqualifies AMAs from moral agency becomes the safeguard that preserves our own.

REFERENCES

Behdadi, Dorna, and Christian Munthe. 2020. "A Normative Approach to Artificial Moral Agency." *Minds and Machines* 30:195-218. https://doi.org/10.1007/s11023-020-09525-8.

Borgmann, Albert. 1984. "Focal Things and Practices." In *Technology and the Character of Contemporary Life: A Philosophical Inquiry*, 1-16. N.p.: University of Chicago Press.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. N.p.: Oxford University Press.

Bryson, Joanna. 2010. "Robots should be slaves." In *Natural Language Processing*. https://doi.org/10.1075/nlp.8.11bry.

Cervantes, Salvador, José A. Cervantes, Sonia López, Luis F. Rodríguez, Francisco Cervantes, and Félix Ramos. 2020. "Artifcial Moral Agents: A Survey of the Current Status." *Science and Engineering Ethics* 26:501-532. https://doi.org/10.1007/s11948-019-00151-x.

Chalmers, David. 2017. "The Hard Problem of Consciousness." In *The Blackwell Companion to Consciousness*, edited by Susan Schneider and Max Velmans. N.p.: John Wiley & Sons, Incorporated. https://doi.org/10.1002/9781119132363.ch3.

Coeckelbergh, Mark. 2020. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26:2051-2068. https://doi.org/10.1007/s11948-019-00146-8.

Eisikovits, Nir, and Dan Feldman. 2022. "AI and Phronesis." *Moral Philosophy and Politics* 9 (2): 181-199. https://doi.org/10.1515/mopp-2021-0026.

Fine, Gail. 1996. *Aristotle: Introductory Readings*. Translated by Terence Irwin and Gail Fine. N.p.: Hackett Publishing Company, Incorporated.

Fischer, John M., and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. N.p.: Cambridge University Press.

Himma, Kenneth E. 2009. "Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology* 11:19-29. DOI 10.1007/s10676-008-9167-5.

Liu, Wei. 2012. "Creating Character: Aristotle on Habituation, the Cognitive Power of Emotion, and the Role of Prudence." *Frontiers of Philosophy in China* 7 (4): 533-549.

Matthias, Andreas. 2004. "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics and Information Technology* 6:175-183.

McDowell, John. 1979. "Virtue and Reason." *The Monist* 62, no. 3 (July): 331-350. https://doi.org/10.5840/monist197962319.

Rudy-Hiller, Fernando. 2018. "The Epistemic Condition for Moral Responsibility (Stanford Encyclopedia of Philosophy)." Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/moral-responsibility-epistemic/.

Sebo, Jeff, and Robert Long. 2023. "Moral Consideration for AI Systems by 2030." *AI and Ethics* 5:591-606. https://doi.org/10.1007/s43681-023-00379-1.

Sullins, John. 2021. "Artificial Phronesis: What It Is and What It Is Not." In *Science, Technology, and Virtues: Contemporary Perspectives*, edited by Emanuele Ratti and Thomas A. Stapleford, 136-146. N.p.: Oxford University Press.

Talbert, Matthew. 2019. "Moral Responsibility (Stanford Encyclopedia of Philosophy)." Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/moral-responsibility/.

Tsai, Cheng-hung, and Hsiu-lin Ku. 2025. "Why AI may undermine phronesis and what to do about it." *AI and Ethics* 5:3079-3086. https://doi.org/10.1007/s43681-024-00617-0.

SEEING REALITY IN LIGHT OF LOVE
AN ANALYSIS OF MURDOCHIAN LOVE

HAZEL QING ZHAO

§ 1: INTRODUCTION

In *The Sovereignty of Good*, Murdoch proposes that love be rendered as a central concept in morality. True love, she repeatedly claims, involves attending to the *reality* of one's beloved. This connection between love and reality is illustrated in the famous example of the mother-in-law (M) and daughter-in-law (D). M initially holds an unfavourable view of D, seeing her as "vulgar," "rude," and "juvenile," but later realises that this negative perception was driven primarily by her own jealousy.[1] Consequently, M reassesses D's personality, which eventually allows her to see D in a positive light: D is "not vulgar but refreshingly simple," "not undignified but spontaneous," "not juvenile but delightfully youthful."[2] Crucially, M's reappraisal of D, now purged of jealousy, is said to be a more realistic view enabled by love: "When M is just and loving she sees D as she really is."[3] It is a "…loving gaze directed upon" D's "individual reality," a gaze that uncovers the truth of D's character.[4]

