FRAGMENTS

McGill Undergraduate Journal of Philosophy

Volume 34

2020

Editorial Notes

Contents

Womanhood

Anna Yedrikova

Abstract: Womanhood, as well as women's unique phenomenological experiences, have been long-neglected throughout the history of Western philosophy, a gap that more modern theorists have sought to fill. One of the crucial differences in men and women's experiences lies within the concept of transcendence; while philosophy traditionally holds the man to be the subject, women are objectified by society and treated less like autonomous agents, and therefore tied more closely to immanence. This paper examines specifically Merleau-Ponty's theories about the situatedness of the body, and how both Simone de Beauvoir and Iris Marion Young address the oversights in his work when it is applied to females; Simone de Beauvoir posits that a woman's subjugation is socially constructed, not innate, and Young builds off of her conclusions to claim that women's limited mobility and self-actualization is not the result of biological distinctions between the sexes, but of socialization into objectification.

The issue of womanhood has been a notorious blindspot throughout the history of philosophy, and existentialism is no exception; a field dominated by the intellectual contributions of European men, it has been molded by the perspectives and implicit biases of said men. Simone de Beauvoir and Iris Marion Young, in *The Second Sex* and 'Throwing Like a Girl', attempt to address these oversights, and especially address how women, reduced to immanence by society at large, feel about their own transcendence. A central

point in both authors' works is the idea of woman as object, rather than a fully-formed subject capable of decisive physical and intellectual action, and how this affects women's self-perception and behavior as a result.

The axis de Beauvoir's introduction to The Second Sex revolves around is the idea of woman as Other, while man is default. "The relation of the two sexes is not that of two electrical poles: the man represents both the positive and the neuter to such an extent that in French *hommes* designates human beings, the particular meaning of the word *vir* being assimilated into the general meaning of the word 'homo' ".[1] A concept that Young will later address in 'Throwing Like a Girl', de Beauvoir similarly describes how the male body is seen as the default and the female body as an aberration, citing philosophers such as Aristotle, who considered the female a defective male. To de Beauvoir, women are constantly denied full personhood; they are objects to men's subjects, the Other they define themselves by, and therefore cannot exist except in opposition to them. "He is the Subject; he is the Absolute. She is the Other".[2] She goes on to point out that this is a fairly universal tendency in human beings, that all cultures and religions, for example, have defined themselves as an in-group that necessarily excludes the Other, but women are unique in that their position as such is not subject to reciprocation. "No subject spontaneously posits itself and at once as the inessential from the outset, it is not the Other who, defining itself as Other, defines the One; the other is posited as Other by the One positing itself as One".[3] Women cannot conceptualize men as Other in the same way men view them as Other, because of their dehumanization at their hands.

Beauvoir has several suggestions for the root of women's inescapable conception of themselves as Other. Instead of there being a specific instance of conquering and subsequent oppression, "as far back as history can be traced, [women] have always been subordinate to men, their dependence is not the consequence of an event or a becoming, it did not *happen*".[4] In addition, unlike other marginalized groups, women have no shared culture or history; they "live dispersed amongst men, tied by homes, work, economic interests, and social conditions to certain men — fathers or husbands — more closely than to

[1] de Beauvoir, *The Second Sex*, 5.
[2] Ibid., 6.
[3] Ibid., 7
[4] Ibid., 8.

other women".[5] Because the separation of the sexes is such an inherent, biological distinction, women's liberation is harder to achieve than, say, that of conquered ethnicities, as they are the only group to live in such intimate proximity to their oppressors; nor are they capable of separating fully from men, or exterminating them. Without any past freedom to fall back on, or sense of identification with other women, it is more difficult to make progress.

Apart from biology, de Beauvoir also clearly identifies another factor that makes women such a profound Other — the development of a world where they are systemically treated as such. "Lawmakers, priests, philosophers, writers, and scholars have gone to great lengths to prove that women's subordinate condition was willed in heaven and profitable on earth".[6] Women are considered the inferior sex, and therefore denied many civil rights and privileges granted to men; as a result of these conditions, they are allegedly stunted in the same way as the American black, living up to the expectations set for them. Another crucial distinction is that women are chained to immanence, kept inside the home and expected to devote themselves to housework and childrearing, removed from the public and intellectual spheres and cut off from developing a sense of transcendence: "what singularly defines the situation of woman is that being, like all humans, an autonomous freedom, she discovers and chooses herself in a world where men force her to assume herself as Other: an attempt is made to freeze her as an object and doom her to immanence, since her transcendence will be forever be transcended by another essential and sovereign consciousness".[7] The woman clearly perceives herself as a subject, and must reconcile that with a world that is determined to make her into an object, which leads into the struggles with fully-formed female physicality Iris Marion Young's work covers.

Iris Marion Young's seminal paper 'Throwing Like a Girl' addresses an age-old stereotype; young girls, compared to young boys, consistently show far less skill at physical tasks such as throwing a ball. While boys will put their entire bodies into it, extending the arm and twisting from the hip, a girl will usually only throw from the elbow, leading to less force behind the toss. There are some physiological reasons for this distinction — men as a whole

[5]de Beauvoir *The Second Sex.*, 8.
[6]Ibid., 11.
[7]Ibid., 17

tend to have more testosterone and muscle mass than women, giving them greater strength — but Young suggests that another factor also significantly plays into this phenomenon, socialization.

Young uses de Beauvoir's concept of a woman's situatedness in order to frame her theories about women's bodily movements: "every human existence is defined by its situation; the particular existence of the female person is no less defined by the historical, cultural, social, and economic limits of her situation".[8] She specifically, when defining womanhood as a subject of analysis, rejects the idea of the ineffable 'feminine essence', and instead uses de Beauvoir's framework; "in accordance with Beauvoir's understanding, I take 'femininity' to designate not a mysterious quality or essence that all women have by virtue of their being biologically female. It is, rather, a set of structures and conditions that delimit the typical situation of being a woman in a particular society, as well as the typical way in which this situation is lived by the women themselves".[9] However, while she does consider herself indebted to de Beauvoir's rejection of spiritual essentialism, and her analysis of how environment molds women, she still seeks to address a gap in her work, that of the physicality of the female body, "by largely ignoring the situatedness of the woman's actual bodily movement and orientation to its surroundings and world, Beauvoir tends to create the impression that it is woman's anatomy and physiology as such that at least in part determine her unfree status".[10]

Young leans on Merleau-Ponty's conceptualizations of the body and intentionality in order to bolster her points about the intentional crippling of women's bodily confidence. In Merleau-Ponty's *Phenomenology of Perception*, unlike other theorists, he does not consider the root of subjectivity to be within the mind or consciousness, but instead within the body: "the body is the first locus of intentionality, as pure presence to the world and openness upon its possibilities".[11] As the body orients itself, it expresses that intentionality through action. However, the female body only has an ambiguous transcendence, "a transcendence that is laden with immanence", because of their lack of true intentional movement and trust in their body's capabilities.[12]

[8]Young, 'Throwing Like a Girl', 3.
[9]Ibid., 5
[10]Ibid., 3
[11]Ibid., 9.
[12]Young, 'Throwing Like a Girl', 10

Another key theory of Merleau-Ponty's is intentionality in motility: "the possibilities that are opened up in the world depend on the mode and limits of the bodily 'I can' ".[13] Women are also prevented, thanks to their socialization, from totally expressing this, due to their distrust in their abilities to perform physical tasks and utilize their bodies to their full extent. "Feminine bodily existence is an inhibited intentionality, which simultaneously reaches toward a projected end with an 'I can' and withholds its full bodily commitment to that end in a self-imposed 'I cannot' ".[14] While uninhibited intentionality requires a connection between intent and action, women frequently do not fully carry out the tasks they set out to perform, due to their hesitancy. There is a gap between the physical action and a woman's trust in her own ability to fulfill it, causing the phenomenon of inhibition.