However, things are not quite simple. Whilst the paradigmatic view in *The Sovereignty of Good* seems to be that "Love" and "Real" are "closely connected" concepts,[5] the ambiguity involved in Murdoch's

1 Iris Murdoch, *The Sovereignty of Good* (New York: Routledge, 2014), 17.
2 Murdoch, 17.
3 Murdoch, 36.
4 Murdoch, 33.
5 Murdoch, 41.

characterization of love has prompted various interpretations. In this paper, I begin by unpacking this ambiguity, showing that Murdoch leaves the relationship between love and reality unclear. Then, I propose that this ambiguity allows for two possible kinds of interpretations of her view on love: the objective view, and the positive light view. Finally, I argue that the positive light view offers a better interpretation of Murdoch's position, concluding that love, while rooted in reality, also exceeds it and transcends it.

§ 2: THE AMBIGUOUS RELATIONSHIP BETWEEN LOVE AND REALITY

To illustrate in more detail the relationship between love and reality, I will start with a closer analysis of the mother-in-law example. Consider M's distortion of D's reality by inappropriately describing her as "vulgar," "rude," and "juvenile." This distortion was motivated by jealousy, which is, crucially, an emotion derived from the human *ego*. "The fat, relentless ego," ever-present in human consciousness, seeks naturally to protect its high status, causing one to appropriate another person's reality to fit with one's own wishes.[6] Indeed, given that jealousy is incompatible with moral righteousness, M's distortion of D's reality is depicted as an instance of seeking to protect her ego by retaining a deluded conception of herself as a morally acceptable person. Given that the ego is always alert to the threats against its own importance, it is not hard to see that it blocks the way of love, which requires other-centeredness. Love comes forward when self-importance is subdued, giving way to a humble recognition of one's own flaws - "I may be prejudiced," "I am certainly jealous"[7] - M admits to herself, subsequently altering her view to "see D" more "loving(ly)."[8] In love, M no longer appropriates D's reality to elevate her ego. Love takes M out of selfish concerns that once motivated a distortion of who D is, enabling M to learn about D's independent, individual reality. The process through which M comes to suppress her ego for a clearer vision of D is termed "unselfing."[9] Love emerges only in an unselfed state.

In short, love enables M to see D "as she really is."[10] M's perception of D is now made more accurate and clear. As Murdoch

6 Murdoch, 51.
7 Murdoch, 17.
8 Murdoch, 36.
9 Murdoch, 82.
10 Murdoch, 36.

puts it, a "refined and honest perception of what is really the case" is the outcome of unselfing.[11] Genuine love for another person is the background condition that enables unselfing, and hence, enables the improved perception that is more truthful. If we accept the analysis presented so far, we should conclude that love *grounds*, and, is a *precondition* for a clear perception of reality. However, through a closer reading, I will show that the relationship between love and reality is a lot more ambiguous. This is due to Murdoch's somewhat unsystematic characterization of love. At times, rather than expressing that love *grounds* the perception of reality, she argues that love itself *is* the perception of reality:

- "Love is knowledge of the individual… Love… is the discovery of reality."[12]

- "The direction of attention is… outward, away from self… towards the great surprising variety of the world, and the ability to direct attention is love."[13]

At other points, Murdoch seems to see love as neither identical to the perception of reality, nor a precondition that grounds it. Rather, love is characterised as involving but *exceeding* a mere perception of reality. Whilst love and the vision of the real are both in place, love yields a new power that enriches what we see with emotional depths:

- "What M is *ex hypothesi* attempting to do is not just to see D accurately but to see her justly or lovingly."[14]

- "…what looks like mere accuracy at one end looks more like… love at the other."[15]

In engagement with Murdoch's works, it is not uncommon for readers to feel that she "does not present systematic, clear arguments", and to feel "somewhat mystified" as a result.[16] This ambiguity could be conceived of as a pitfall in Murdoch's writing, yet I suggest that simultaneously, it opens room for critical reflection upon the relationship between love and the perception of reality. In what follows, I will begin by outlining the two dominant interpretations that seek to clarify this relationship: the objective view and the positive light view. Then, I argue that the positive light view offers a better account of love.

[11] Murdoch, 37.
[12] Murdoch, 27.
[13] Murdoch, 65.
[14] Murdoch, 22.
[15] Murdoch, 87.
[16] Heather Widdows, *The moral vision of Iris Murdoch* (Aldershot, England; Burlington, VT: Ashgate, 2005), 11.