Merleau-Ponty's last concept is that of unity in motion: "By projecting an aim towards which it moves, the body brings unity to and unites itself with its surroundings; through the vectors of its projected possibilities it sets things in relation to one another and to itself".[15] But the female body is characterized by its discontinuous motion instead; they tend to only use one part of their bodies to accomplish a physical task, and "the part of the body that is transcending toward an aim is in relative disunity from those that remain immobile".[16] This is where we reach Young's primary point, building off both of Merleau-Ponty and de Beauvoir. "According to Merleau-Ponty, for the body to exist as a transcendent presence, to the world and the immediate enactment of intentions, it cannot exist as an object. As subject, the body is referred not onto itself, but onto the world's possibilities".[17] Woman, however, is not capable of fully conceptualizing herself as a subject, and therefore expressing intentionality towards an object, when she is framed as an object in the wider world and sees herself as the object of motion, not as a subject. Through the lens of this objectification, we can unite all three obstacles to women's transcendence under a common cause.

Young also examines Merleau-Ponty's concept of phenomenal and objective spaces: "feminine existence lives space as *enclosed* or confining, as having a

[13]Ibid., 10.
[14]Ibid., 10
[15]Ibid., 11.
[16]Ibid., 12
[17]Ibid., 12

dual structure, and the woman experiences herself as *positioned* in space".[18]. Women have been traditionally consigned to indoor spaces, within the home; not only that, but they are taught not to take up too much room with their bodies, and instead keep their limbs close to themselves. As well, "in feminine existence there is a double spatiality, as the space of the here is distinct from the space of the yonder".[19] Unlike in Merleau-Ponty's descriptions, where bodily movements link here and yonder, women are trained to perceive the yonder as a space they are not allowed to access.

Lastly, there is a distinction between how the male body and the female body are positioned in space "because the body as lived is not an object, it cannot be said to exist in space as water is *in* the glass".[20] But the female body, which is far more inhibited and hesitant in its motions, cannot be described as such, "to the extent, that is, that feminine bodily existence is self-referred and thus lives itself as an object, the feminine body does exist in space".[21] Kept within the home, discouraged from seeing itself as capable of affecting motion or trusted to carry out physical tasks, the female body is more object than subject as described in Merleau-Ponty's works. Without being allowed the physical freedom of men, women cannot accurately fit into Merleau-Ponty's model.

Both de Beauvoir and Young, in their works, seek to fill a gap that has long plagued philosophy — the idea of the male as the default — and point out the flaws in this approach. Beauvoir, in *The Second Sex*, explores the concept of the woman both as an Other, and as the object to the male subject, an aberration from the natural male form. Young builds off of de Beauvoir's work in her paper, using her concept of situatedness to move on to her critique of Merleau-Ponty. While Merleau-Ponty's observations may be accurate for the male subject, which he assumes to be the only relevant one, Young painstakingly points out that women's socialization leads them to see themselves as far less capable of enacting physical change than men, even as objects rather than subjects, and therefore prevents their bodies from being the locus of consciousness like the male body is. These theorists portray a missing perspective– that of the woman, as a worthwhile topic of discussion and a lens the world can be viewed through.

[18]Young, 'Throwing Like a Girl', 13
[19]Ibid., 14.
[20]Ibid., 15
[21]Ibid., 15.

Bibliography

de Beauvoir, Simone. *Selections from the Second Sex*, C. Borde and S. Malovany-Chevallier (Trans.). Vintage Books, 2011.

Young, Iris Marion. "Throwing Like a Girl". *On Female Body Experience*, Oxford University Press, 2005.

Marx and the Downward Spiral of Capitalism

John Jakob Etter

Abstract: This paper uses a Marxist framework to discuss the relationship between the nature of capitalism, the wealth gap, and modern economic systems such as widespread credit networks and investments. Credit and investments are presented as a means of production and as commodities in Marxist terminology and are related to the nature of capitalism and the wealth gap in terms of their exploitative nature. Ultimately, this paper argues for an understanding of credit networks and investing as entrenching the wealth gap, which is intrinsic to the nature of capitalism.

Introduction — Capitalism, Exploitation, and Class

The nature of capitalism, the relationship of capitalism to exploitation and income inequality, and how governments should regulate the market are all questions that are still being debated today. Many have turned to Karl Marx for answers to these questions. However, when Marx wrote the first volume of *Capital* in 1867, he was living in a very different world than we experience today. Individuals are even further removed from the products of their labour than they were in Marx's time. Modern economic systems such as extensive networks of personal credit and widespread investment (i.e. stock trading and venture capitalism) have expanded exponentially over the

last few decades. This expansion of capitalism to a historically unprecedented scale has resulted in huge wealth concentration disparities between the top 1% and the middle and working classes, referred to from here on as the wealth gap. For example, in America, three men (Jeff Bezos, Bill Gates, and Warren Buffet) have more wealth than the lower 50% of Americans combined ($350 billion vs. $250 billion).[1] However, whether this is an inevitable result of modern capitalism, or rather only one of many possible outcomes, is not agreed upon. Furthermore, as Ivan Ascher notes, "the very phenomenology of capital seems to have changed [from Marx's time]".[2] Is Marx's terminology of use-value, exchange-value, and commodities still relevant to these questions? I argue yes.

In this paper, I propose an understanding of credit networks and investments in a Marxist framework, in order to argue that these modern economics systems entrench the wealth gap by increasing the exploitation of workers, and that an extreme wealth gap is an inevitable consequence of capitalism. In the conclusion, I explore how this understanding of modern capitalism serves as a ground zero for discussion about the responsibility of the state in a capitalist society.

Modern economic systems in a Marxist framework

First, extensive networks of credit and lending, examples of which include personal credit cards, payday loans, and student loans, can be understood in a Marxist framework as a means of production, because money itself has become a means of production. The possession of money confers lending or crediting power, and thus can be used to create surplus-value for the capitalist in the form of interest. One could immediately object to this understanding by noting that according to Marx, surplus-value cannot appear simply from the circulation of money.[3] Instead, there are specific required conditions to create surplus-value, namely, the addition of labour to the means of production. The creation of surplus-value by 'money-money'[4] circulation

[1] Chuck Collins & Josh Hoxie, *Billionaire Bonanza: Inherited Wealth Dynasties in the 21st-Century U.S.*, Institute for Policy Studies (2018), https://inequality.org/great-divide/billionaire-bonanza-2018-inherited-wealth-dynasties-in-the-21st-century-u-s/.

[2] Ivan Ascher, *Portfolio Society* (New Your City: Zone Books, 2016), 35.

[3] Karl Marx, *Capital*, trans. Ben Fowkes, Volume I ((1867) 1976), 268.