§ 3. LOVE: THE OBJECTIVE VIEW V.S. THE POSITIVE LIGHT VIEW

On the objective view, to love someone is to confront the full reality of who they are in a detailed and objective manner. Wolf endorses the objective view as an interpretation of Murdoch, proposing that true love requires loving "unreservedly despite the full awareness of faults in the objects of… [one's]… love."[17] The objective view can accommodate two ways of understanding love. First, it is compatible with the claim that love itself *is* the perception of reality. The person capable of careful, objective attention to the details and nuances of reality embodies love. This is to say that when M manifests love by virtue of the fact that she attends to the reality of who D objectively is, she sees D clearly and accurately, and to see D this way *is* to love D. Alternatively, the objective view is compatible with viewing love as a *precondition* that *grounds* the perception of reality; that is to say, attention to reality can only happen when love is present. There cannot "be such a thing as careful attention — which aims to see its object clearly, accurately, objectively — in the absence of love."[18] Or, as Mason writes, "… love has an irreducible epistemic role… reality is… perceptible only to the person who lovingly attends to it."[19] Love sets up the background within which objective vision of reality takes place.

The positive light view of love stands in contrast to the objective view. In this view, love involves attending to another person but does not necessitate full confrontation with the *entirety* of their objective, present reality. Whilst a sufficient level of objectivity is still in place to enable genuine receptivity towards the beloved's reality, it allows that the lover's vision of their beloved is overall more positive than strictly objective, seeing their beloved as promising more than they currently are. This involves, as Driver puts, "[overlooking] at least minor flaws" in the object of one's love.[20]

[17] Susan Wolf, "Loving Attention: Lessons in Love from The Philadelphia Story," in *Understanding Love: Philosophy, Film, and Fiction*, ed. Susan Wolf and Christopher Grau (Oxford: Oxford Academic, 2014), 169, https://doi.org/10.1093/acprof:oso/9780195384512.003.0017.

[18] Wolf, 174.

[19] Cathy Mason, "Iris Murdoch and the Epistemic Significance of Love," in *New Philosophical Essays on Love and Loving*, ed. Simon Cushing (Cham, Switzerland: Palgrave Macmillan, 2021), 10.

[20] Julia Driver, "Love and Unselfing in Iris Murdoch," *Royal Institute of Philosophy Supplement* 87 (2020): 177.

Whilst we might debate about what kind of flaws count as "minor" (which will set a limit on the permissiveness of a loving gaze), as far as some flaws are overlooked, the small element of idealisation involved here nevertheless means that love is neither identical to nor a mere precondition to the perception of the real. The positive light view, whilst certainly not blind to reality, is compatible with the view that love involves but *exceeds* reality.

It should be emphasised that both the objective view and the positive light view understand love as requiring attending to the objective, present reality of the beloved in some way. However, they indeed demand different *kinds* of attention to the reality in question. The objective view of love involves attending to reality accurately, objectively, and in a detailed manner; love is either this detailed perception itself or a necessary means to achieve detailed perception. One sees relatively accurately their beloved's strengths and weaknesses, and the amount of joy or pain they experience. With an objective understanding of their beloved, they love their beloved as they are. The positive light view of love, by contrast, is not as rigorous about objectivity or accuracy *per se*. It still requires one to be attentive enough towards the beloved's experience to "share their contexts,"[21] which enables seeing a comprehensive picture of them. However, the picture seen here is less 'as it is' than the picture seen from the objective view — it speaks to a more general, universalising impression, while leaving room for appropriate imaginary content regarding future potentials. Think of the helpful example given by Driver, where parents might be moved by love to see their child as a more talented artist than he objectively is at present.[22] For, rather than dismissing the child's doodles as some incomprehensible abstractions, love enables the parents to pick up subtle signs, and see in the child the potential for artistic growth. Similarly, we have reasons to think that a mother can be said to love her child even prior to the child's birth, even though she has not yet attained much knowledge of what the child is like.[23]

At first glance, we may be inclined to say that the positive light view digresses from the paradigmatic view in SOG. After all, one could argue that in claiming M sees D "not just accurately but lovingly," the contrast here between love and the real nevertheless supposes that M sees D *both* lovingly and realistically. As Robjant

[21] Driver, 172.
[22] Driver, 177.
[23] Christopher Cordner, "Love," in *The Murdochian Mind*, ed. Silvia Panizza and Mark Hopwood (London: Routledge, 2022), 175.