[4] 'Money-money' refers to the circulation of money in the economy where profit is created without transiting through the intermediate form of a commodity (which would

is simply usurer's capital and is "inexplicable from the standpoint of the exchange of commodities".[5] At first glance, these statements seem to invalidate credit as a means of production, because the profit from crediting and lending comes out of usury. However, upon deeper examination, credit does fulfill the requirement of creating surplus-value through the addition of labour to the means of production, because it hijacks the labour of the borrower. This occurs because the money that the borrower uses to pay interest charges comes from their labour, and thus has the quality of containing labour. Credit companies could tell borrowers that they owe a certain amount of interest, but if the borrower was not labouring for a wage, their labour couldn't be hijacked, and they wouldn't be able to pay the interest charges. Therefore, there would be no surplus value created for the capitalist (the one who possesses the money, crediting power, and means of production) without the input of labour. As such, credit and lending fulfill the requirement of their surplus-value originating from labour-power and can be viewed as a means of production. Understanding money as a means of production also relates to entrenching the wealth gap, because the more money you have, the easier it is to produce surplus-value.

While credit takes the form of a means of production, investments such as stocks and bonds can be understood in a Marxist framework as commodities.[6] Marx defines a commodity as "an external object, a thing which through its qualities satisfies human needs of whatever kind".[7] Investments fit this definition, because they are external, both to humans and to money (i.e. separated from both), and they fulfill human needs of many kinds. For example, government bonds allow the financing of military expenditure, municipal bonds allow the creation of roads and public works, and personal stocks provide leverage over the direction of the company.[8] Furthermore, investments meet these criteria because the need fulfilled by commodities is wide-ranging and can "arise [from]... the imagination" or can fulfill needs "directly as a means of subsistence... or indirectly as a means of production".[9]

---

be 'money-commodity-money'). Marx, *Capital* Volume I, 249, 267.

[5]Marx, *Capital* Volume I, 267.

[6]Ascher, *Portfolio Society*, 35.

[7]Marx, *Capital* Volume I, 125.

[8]Ascher, *Portfolio Society*, 35.

[9]Marx, *Capital* Volume I, 125.

The use-value of investments is their usefulness in fulfilling needs, as previously described. However, commodities must also have exchange-value, and this is where one might initially object to viewing investments as commodities, because the medium of equivalence that renders commodities exchangeable must be the quality of containing of human labour.[10] The exchange-value of a commodity corresponds to the amount of "congealed labour time", or the amount of socially average labour time needed to produce the commodity.[11] However, a consideration of the exchange-value for investments reveals that they do indeed fulfill this requirement. Ascher proposes the equivalence for investments as grounded in amount of risk that they possess; investments with low risk and high returns have higher exchange-value than an investment with high risk and high returns.[12] The risk of an investment is based on how much profit the investment is likely to return, or more specifically, how much labour investors think they will be able to extract from their investment. A low risk investment is one in which the company or group which is invested in is expected to reliably produce surplus-value by way of labour power. Therefore, the equivalence for investments is grounded in how much labour is expected to come from them, parallel to the equivalence for commodities whose exchange-value is grounded in how much labour goes into them. By nature of their use-value, exchange-value, and equivalency, then, investments can be considered as commodities, rendering the Marxist terminology still relevant to even widely expanded modern systems of investments such as Wall Street stock trading. Furthermore, understanding investments as commodities supports the understanding of money itself as a means of production, because possessing money lends itself to the creation of more money by means of allowing crediting and investing.

[10] Ibid., 130.
[11] Ibid.
[12] Ascher, *Portfolio Society*, 41.

The relationship between modern economic systems of credit and investments and the exploitation of workers

Both investments and extensive credit networks serve to increase and hasten the wealth gap and concentrate wealth in the hands of capitalists by increasing the efficiency of capitalism's exploitation of the worker. The exploitation of the worker that is the life-blood of capitalism takes place mainly in two forms: first, by coercing workers into labour,[13] and second, by severing the worker from the right to the product or surplus-value that they produce.[14] An example of exploitation of both forms is a worker in a toy factory who must perform wage-labour to sustain themselves, but cannot afford to purchase the toys they produce for their own children.[15] They are coerced into working for subsistence, and they are also removed from the product of their labour, and so are exploited in both ways. Having a claim to the labour product can refer to either the commodity produced, or the surplus-value produced.

Investments increase the efficiency of the second form of exploitation (severance from labour products), while credit and lending increase the efficiency of both (coercion and severance from labour products). It is important to bear in mind that, while these forms of exploitation often do apply in particular cases of individual capitalists and workers, for the purposes of this paper, exploitation is considered between classes (capitalists and workers) instead of between individuals. This means that even if a worker is exploited by capitalist A, and credit additionally allows capitalist B to exploit them, the overall efficiency of the exploitation increases even though the relationship between the worker and capitalist A may not change. As I show, this is the form that increased exploitation by crediting and investments take.

First, crediting and lending allow capitalists to not only take surplus-value from the worker that is generated during the working day, but also to siphon back some of the wages awarded to the worker by way of charging interest, thus maximizing the surplus-value produced for the capitalist. For example, a worker who works for eight hours and generates $200 in value for the capitalist might get paid $100, thus generating $100 that goes to the capitalist. However, if the worker also has taken out personal credit or loans, they will need to

[13]Marx, *Capital* Volume I, 303.

[14]Ibid., 300.

[15]Kieran Allen, *Marx and the Alternative to Capitalism* (London: Pluto Press, 2011), 47.

pay interest charges. If the worker pays $3 from their daily wages in interest charges, they really only get paid $97, while the capitalist class receives $103 in surplus-value. Thus, the worker is further alienated from the product of their labour. Additionally, credit serves to increase the coercion on the worker, because debt serves as a work imperative. In a system where the best arrangement for capitalists is to extract as much work as possible and paying workers as little as possible (while ensuring their subsistence and continued labour power)[16] any chance event such as sickness or car problems could push a worker into needing to use credit. Once credit is taken out, the work imperative is increased, because there is only a downward spiral waiting for the low-wage worker who cannot pay off their debt. Therefore, credit both extracts more surplus value by separating the worker from their product (in this case, their wage), and strengthens the coercion to work, the two main forms of capitalist exploitation.

Second, investments divorce workers from the surplus-value they create by conferring the right to surplus-value produced to the owner of the investment. In the form of already existing companies, capitalists who possess enough money to buy investments (i.e. stocks) gain rights to the profits of the company, even while the workers share no part in the profits. Similarly, in the form of starting companies, capitalists use investments as a commodity to finance the start-up process (i.e. venture capitalism). Therefore, no matter how successful the workers are in creating surplus-value, the original capitalist who provided the investment will always have a claim to the profits, in addition to the capitalist(s) managing the operations. This leaves the worker with an even slimmer portion of the surplus-value. Both stocks and venture capitalism serve to increase the exploitation of the worker, as well as entrench the wealth gap between the classes. They make it easier for the capitalist to exploit labour power, because these investments are usually only available to those already possessing accumulated capital (i.e. capitalists).

The intrinsic nature of the wealth gap in capitalism

I have argued that credit and investments fit into a Marxist framework and that they increase the efficiency of worker exploitation, but does this necessarily mean that they drive inevitable income inequality and extreme

[16]Marx, *Capital* Volume I, 377.

wealth concentration in the hands of a few? I argue yes, drawing on the nature of capitalism as outlined by Marx and refuting a few common counter-arguments.

First, the exploitation of the worker is one of the primary characteristics of capitalism, and it is also what drives the wealth gap. The more that capitalists have sole claim to surplus-value created by the workers, the more wealth they will amass, while the wealth of the workers remains relatively constant in relation to that of the capitalists. Inherently, the more this gap grows, the more capitalists are rewarded. Capitalism is a drive for profit, but all profits are inherently linked to the degree of exploitation of the worker- the higher the profit margin for a company, the greater the severance of the worker from the surplus-value they create and thus the greater the exploitation. This understanding is supported by Marx: "the rate of surplus-value is an exact expression for the degree of exploitation of labour-power".[17]

One common counter-argument to the necessity of a wealth gap in capitalism draws on the morality of the capitalist; arguing that ethics or personal morality will cause individual capitalists to not extract as much as possible at the expense of the workers. However, this argument fails to take into account the very nature of capitalism: that it concerns a process and class, rather than individual capitalists,[18] and that capitalism is inherently vampiric, sucking as much surplus-value from workers as it can.[19] Additionally, even while the individual cannot be "responsible for [the] relations whose creature he remains, socially speaking", each capitalist is pushed to conform to the nature of capitalism because of the constant pressure from competition.[20] As Kieran Allen puts it, "[each capitalist] can always be eliminated by their rivals".[21] In other words, if one capitalist wavers and fails to abide by the vampiric nature of capitalism, which "will not let go 'while there remains a single muscle, sinew, or drop of blood to be exploited' ", they will fail to perpetuate their

[17]Marx, *Capital*, trans. Samuel Moore and Edward Aveling, Vol. I ((1867) 1887), 159 (footnote 7).

[18]Marx, *Capital* Vol. I, 92 (Preface to the 1st edition) "My standpoint ... can less than any other make the individual responsible for relationships whose creature he remains, socially speaking, however much he may subjectively raise himself above them."

[19]Marx, *Capital*, Vol. I, 342

[20]Ibid., 92

[21]Allen, *Marx and the Alternative to Capitalism*, 32.

position, and will be replaced by other capitalists who have no moral qualms about abiding by the nature of capitalism.[22]

Another counter-argument centers around the theory of trickle-down economics: that more money in the hands of the capitalist will 'trickle-down' and lead to more money in the hands of the workers. I point out three flaws in this theory. First, there will not be any 'trickle-down' effect because this ignores the nature of capitalism; if there is any extra surplus-value, the capitalist will retain it for themselves rather than let it pass to the worker.[23] Second, even if the working class has an increase in accumulated capital due to more jobs being created, or any other reason, this will not narrow the wealth gap because any increase in wealth of the workers is necessarily accompanied by a proportional or greater increase in wealth of the capitalists. In other words, even if workers receive greater absolute surplus-value, they receive the same portion, and the wealth gap is not diminished. Therefore, even if trickle-down economics creates jobs for the working class, it doesn't address the exploitative nature of capitalism or begin to suture the wealth gap together.

Trickle-down economics also often purports to allow for upward mobility, claiming that even if the exploitative nature of capitalism is unchanged, individuals will benefit because they can become capitalists, and therefore, the wealth gap isn't really a problem. However, this fails to consider the historical labour power that serves as the roots for modern capitalism and means of production. It is very difficult for the worker, who receives only wages, to get a foothold in the means of production or to become a capitalist, because they don't have access to the historically accumulated capital that is the basis for modern capital.[24] Furthermore, even if a select few individual workers manage to break free from the exploitation of credit and investment that is needed to achieve upward mobility, this does not change the relationship between the working *class* and the capitalists, it merely changes the position of a few individuals within the system.

---

[22]Marx, *Capital*, Vol I, 416.

[23]Ibid., 342

[24]Richard Peet, "Inequality and Poverty: A Marxist-Geographic Theory," *Annals of the Association of American Geographers* Vol. 65, No. 4 (Dec. 1975), https://www.jstor.org/stable/2562423, 565.

Conclusion

I have argued that the vampiric and exploitative nature of capitalism inherently drives society to an extreme wealth gap, and modern economic systems like credit and investing serve to hasten and exacerbate the process. These systems strengthen and entrench the wealth gap because they enhance the exploitative nature of capitalism.

Lastly, I briefly explore the extensions of my arguments with regard to the responsibility of the state to regulate capitalism. Clearly, capitalism left unfettered will not move to diminishes levels of exploitation or a more equitable wealth distribution of its own accord. However, the state can regulate capitalism in a way that reduces its negative consequences, just as it did in the early days of capitalism with legislation about the limits of the working day.[25] In fact, I take the arguments of this paper to serve as a mandate for such state regulation of capitalism. Any modern state that seeks to achieve justice or equality, modern notions of which are widely understood as providing some degree of equal opportunity, must hold the exploitation of capitalism in check. Multiple mechanisms for this regulation are possible, such as: progressive wealth taxes, protections against the monopolization of industry, and legislation protecting worker's rights to a living wage and reasonable working hours. If capitalism runs its course unchecked, continually accelerated by modern economic systems, there is no foreseeable outcome other than extreme disparity in wealth concentration wherein the rich have more money than is possible to spend while the middle and lower classes toil for bare subsistence, and one's lot in life is tremendously determined by the socioeconomic class they are born into. The understanding of modern capitalist exploitation that I have argued for in this paper serves as a starting point for moving forward and thinking about these issues.

[25]Marx, *Capital*, Vol. I, 390-91

Bibliography

Allen, Kieran. *Marx and the Alternative to Capitalism.* London: Pluto Press, 2011.

Ascher, Ivan *Portfolio Society.* New York City: Zone Books, 2016.

Chuck, Collins & Hoxie, Josh. *Billionaire Bonanza: Inherited Wealth Dynasties in the 21st-Century* U.S. Institute for Policy Studies (2018). https://inequality.org/great-divide/billionaire-bonanza-2018-inherited-wealth-dynasties-in-the-21st-century-u-s/.

Marx, Karl. *Capital.* Translated by Ben Fowkes. Volume I, (1867) 1976.

Marx, Karl. *Capital.* Translated by Samuel Moore and Edward Aveling. Volume I, (1867) 1887. https://www.marxists.org/archive/marx/works/download/pdf/Capital-Volume-I.pdf

Peet, Richard. "Inequality and Poverty: A Marxist-Geographic Theory." *Annals of the Association of American Geographers* Vol. 65, No. 4 (Dec. 1975): 564-71. https://www.jstor.org/stable/2562423.

Identity and Amputation in Saidiya Hartman's *Scenes of Subjection*

Navoneel Chakraborty

Abstract: This paper explores the concept of amputation in Saidiya Hartman's Scenes of Subjection, the processes used to perpetuate the concept, and its effect on black identity, both social and communal. As such, it considers the temporal and mnemonic effects of slavery, the middle passage, and centuries of discrimination, both internal and external to the black community that have worked to prevent the creation of a black identity in the Americas. To this end, I briefly explore the history of slavery and black culture, along with the mechanisms used by the dominating classes to maintain a social hierarchy that disenfranchises African-Americans, and the response to this; the creation of a unique black identity in the aftermath of slavery, segregation and discrimination.

In *Scenes of Subjection*, Saidiya Hartman utilizes the concept of amputation to describe the breach, and loss, caused by slavery to address the disconnect in kinship and natal alienation.[1] To Hartman, this resulted in the discontinuity of memory, a manner in which the dominators separated the enslaved from

[1]Saidiya V. Hartman, *Scenes of subjection: terror, slavery, and self-making in nineteenth century America.* (New York: Oxford University Press. 1997).

their communities, and the internal and external imposition of a social hierarchy in which the enslaved were unable to form communities, that included both the living and the dead.[2] However, she fails to address the effect this mnemonic amputation had upon black identity during, and after slavery, especially given how racial identity has been linked to the psychological health and well-being of African-Americans.[3] For the purposes of this paper, I will attempt to demonstrate how mnemonic and communal 'amputation' also led to the creation of a black identity unique to the enslaved and resulted in the further creation of societal and legal structures that were oppressive to racialized subjects.