puts it, "M before D confronts a reality."[24] The liberation from the ego is conceived of as the path towards an increasingly clear, true, and accurate vision of the current state of the world, things, and other people. The positive light view, however, regards love as presenting an overriding force in informing M's vision of D, allowing her to attain a renewed interpretation that recognises the positive potential in D - pleasantly "simple," "spontaneous," and "youthful." If we accept this, it follows that it may be hard to tell whether M's loving vision of D is the most accurate and realistic reflection of who D currently is. Perhaps "impartial observers" who feel no particular love towards D may not share M's vision of D; they might, instead, "agree with M's earlier vision, seeing D as vulgar, juvenile, and so on."[25] Why should we think that love must be something *more than* this perception of the real?

§ 4: JUSTIFICATION FOR THE POSITIVE LIGHT VIEW

Before I present my analysis, I shall briefly reformulate the three ways of understanding Murdoch's characterization of the relationship between love and accurate perception of reality:

(1) Love itself *is* the accurate perception of the reality of the beloved;

(2) Love *grounds*, and, is a *precondition* for the perception of the reality of the beloved;

(3) Love involves but *exceeds* the perception of the reality of the beloved.

Perhaps one is tempted to say that, in SOG, Murdoch overall leans towards the objective view of love - "Real" and "Love" are "closely connected."[26] The objective view presents an ideal picture which grants both love and accurate perception of the present state of the beloved. This view is compatible with both (1) and (2). Upon further

[24] David Robjant, "Is Iris Murdoch a Closet Existentialist? Some Trouble with Vision, Choice and Exegesis," *European Journal of Philosophy* 21, no. 3 (2013): 483.

[25] Paul Katsafanas, "Depth, Articulacy, and the Ego: Murdoch on Moral Vision," in *Iris Murdoch's The Sovereignty of Good*, ed. Carla Bagnoli and Bradford Cokelet, Anniversaries Series (Cambridge: Cambridge University Press, forthcoming).

[26] Murdoch, *Sovereignty of Good*, 41.

examination, however, I argue that (1) lacks plausibility. Namely, love cannot be wholly identical with accurate vision - it needs to be something more. As Cordner argues, mere accuracy of vision is insufficient for love. Accurate vision without the right orientation can easily corrupt into a cold, "acute eye for others' weaknesses", where the "aim" is to "exploit them" for one's own gain.[27] It is, therefore, hard to see why accuracy and objectivity *per se* entail love.

However, didn't Murdoch, through the mother and daughter-in-law example, demonstrate that accurate vision is attainable only to those who are *already* unselfed (i.e. taken out of self-concern)? Shouldn't this lead us to think that a selfish and manipulative person cannot even achieve accuracy of vision in the first place? I argue we have reasons not to think so. It is vital to note that unselfing as a criterion to enable accurate perception "is mostly a negative one."[28] As Murdoch herself concedes, the ego is inherent to and ever-present in the human consciousness; its forces are impossible to fully counteract.[29] As long as we remain alive, the ego will continue posing selfish forces, yielding jealousy, resentment, and pride that blots out our vision of the real. Understood this way, we could conclude that one's vision can lean more towards being unselfed than egocentric, but not that it can be entirely unselfed. As such, what is shown here is merely that unselfing is good for improving the accuracy of vision. Since absolutely unselfed, accurate vision is already a rejected possibility in Murdoch's understanding, it follows that all vision can only attain relative accuracy, and there is no reason to think that relatively accurate vision is only attainable by someone who has unselfed. With this in mind, we are now in a position to believe that selfish people who aim to exploit others are just as capable of a relatively accurate vision of another person's reality. After all, an objective perception of others' motives, fears and weaknesses is essential for deploying them for one's own purpose. Selfish people might even see accurately the strengths and goodness in another person, yet remain unmoved by what they see. (1) cannot be the most plausible, because accurate, truthful vision itself does not necessarily involve a loving attitude.

(2) can be ruled out on similar grounds. Namely, given that there can be unloving forms of accurate perception, it means that accurate

[27] Christopher Cordner, "Lessons of Murdochian Attention," *Sophia* 55, no. 2 (2016): 201.

[28] Silvia Caprioglio Panizza, "Perception, Self, and Zen: On Iris Murdoch and the Taming of Simone Weil," *Philosophies* 8, no. 4 (2023): 3, https://doi.org/10.3390/philosophies8040064.