To this end, I consider identity to be a fundamental part of meaning-making, or self-making, and how its denial would contribute to the domination of the enslaved by their captors. Identity, in this paper, will be primarily defined as the sociological concept of self-making, as conceptualized by Peter Weinreich:

> "A person's identity is defined as the totality of one's self-construal, in which how one construes oneself in the present expresses the continuity between how one construes oneself as one was in the past and how one construes oneself as one aspires to be in the future"; this allows for definitions of aspects of identity, such as: "One's ethnic identity is defined as that part of the totality of one's self-construal made up of those dimensions that express the continuity between one's construal of past ancestry and one's future aspirations in relation to ethnicity".[4]

Given this definition of identity, the role played by discontinuity becomes evident in "amputating"' the connection that the enslaved would feel with their communities, their traditions, their ancestry, and their 'home' of Africa. Discontinuity plays this role in hacking off this limb, and leaving behind a "phantom" memory, through natal alienation, kinless-ness, and the ghostly memory of the middle passage and of the African heritage of the enslaved

[2]Ibid.

[3]Ruth Horowitz. "Racial Aspects of Self-Identification in Nursery School Children." *The Journal of Psychology* Vol. 7, no.1 (1939): 91–99. doi: 10.1080/00223980.1939.9917623

[4]Peter Weinreich. "The operationalisation of identity theory in racial and ethnic relations." *Theories of Race and Ethnic Relations*, (1986): 299–320. doi: 10.1017/cbo9780511557828.016

subject.[5]

Natal alienation was the practice of separating mothers from their children, both physically and affectively. Physically, either mothers were prevented from caring for their children, or through means of sale. Affectively, the pain accorded to the care of a subject that was doomed to be not-quite-a-person, an individual condemned to a life of terror, similarly caused a form of alienation, as both children and mothers sought to spare themselves, or their children, the pain of losing loved ones, or seeing them be terrorized. This concept is aptly summed up in Toni Morrison's *Beloved*, as Sethe would rather see her children dead than in the custody of the slave-owning schoolteacher.[6]

This form of ancestral discontinuity then contributed to the kinless nature of slave identity. Being separated from their mothers at birth, and growing up without a traditional, heteronormative family would often lead to a sense of loss, furthered by the imposition of a backwards status[7], and cemented in the hopes of a life free of this pain.

> "Whether figured as 'life in Africa when they (we) were free' or embodied by ... unviolated natality ... or an understanding of the self in relation to the millions gone and/or those on the other side of the Atlantic"[8]

The lack of a community around which the enslaved could find solace and comfort contributed to this phantom limb, itself an extension of the hope of one day finding subjectivity and communal identity free of the terror of slavery.

This communal identity was then most often attributed to the foundational status of Africa in memory[9], as a continuing narrative that strove to connect the enslaved to an aspect of their existence that was, in their eyes, free. It was in the establishment of a shared history that the oppressed other may begin to assert their subjectivity and demand recognition and reciprocity.[10]

---

[5]Hartman, *Scenes of Subjection*.

[6]Toni Morrison. *Beloved*, (S.l.: Vintage Classics. 2004).

[7]Hortense Spillers. "Mamas Baby, Papas Maybe: An American Grammar Book." *Diacritics* Vol. 17, no. 2 (1987): 64. doi: 10.2307/464747

[8]Hartman, *Scenes of Subjection*.

[9]Ibid.

[10]Simone de Beauvoir. *The Second Sex.*Trans. Constance Bord & Sheila Malovany-

To this end, they strove to establish a connection with their history, through practices rooted in tradition, such as the use of overturned pots, or through a remembrance of the dead, as links in the chain that bound them to their ancestors.[11] The lack of proper burial, or death rituals, abruptly ended not only the lives, but the memory of those lost during slavery and the middle passage. As Hartman herself remarks, to remember the dead is to mend ruptured lines of descent and filiation.[12] Thus, this was a form of remembering that recognized this loss in its constitution of community.

The constitution of community was, however, also fraught with danger, stemming from the consequences of being discovered. This not only weakened any attempts at redress through remembrance, but the attempt to create a community in an environment that was legally and socially intolerable of such a concept was terrifying in and of itself. This was then aided by the threat of informants, or those amongst the enslaved who were subject to tortures until forced to inform. More often than not, this would in turn result in the exclusion of said enslaved subject, inadvertently contributing to a weakened community that was few in numbers.[13]

While focusing on the legal and social ramifications of such a phenomenon however, I find that Saidiya Hartman neglects the idea of personal identity, and the role that it played in the subjection of racialized individuals.

In studies done on identity, a particular distinction is made between personal identities and social identities, wherein the latter explicates the relationship between different social identity groups, known as in-groups and out-groups, and the former helps provide meaning for the self . However, it is undeniable that the two are closely linked, and Saidiya Hartman's focus on a social and communal identity neglects the effects that such amputation can have upon a personal identity, namely, distress and detriment to one's mental health and ability to create meaning for the self. In particular, this distress is

---

Chevallier. (London: Vintage Books: 2015), for clearer analysis on the effect a shared history has on the struggle against oppression. While Beauvoir is primarily speaking about feminist movements, it is my belief that the mechanisms present in the structural discrimination and the attempt to dismantle it are applicable to situations that extend outside of just feminism.

[11]Hartman, *Scenes of Subjection.*

[12]Saidiya Harman. "The Time of Slavery." *The South Atlantic Quarterly* Vol. 101, no. 4 (2002): 757-777. https://www.muse.jhu.edu/article/39111.

[13]Hartman, *Scenes of Subjection.*

caused if the feedback from others through perceptions of the self or reflected appraisals is incongruent with one's identity[14], a phenomenon that is bound to occur given the imposition of non-agency upon the subjugated that was so at odds with their subjectivity and activity. While her conceptualization of communal identity does well to consider the social categories that help determine a sense of belonging and the characteristics that form a part of self-conceptualization, it remains vital to consider that this social identity is at its strongest when it is acquired - namely, when an individual determines for himself this identity[15]. Meaning, in social identity, is created over time through culture and history[16], but a personal identity has a greater ability to affect this social identity[17], and the determination of in-group and out-group categorizations of social identity. However, the acquisition is of this identity, indeed, any identity is made even more difficult in a racializing community due to the imposition or ascribing of an identity to a racialized subject by the primarily white society, resulting in the imposition of a meaning, best summed up through the concept of "lateness".[18]

Thus, in the denial of a self-determined identity and meaning for racialized and enslaved subjects, both legally and socially, the apparatus for subjection not only denied enslaved the ability to create their own meaning in the world, but also set-up a system of oppression that defined its inclusion of the dominators, by its exclusion of the dominated[19], setting up in-group and out-

[14]Hogg, Terry & White. *A Critical Comparison of Identity Theory with Social Identity Theory.*

[15]Leonie Huddy, "From Social to Political Identity: A Critical Examination of Social Identity Theory." *Political Psychology* Vol. 22, no.1, (2001): 127-156. Retrieved from www.jstor.org/stable/3791909.

[16]Taylor, Charles. "The Politics of Recognition." *Multiculturalism*, (1994): pp. 25–74., doi: 10.2307/j.ctt7snkj.6.