[29] Murdoch, *Sovereignty of Good*, 53.

perception does not necessarily require love as a precondition. Since both (1) and (2) have been rejected, the objective view has not found justification. Love, then, is more plausibly understood as involving but exceeding the clear perception of reality. The objection to (2) shows us that love cannot ground an accurate perception of one's beloved. However, this does not threaten the fact that love could still involve accurate vision. Love certainly inspires in us a curiosity towards, and a desire for knowledge of, our beloved, which in turn leads us to become more attentive to their reality. As Cordner puts it, love involves an openness and receptiveness towards one's beloved.[30] It is natural for one to, through love, gain clearer knowledge of what their beloved is like, and be moved to attend to their needs. Accurate vision does not entail love, but love nevertheless motivates one to see *increasingly* the reality of another person: the lover discerns the subtle sadness of his beloved when no one else notices; the sister knows that her younger brother is struggling even though he did not ask for help.

Love, then, is at least compatible with seeing a version of our beloved that is increasingly real. But the sense in which it is real is quite nuanced, because we still have good reasons to question how objective and accurate this love-inspired vision can be. Namely, given that the "unself" - the state in which love emerges - is an ongoing process towards an ideal that is never fully achievable, the extent of accuracy will remain limited. And, we could take our skepticism even further. We might think that love does not really result in more accurate or objective vision. In a sense, it is hard to pinpoint whether a loving vision is accurate enough, given that we, as human beings, exist as independent entities who will never fully understand each other. Indeed, Piety argues that in the mother-in-law example, the only source of objective knowledge has been D's unaltered behaviour.[31] Even when M became more unselfed and attentive, she was still attaching *interpretations* to D's behaviour which cannot count strictly as objective knowledge. Consequently, M's reappraisal of D serves to reveal not something about D, but something about M herself — "that is, that she was loving."[32] Moran expresses similar views. He thinks that M's later favourable description of D merely shows a change in M's personal orientation rather than an increase in M's accuracy of

[30] Cordner, "Lessons of Murdochian Attention," 211.
[31] Marilyn Gaye Piety, "Kierkegaard and Murdoch on Knowledge of the Good," in *Why Kierkegaard Matters: A Festschrift in Honor of Robert L. Perkins*, ed. Robert L. Perkins, Mary Amanda Jolley, and Edward L. Rowell (Macon: Mercer University Press, 2010), 210-11.
[32] Piety, 213.

vision.[33]Both Piety and Moran think that whilst M's loving vision of D is unselfed, attentive, and receptive towards D's reality, there can be no certainty as to whether the reality of D perceived by M is accurate and objective.

Given these doubts, should we altogether reject the view that love can lead to a clearer perception of the beloved's reality? I suggest we need not. Building on Cordner's account of love as a receptivity towards the beloved, I propose that, through love, one is receptive towards not just a factual reality, but also a felt, phenomenal reality. When we love someone, they become more *real* to us in that they become more *significant* to us. This perceived significance diminishes our sense of self-importance, and moves us to desire what is good for our beloved for their own sake. The beloved is made more real to us, for, in our loving gaze, their needs appear with more urgency, their joy and sorrow more noticeable, and their entire being precious and irreplaceable. This is what I mean by a receptivity towards the felt reality of our beloved. Seeing the beloved in this way does not require a complete adherence to absolute factual knowledge about the beloved. For, as previously mentioned, loving vision is compatible with overlooking minor flaws, and recognising potential for betterment that is yet to be actualised. It is precisely this type of love that adequately explains cases where parents lovingly discern the artistic talents of their child, or care about their newborn tenderly, without yet knowing what kind of person they will become. As such, in viewing our beloved as significant in this manner, our love is rooted in reality yet exceeds and transcends it, because we see through them to the inherent preciousness and value they instantiate, where this preciousness is not tied wholly to factual knowledge.

This, I argue, is the way the positive light view is to be justified. Loving vision involves but surpasses the real, because the experience of love is fundamentally *personal*. As Murdoch herself proposes, the reality of any object or person is "infinitely to be learned, as an individual object of love… a painter might say, 'you don't know what 'red' means'," [34] because she has developed a unique vision of red that no one else can fully share. Or again, "the movement of understanding is onward into increasing privacy… and not back towards a genesis in the rulings of an impersonal public language." [35] In one's vision of reality, the element of the personal is inevitable. A love-inspired vision

[33] Richard Moran, "Iris Murdoch and Existentialism," in *The Philosophical Imagination* (New York: Oxford University Press, 2017), 190.
[34] Murdoch, *Sovereignty of Good*, 29.
[35] Murdoch, *Sovereignty of Good*, 28.

is a particular vision rather than a universal one. As Cordner beautifully puts it: "it is *John's* smile she loves… not 'a' smile that… happens to be attached to John… she finds John *presenced-in* his smile."[36] Thus, we are not seeing more clearly in the sense of gathering ever-more-detailed and accurate information about our beloved. Instead, love opens a space for the real by making the felt reality of our beloved more significant to us. Simultaneously, the profoundly personal nature of love takes us to see the beloved as embodying an inherent goodness that reaches deeper than the apparent qualities they possess. What is primary here is a knowledge 'of' the ontological importance of the beloved and not a knowledge 'about' them, although the latter may well be a spillover effect of the former, because love motivates a curiosity in the lover to learn more about the beloved. This is the way in which love involves, but also exceeds, the perception of the real. To put it more precisely, love involves perceiving a certain level of the beloved's objective reality, but exceeds it by entering a more phenomenal, personal realm where their inherent preciousness comes to light.

However, a further objection may be raised here. Specifically, the arguments above express that, while love is directed towards a beloved who exists with a factual, objective reality, the lover's vision is based more in a phenomenal, felt reality of the beloved. If this is true, the aforementioned skeptical challenge appears unresolved. For, given that the phenomenal reality takes precedence, one might worry that this love does not involve sufficient objectivity at all. Basing our love upon certain current facts about the beloved seems required to warrant that it is towards them, rather than someone else, that we direct our love. If no sufficient objectivity is involved, we will be loving a mere illusion of the beloved that deviates from their actuality, which would also count as an egotistical distortion that Murdoch herself criticises. If the positive light view of love is to be better justified, then, it is not enough that one has access to the phenomenal reality of the beloved, referring to mere potentials within or the overall significance of them. Rather, it must be demonstrated that an appropriately objective foundation remains established in the background, because objectivity is crucial for warranting that the beloved individual is the true object of our love.

In defence of my account, I argue that the objective foundation required is still in place even if we accept the positive light view. Namely, the worry is that perceiving a felt reality might miss the kind of objectivity to ensure that our love is directed at a real, particular

[36] Cordner, "Love," 175.

person rather than an illusion or fantasy. But I think a loving vision that perceives an overall significance and positive potential still has a firm basis in the objective reality, although such objectivity is limited due to various factors, such as the impossibility of fully unselfing, and the general difficulty in deriving accurate interpretations of other people who are not ourselves. To illustrate this point more clearly, recall the examples I have provided above. When the parents see promise in the child's scribbles, they are inferring the child's talents based on a concrete piece of drawing. There is no illusion in the sense that the parents see a talent with no basis in objective reality. It is just that love informs their vision to make the talent more noticeable.

The same applies to my example of perceiving the feelings of the beloved. Before I proceed with this, an understanding of Blum's account of responsiveness will be useful. Blum comments that, in Murdoch's view, "knowing" another person involves a responsiveness, where one allows "the perception of another person's distress to affect oneself."[37] However, he follows this up by noting that "experiencing the *same* feeling as the other person is neither necessary nor sufficient for responsiveness."[38] It is not sufficient, because having the same feeling does not entail altruistic concerns for the suffering person. Nor is it necessary, because one can have deep and loving care towards the suffering other without sharing the same emotional experiences. When I argue that the beloved's suffering becomes more urgent or noticeable in the lover's view, I mean it not in an illusory sense where the lover has made an epistemic error. The lover is still responding to the beloved's actual states of suffering at a particular point, which is grounded in objective reality. What I propose is that, because having the exact *same* painful feelings as our beloved is unlikely, love may heighten our perception of their painful feelings as more urgent on a phenomenal level. A perceived urgency is still far from an epistemic error. And the outcome of this perception is still warranted, because it is the loving responsiveness to the beloved's actual states of suffering.