[17]Given the nature as an identity that is taken up by or imposed upon an individual as a part of a group, it seems to me that from a subjective point of view, that personal identity occupies greater relevance. Once categorized, it may be impossible to break out of the group identity, but it is the incongruence between the identity ascribed and that acquired that causes the greatest distress. Therefore, personal identity has the ability to contribute to and detract from not only mental health, but to a social identity.

[18]Frantz Fanon, *Black Skin, White Masks.* (London: Pluto Press: 1986). The subject of lateness, in the case of racialized subjects, is the concept of a social memory, informed by history, that results in limited meaning making for the racialised subject, thereby detracting from their agency and subjectivity. See Fanon, *Black Skin, White Masks* for more details.

[19]David Cutler, Edward Glaeser, & Jacob Vigdor. *The Rise and Decline of the American*

group distinctions that have become difficult to undo. Identities are primarily determined through the lens of the other[20], in that our meaning is determined through our interaction with the other. Much like how the role of a mother is determined through its connection to the role of the father, whiteness was defined by its lack of blackness, and to be American meant to be white, even during and after reconstruction. This exclusion from national identity served to segregate and disadvantage racialized subjects while creating even greater issues in terms of identity. Here, we find that while black people were no longer American, they could not be considered African either, as was the case for many who returned to the continent.

The rapid disruption in the sense of a personal and national identity began with slavery, as yet another method of removing the agency of the enslaved and subjugating them to the property owners. In order to accomplish this, slave owners sought to eliminate certain outward expressions of African practices, going so far so as to wipe out any remnants of an African identity:

> "There is ample evidence to show that the slave masters went out of their way to break down the captives' identity with false substitutes. Such substitutes were not just restricted to whippings, mounted gun patrols, rapes, and other forms of punishment, but also to the disavowal of African images, symbols, and rituals. The desire of slave owners was to make the African feel inferior and dependent on the master".[21]

This exclusion of a black society was then used abundantly in the United States during slavery, and persisted long after emancipation. Jim Crow laws, miscegenation laws, and segregation all operated on the concept of the black person being equal but separate.[22] More often than not, separate also meant separate from the national, American identity, resulting in a struggle for those who were previously enslaved and their descendants to "define their place in North America".[23] Furthermore, the continuation of categorization

Ghetto. (National Bureau of Economic Research: 1996)

[20]Hogg, Terry & White. "Critical Comparison of Identity Theory".

[21]Francis Ngaboh-Smart, "The Politics of Black Identity: Slave Ship and Woza Albert!" *Journal of African Cultural Studies* Vol. 12, no.2, (1999): 167-185. Retrieved from www.jstor.org/stable/1771870

[22]Hartman, *Scenes of Subjection*

[23]Brian Thomas, "Struggling With the Past: Some Views of African-American Identity." *International Journal of Historical Archaeology* Vol. 6, no.2, (2002): 143-151. Retrieved

sharpened intergroup boundaries[24], paving the way for both a communal and personal self-enhancement, typically through comparisons with out-groups. In a society founded on the superiority of a white community, this undoubtedly meant the diminution of black identity and capability, leading to a vicious cycle that chipped away at black identity, as group members are more willing to discard membership in a group of low status[25], further diving community and ascribing these groups to a perpetually lowly status, as those able to change this status quo are integrated into the groups that are seen as being better.[26] This struggle to create a stable identity for the enslaved and their descendants can then be directly traced to the phenomena that Hartman terms 'amputation.'

These practices not only affected American perceptions of any black identity, but one that helped cement the identity of former slaves within Africa. In *Lose Your Mother*, Hartman describes her experiences in Ghana, where she was not seen as Ghanaian, but rather, another American tourist, and worst still, the descendant of slaves.*Saidiya Hartman. Lose Your Mother: a journey along the Atlantic slave route. (New York: Farrar, Straus and Giroux: 2007).* This genealogy was not only frowned upon in Ghana, but also served as the basis for the lack of solidarity felt between two communities that should be tied together by the horrors of the slave trade. This was in part due to those who were enslaved in the Africas. More often than not, they were members of the lower rungs of society, outcasts, criminals, the undesirables.

> "People pride themselves that their great-grandfathers kept slaves, and were not among the numerous slaves that abounded,' ... 'To be called a slave is an insignia of shame.' The dishonour of the slave had persisted, as had the dignity and self-respect of the affluent and the powerful".[27]

Thus, through Saidiya Hartman's account, we are able to see the 'amputation'

---

from www.jstor.org/stable/20852996

[24]Hogg, Terry & White."Critical Comparison of Identity Theory"

[25]Huddy, *A Critical Examination of Social Identity Theory.*

[26]The ever-changing nature of racialization led to such a phenomena, where throughout history, groups of people have been ascribed a status of being non-white, near-white, and white, depending on the utility of such categorizations, and its ability to maintain the superiority of the reigning racial power. See Gualteri (2009) for greater explication of this process in Arab-Americans, or Hartigan (1997) for its effect.

[27]Ibid.

of any connection with African kin that not only stems from the middle passage, but from the exclusion of those who were enslaved from society in the first place. To the extent that "numbers of blacks avoided using the term African ... because to continue to refer to oneself as African might encourage colonizationists to believe one wanted to be shipped back to Africa".[28] Naturally, this failed to be an option due to the lack of a black identity even in the perceived 'home' that was Africa.

This phenomenon of dissociation would eventually lead to the creation of a unique black culture in North America, one that directly stemmed from the ruptures of the middle passage, and the sense of dis-belonging within both Northern American and African societies. The identification and creation of such an identity was then essential, as Saidiya Hartman would argue, to beginning a process of redress for the wrongs suffered under slavery[29], and discrimination during the reconstruction era. The impact of this unique identity and its formation remains a poignant question and would require a greater undertaking to comprehend. However, it is important to note how inter-subjective relations play a role in the meaning making of an individual, and thus, despite her lack of acknowledgement of personal identity, that of a communal and interpersonal identity and community serves greatly in the understanding of this complex and layered issue[30]. Indeed, much in the manner that blackness cannot be understood without consideration of the afterlife of slavery, blackness can also no longer be considered without its primacy, vibrancy and generative capacity.[31]

[28]Thomas, *Some Views of African-American Identity*

[29]Hartman *Scenes of Subjection*

[30]Ibid.

[31]Alessandra Raengo, "*Dreams are colder than Death* and the Gathering of Black Sociality." *Black Camera* Vol. 8, no.2, (2017): 120. doi: 10.2979/blackcamera.8.2.07. Along with the article by Raengo, see film Arthur Jafa, *Dreams are Colder than Death* (2014) for an analysis of modern black culture, its history, and an interpretation of what it means to be black.

Bibliography

Beauvoir, Simone de, Borde, C., & Malovany-Chevallier, S. (2015). *The Second Sex*. London: Vintage Books.

Cutler, D. M., Glaeser, E. L., & Vigdor, J. L. (1996). *The Rise and Decline of the American Ghetto*. National Bureau of Economic Research.

Fanon, F. (1986). Black Skin, White Masks. London: Pluto Press.

Gualtieri, S. (2009). *Between Arab and White: Race and Ethnicity in the Early Syrian American Diaspora*. University of California Press. Retrieved from www.jstor.org/stable/10.1525/j.ctt1pnccz

Hartigan, J. (1997). "Establishing the Fact of Whiteness." *American Anthropologist*, Vol. 99(3), new series, 495-505. Retrieved from www.jstor.org/stable/681737

Hartman, S. V. (1997). *Scenes of subjection: terror, slavery, and self-making in nineteenth century America*. New York: Oxford University Press.

Hartman, S. V. (2007). *Lose your mother a journey along the Atlantic slave route*. New York: Farrar, Straus and Giroux.

Hartman, S.V. (2002). "The Time of Slavery". *The South Atlantic Quarterly* Vol. 101(4), 757- 777. https://www.muse.jhu.edu/article/39111.

Hogg, M., Terry, D., & White, K. (1995). "A Tale of Two Theories: A Critical Comparison of Identity Theory with Social Identity Theory." *Social Psychology Quarterly*, Vol. 58(4), 255-269. Retrieved from www.jstor.org/stable/2787127

Horowitz, R. E. (1939)." Racial Aspects of Self-Identification in Nursery School Children'.' *The Journal of Psychology*, Vol. 7(1), 91–99. doi: 10.1080/00223980.1939.9917623

Huddy, L. (2001). "From Social to Political Identity: A Critical Examination of Social Identity Theory". *Political Psychology*, Vol. 22(1), 127-156. Retrieved from www.jstor.org/stable/3791909

Jafa, A., Joseph, K., Mader, A., Johnes, A., Anyanwu, L. (Producers), & Jafa, A. (Director). (2014) *Dreams are Colder than Death* [Motion picture]. United States.

Morrison, T. (2004). *Beloved.* S.l.: Vintage Classics.

Ngaboh-Smart, F. (1999). "The Politics of Black Identity: Slave Ship and Woza Albert!" *Journal of African Cultural Studies*, Vol. 12(2), 167-185. Retrieved from www.jstor.org/stable/1771870

Raengo, A. (2017). "*Dreams are colder than Death* and the Gathering of Black Sociality." *Black Camera* Vol. 8, no.2, (2017): 120. doi: 10.2979/blackcamera.8.2.07

Spillers, H. J. (1987). "Mamas Baby, Papas Maybe: An American Grammar Book". *Diacritics*, Vol. 17(2), 64. doi: 10.2307/464747

Taylor, Charles. (1994) "The Politics of Recognition." *Multiculturalism*, 1994, pp. 25–74., doi:10.2307/j.ctt7snkj.6.

Thomas, B. (2002). "Struggling With the Past: Some Views of African-American Identity". *International Journal of Historical Archaeology*, Vol. 6(2), 143-151. Retrieved from www.jstor.org/stable/20852996

Weinreich, P. (1986). "The operationalisation of identity theory in racial and ethnic relations." *Theories of Race and Ethnic Relations*, 299–320. doi: 10.1017/cbo9780511557828.016

A Lewisian Account of Hinge Commitments

Helena Lang

Abstract: In *Epistemic Angst*, Duncan Pritchard presents an interpretation of Wittgenstein's *On Certainty* as a reply to the skeptical argument. In this paper, I argue for Lewisian epistemic contextualism (henceforth LEC) as a fruitful reading of Duncan Pritchard's flavour of Wittgensteinian hinge epistemology (henceforth PW, based on his 2015) as outlined in *On Certainty*. First, I sketch the skeptical paradox and G. E. Moore's solution to it. Then, I outline Wittgenstein's critique of Moore's solution as spelled out in On Certainty. After, I lay out PW as a response to the 'closure problem' faced by this Wittgensteinian critique. Finally, based on Lewis's 1996, I develop a contextualist account of PW that models hinge propositions as propositions that are only properly conceived as hinge propositions in contexts with skeptical standards, but are correctly conceived and known as everyday propositions in contexts with everyday epistemic standards.

1. Introduction

In this paper, I argue for Lewisian epistemic contextualism (henceforth LEC) as a fruitful reading of Duncan Pritchard's flavour of Wittgensteinian hinge epistemology (henceforth PW, based on his 2015). First, I sketch the sceptical paradox and Moore's solution to it. Then, I outline Wittgenstein's critique of

Moore's solution. After, I lay out PW as a response to the 'closure problem' faced by this Wittgensteinian critique. Finally, based on Lewis's 1996, I develop a contextualist explanation of PW. Defending this account, I consider the objection that LEC fails to argue that hinge commitments are rationally grounded knowledge (even in everyday contexts), to which I reply that the usefulness of the condition of rational ground is questionable. I reply to another objection which is based on Pritchard's strict no-belief condition and argue that the latter is implausible in its own right. I finally argue that LEC's account of the intuitions PW tries to explain is superior since LEC explains how hinge commitments can become knowledge and since it provides an error theory as to why the sceptical paradox is puzzling.

2. The skeptical argument and Moore's response

First, consider the skeptical argument:

(1) If I know that I have hands (henceforth 'that h'), I know that I am not a brain in a vat.

(2) I do not know that I am not a brain in a vat.

(3) Therefore, I do not know that h.

(1) is very plausible because it is derived from the closure principle which states that deduction is a normal way for us to enlarge our knowledge: If I know that p and I know that p entails q, then I am in a position to know q. (2) also seems to be true, because in its strongest form, the scenario of my being a brain in a vat (one of the skeptical scenarios) is supposed to be internally indistinguishable from "normal perceptual conditions".[1] (3) validly follows from the other two premises. But it is highly unattractive to accept (3), because it deeply conflicts with our intuition that we do know many propositions about the world, such as h. Thus, one method to argue against (3) as a conclusion is to motivate the falsity of (1) or (2). One such argument for rejecting (2) was offered by G.E. Moore:

(1) If I know that I have hands (henceforth 'that h'), I know that I am not a brain in a vat.

[1]Duncan Pritchard, *Epistemic Angst: Radical Skepticism and the Groundlessness of Our Believing* (Princeton: Princeton University Press, 2015), 11.

(2') I know that h.

(3') Therefore, I know that I am not a brain in a vat.

As to why (2') is true, Moore insists he is more certain of this proposition than he could be of anything.[2] We can see (2') here being used as an epistemic foundation, on the basis of which, together with using (1), other propositions can then be inferred.[3]

3. Wittgenstein's critique of Moore's response

Wittgenstein criticized the above response. According to him, there is a problem with both this kind of argument against the skeptic and with the skeptical argument itself.[4] First, his critique of Moore's response: As seen before, Moore takes (2') to play a foundational epistemic role in his response to the skeptic. This seems alright because Moore takes himself to be most certain of (2') and a belief's being most certain seems to entail that this belief is very well supported. But Wittgenstein thinks that if Moore says that he is most certain of (2'), this premise cannot be rationally supported.[5] Wittgenstein thus implicitly endorses this principle:

(RS) A belief is rationally supported iff one is more certain of the reason for one's belief than of the belief itself.[6]

For example, Wittgenstein even rejects the 'sight' of his hands as evidence for h, because this evidence is not more certain than his belief that h.[7] It follows that, as nothing could be more certain than Moorean propositions such as (2'), and things that are certain must be backed up by rational support, Moorean propositions cannot both be, in this specific sense, rationally supported and fulfill the role of an epistemic basis.

Second, Wittgenstein's critique of the skeptical argument: h is a proposition which cannot be rationally doubted in virtue of its being most certain. This is because any grounds we would have for that doubt would have to be more

[2] Pritchard, 64.

[3] Ibid.

[4] Ludwig Wittgenstein, *On Certainty*, trans. G. E. M. Anscombe and G. H. von Wright (New York: Harper, 1969), para. 250.

[5] Pritchard, 64.

[6] Ibid., 65.