§ 5: CONCLUSION

Overall, my view is that a loving vision is layered, with objectivity present within but not exhaustive of it. One might think that the positive light view I defend departs from Murdoch's paradigmatic

[37] Lawrence A. Blum, *Moral Perception and Particularity* (Cambridge: Cambridge University Press, 1994), 190, emphasis added.
[38] Blum, 190.

view, which is that love enables us to see the beloved reality as it is. However, given that Murdoch thinks that love in a state of absolute "unself" is an ideal end point that is humanly impossible, and that only such a state allows for a true perception of reality, it follows that love can, at best, perceive what is relatively real. And this, perhaps, is why Murdoch also refrains from endorsing a universalised, singular vision. When she writes that "the movement of understanding is onward into increasing privacy… and not back towards a genesis in the rulings of an impersonal public language,"[39] she acknowledges the necessary subjectivity inherent in our vision. It could be argued, therefore, that the positive light view may be more consistent with Murdoch's philosophical framework than it initially appears.

In ordinary life, we would think that loving someone does not guarantee an accurate and objective vision of them. There is, of course, no harm in loving someone whilst seeing them accurately, fully and in a detailed manner. Yet, my arguments thus far should offer at least some reasons to think that love need not satisfy this condition to count as good love. Before we end, it may be illuminating to draw on Murdoch's novel to offer final support for the positive light view. While Murdoch did stress that "her novels should not be seen as… theoretical arguments for a position," she does acknowledge that "the novel is itself an ethical form," which should warrant reference to it for further insights into moral matters.[40]

In Murdoch's novel *The Black Prince*, the protagonist, Bradley, falls in love with a girl named Julian.[41] While his love for Julian has made him more attentive and less selfish, it has also brought him to perceive something more in Julian that is not entirely obvious to other people. The ordinary and oftentimes immature young girl that everyone takes Julian to be becomes so wonderful and precious in Bradley's loving gaze. Although this vision may not be free from some level of idealisation, it is undoubtedly loving. In fact, if we take actions to be indications of love, we could even argue that Bradley loves Julian too much. As Nussbaum notes, the intensity of Bradley's love led him to a courageous sacrifice "clearly for Julian's sake" and "extreme in its consequences"[42] - he conceals the fact that Julian's mother was the true murderer of her father by taking the blame himself, which results in his jail sentence.

[39] Murdoch, *Sovereignty of Good*, 28.
[40] Martha C. Nussbaum, "'Faint with Secret Knowledge': Love and Vision in Murdoch's The Black Prince," *Poetics Today* 25, no. 4 (Winter 2004): 692.
[41] Iris Murdoch, *The Black Prince* (New York: Penguin, 2003).
[42] Nussbaum, "Faint with Secret Knowledge," 703.

The novel, as I interpret it, supports the positive light view that love arises from objective reality, but transcends it at the level of subjective experience. It enables us to perceive the beloved's inherent value as a unique individual, a value that lies beyond mere factual knowledge. Even though Bradley appears to miss certain factual details about Julian, the intensity of his love is evident through his responsiveness towards her. Through love, Bradley achieves a heightened awareness of Julian's needs and feelings. This is evident in his decision to conceal the truth of the murder, which was driven by his belief that revealing it would cause Julian significant emotional pain. Both the acknowledgement of Julian's inherent value and the responsiveness towards Julian's potential agony point towards the positive light view of love. The time when Julian questions that Bradley does not see her clearly,[43] what she has in mind is perhaps more akin to a complete factual understanding of her personality. But love takes precedence over such facts. Beyond an infinite collection of facts, Bradley loves Julian because she *is*.

[43] Murdoch, *The Black Prince*, 265.

REFERENCES

Blum, Lawrence A. *Moral perception and particularity*. Cambridge: Cambridge University Press, 1994.

Cordner, Christopher. "Lessons of Murdochian Attention." *Sophia* 55, no. 2 (2016): 197-213.

Cordner, Christopher. "Love." In *The Murdochian Mind*, edited by Silvia Caprioglio Panizza and Mark Hopwood, 169-82. London: Routledge, 2022.

Driver, Julia. "Love and Unselfing in Iris Murdoch." *Royal Institute of Philosophy Supplement* 87 (2020): 169-180.

Katsafanas, Paul. "Depth, articulacy, and the ego: Murdoch on moral vision." In *Iris Murdoch's Sovereignty of Good*, edited by Carla Bagnoli and Brad Cokelet. Anniversaries Series. Cambridge: Cambridge University Press, forthcoming 2025.

Mason, Cathy. "Iris Murdoch and the Epistemic Significance of Love." In *New Philosophical Essays on Love and Loving*, edited by Simon Cushing, 39-62. Cham: Palgrave-Macmillan, 2021.

Moran, Richard. "Iris Murdoch and Existentialism." In *The Philosophical Imagination*, 185-204. New York: Oxford University Press, 2017.

Murdoch, Iris. *The Black Prince*. New York: Penguin, 2003.

Murdoch, Iris. *The Sovereignty of Good*. New York: Routledge, 2014.

Nussbaum, Martha C. " 'Faint with Secret Knowledge': Love and Vision in Murdoch's *The Black Prince*." *Poetics today* 25, no. 4 (2004): 689-710.

Panizza, Silvia Caprioglio. "Perception, Self, and Zen: On Iris Murdoch and the Taming of Simone Weil." *Philosophies* 8, no. 4 (2023). https://doi.org/10.3390/philosophies8040064.

Piety, Marilyn Gaye. "Kierkegaard and Murdoch on Knowledge of the Good." In *Why Kierkegaard Matters: A Festschrift in Honor of Robert L. Perkins*, edited by Robert L. Perkins, Mary Amanda Jolley, and Edward L. Rowell, 204–14. Macon, GA: Mercer University Press, 2010

Robjant, David. "Is Iris Murdoch a Closet Existentialist? Some Trouble with Vision, Choice and Exegesis." *European Journal of Philosophy* 21, no. 3 (2013): 475–94.

Widdows, Heather. *The Moral Vision of Iris Murdoch*. Aldershot, England: Ashgate, 2005.

Wolf, Susan. "Loving Attention: Lessons in Love from The Philadelphia Story." In *Understanding Love: Philosophy, Film, and Fiction*, edited by Susan Wolf and Christopher Grau. Oxford: Oxford University Press, 2014. https://doi.org/10.1093/acprof:oso/9780195384512.003.0017.

WESLEY STONE
Macalester College

Wesley Stone is a senior at Macalester College studying Philosophy and Data Science. His philosophical interests center on ethics, including population ethics and metaethics. Outside of philosophy, he closely follows new developments in economics, politics and AI. After graduation, he plans to pursue an MA in Economics at Duke University, where he expects his philosophical background to provide a useful perspective.

JUNZE CHEN
Emory University

Junze Chen is currently a junior studying Philosophy and German Studies at Emory University. His primary philosophical interests are early modern philosophy (esp. Spinoza), German Idealism, and critical theory. He is also interested in Ancient Greek philosophy and wishes to further explore its reception and reinterpretation in thinkers such as Schelling, Hölderlin, and Heidegger.

WILLIAM THOMAS
Trinity College, Cambridge

William Thomas is an undergraduate studying Philosophy at Trinity College, Cambridge. He is primarily interested in the philosophy of physics, particularly statistical mechanics. He is also interested in metaphysics and formal epistemology, especially self-locating belief and probability, in relation to paradoxes and the evaluation of cosmological theories. Outside of philosophy, he enjoys rugby, karate, and general idling. After university, he hopes to pursue postgraduate studies in Philosophy.

JINGLONG (BENJAMIN) YANG
Vanderbilt University

Jinglong Yang is a fourth-year undergraduate student at Vanderbilt University, working at the intersection of ethics, moral philosophy, metaphysics, and the philosophy of AI. His research is currently focused on examining whether AI systems could genuinely possess normative competence—the ability to recognize and act on the practical reasons that apply to one's action—and what such a possibility would mean for trust, agency, and moral responsibility. In an age of rapid technological advancement, philosophy is urgently called for to guide and check the increasing power of science. His research is an endeavor in this direction. His broader philosophical interests include Spinoza, Aristotle, Iris Murdoch, and Stoicism.

HAZEL QING ZHAO
University of Warwick

Hazel Qing Zhao is a recent graduate (‘25) of the University of Warwick in the UK, where she studied Philosophy with a focus on continental thought, ethics, literature and religion. Iris Murdoch and Simone Weil have had major influences on her work, particularly their mystical notion of attention to the individual. To further her research, Hazel will begin graduate studies at the University of Chicago Divinity School this fall. In her spare time, she enjoys quiet walks in nature and improvising on the piano.

Dianoia fosters open philosophical discussion and writing among undergraduate students. The journal is committed to providing the opportunity for intellectual reflection that bridges the academic disciplines in pursuit of holistic understanding.

The ancient Greek word ‘διάνοια’ translates as ‘thought’ but connotes a discursive direction or search for a higher form of knowledge. We thus hope to direct philosophical discourse toward an intellectually rigorous level of thinking.

Dianoia is sponsored by

The Institute for the Liberal Arts and the
Philosophy Department of Boston College


COLLEGIUM BOSTONIENSE
FUND. MDCCCLXIII