[7] Wittgenstein, *On Certainty*, para. 250.

certain than h itself. Here, Wittgenstein implicitly appeals to a principle parallel to (RS). On his view, reason for doubt must be more certain that what is doubted. Thus, according to him, any total, positive or negative, rational evaluation of our claims is incoherent. Rational evaluation always has to take place against the backdrop of hinge commitments which we are most certain of and which we hold irrationally.[8]

Then Pritchard points out a problem for this Wittgensteinian account of the structure of rational evaluation.[9] This problem follows from the closure principle and the fact that our hinge commitments are held without rational ground.[10] Here is the problem phrased in two questions: (i) why can we not gain rational support for hinge commitments by inferring them from non hinge propositions while the rational support of the latter is preserved? And, (ii) how do we keep 'local' rational support if our hinge commitments are not rationally supported?[11] In other words, why is rational support not transmitted in an inference from ordinary to hinge propositions and how is it generated at some point so that our ordinary beliefs are rationally supported? Pritchard calls this the 'closure problem' for the Wittgensteinian account of rational evaluation.[12] It is sketched here following Neta:[13]

(A) We cannot have rationally grounded knowledge of hinge commitments.

(B) The closure principle: "If S has rationally grounded knowledge that p, and S competently deduces from p that q, thereby forming a belief that q on this basis while retaining her rationally grounded belief that p, then S has rationally grounded knowledge that q".[14]

(C) We have rationally grounded knowledge of everyday propositions.

On Wittgenstein's view, (A)-(C) should all be true. But this clashes with the fact that (A)-(C) can, apparently, not all be true at the same time given that hinge propositions can be deduced from everyday propositions.

[8]Pritchard, 66.
[9]Ibid., 73.
[10]Ibid., 89.
[11]Ibid., 73.
[12]Ibid.
[13]Ram Neta, "An Evidentialist Account of Hinges," *Synthese*, January 30, 2019, 2, https://doi.org/10.1007/s11229-018-02061-0.
[14]Pritchard, 72.

4. PW as a response to the closure problem Wittgenstein faces

In response to this, Pritchard defends PW as an interpretation of Wittgenstein's response to the skeptical argument explained above. PW is supposed to avoid the closure problem in arguing that hinge commitments are not "in the market for rationally grounded knowledge".[15]

PW distinguishes three kinds of hinge commitments. First, there are hinge commitments that are specific to oneself, "personal hinge commitments", such that one has hands or that one has never been to the moon.[16] These are the hinge commitments that Wittgenstein thought were at play in the Moorean argument and our knowledge of these propositions is called into question in the skeptical argument. Pritchard explains their nature further in saying that our personal hinge commitments "codify" one über hinge commitment", namely that "one is not radically and fundamentally mistaken in one's beliefs".[17] His motivation for this is that our personal hinge commitments are not "entirely context-bound".[18] They are such because "to be wrong about something like this would reflect radical and fundamental error".[19] A third kind of hinge commitments are "anti-skeptical hinge commitments".[20] These codify "our attitude to radical skeptical scenarios" and because we take a stance of denial towards skepical hypotheses, they follow from our über hinge commitment but are one instantiation of it concerning a particular case.[21] They also differ from the personal hinge commitments in that most people do not have them before encountering the skeptical argument.[22]

4.1 Hinge commitments: the nonbelief reading

PW tries to establish the case of us not being able to have knowledge of hinge propositions by arguing that we cannot even rationally believe them in a way that would be aiming at knowledge or, in other words, a

[15] Pritchard, 89-90.
[16] Ibid., 96.
[17] Ibid., 96.
[18] Ibid., 95.
[19] Ibid.
[20] Ibid., 97.
[21] Ibid.
[22] Ibid.

necessary condition for it.[23] Pritchard calls this "nonbelief reading" of hinge commitments.[24]

The argument for the nonbelief reading goes as follows[25]:

(I) We are as certain of hinge commitments as we could be of anything.

(II) If we do not have a reason for our belief that p, we cannot coherently believe that p.

(III) We do not have a reason to believe hinge commitments (by I and RS).

(IV) Thus, we cannot coherently believe hinge commitments (by II, III). A fortiori, we cannot know hinge commitments.

4.2 PW on the closure problem

We saw before that (A)-(C) of the closure problem were all entailed by Wittgenstein's account, but incompatible. PW now argues that, given RS, inferring hinge propositions from everyday propositions cannot be rational, so the closure principle in (B) does not apply in such cases. Thus, (A)-(C) can all be true. More precisely, on PW, (A) is true because we do not have rationally grounded knowledge of (all types of) hinge commitments, as has been established by the argument for the nonbelief reading. The hinge commitments are irrationally held as a background for other epistemic activity. (B) holds and is compatible with the conjunction of (A) and (C) because the process of deduction leading to the belief must be rational.[26] But by the notion of rational ground which Pritchard here presupposes, one could not rationally undergo such an inference and then end up with a belief geared towards knowledge as the result. It is in this sense that the hinge commitments are outside the market of knowledge. Finally, (C) is also held true on this account, as long as this knowledge is rationally supported, which is only possible while holding the irrational hinge commitments. Importantly, hinge commitments are not part of the everyday propositions appealed to in (C).

Our propositional attitude towards hinge commitments, according to PW,

[23] Pritchard, 91.
[24] Ibid.
[25] Ibid., 90.
[26] Ibid., 91.

is "a commitment to the target proposition that is incompatible with an attitude of agnosticism about its truth".[27] One might also "endorse" the proposition.[28]

5. A Lewisian epistemic contextualist account of PW

Now I move on to present an attributor contextualist account, based on Lewis's 1996, of PW that I take to provide a fruitful and intuitive explanation of it. This view can explain PW's solution of the closure problem, the mysterious propositional attitude, the locality of rational support, how hinge commitments can become knowledge, and provides a better error theory concerning the skeptical paradox.

Consider Lewisian epistemic contextualism (henceforth LEC). LEC as a view concerns the *semantics* of the verb 'know', but it is still epistemologically interesting because of our seemingly paradoxical usage of the verb 'know' in the skeptical argument, as we saw above.[29] On this view, "x satisfies 'knows p' in context C iff x's evidence e eliminates every not-p-world w, except for those that are properly ignored in C".[30] What constitutes a properly ignored possibility is spelled out by Lewis in several rules. The one most relevant here is the "Rule of Attention": according to this rule, a possibility that is attended to is not properly ignored.[31] In Lewis's words, "no matter how far fetched a certain possibility may be [...], if in this context we are not in fact ignoring it but attending to it, then for us now it is a relevant alternative".[32]

5.1 LEC on the closure problem

First, consider its explanation of the 'closure problem' for Wittgenstein's account, a problem that PW allegedly provides a good response to in arguing that (A)-(C) are compatible. On LEC, (C) is truly asserted in contexts with

[27]Pritchard, 101.

[28]Ibid., 92.

[29]David Lewis, "Elusive Knowledge," *Australasian Journal of Philosophy* Vol. 74, no. 4 (December 1996): 550, https://doi.org/10.1080/00048409612347521.

[30]Michael Blome-Tillmann, "Knowledge and Presuppositions," *Mind* Vol. 118, no. 470 (April 1, 2009): 245, https://doi.org/10.1093/mind/fzp032.

[31]Lewis, "Elusive Knowledge", 559.

[32]Lewis, 559.

low epistemic standards, or so-called everyday contexts. In these contexts, due to our practical interests, we properly ignore alternatives such as worlds in which the skeptical hypothesis is true. Thus, our evidence does not have to eliminate these worlds in order for us to satisfy 'knows p'. Here, I want to motivate the claim that propositions such as h, that is, personal hinge commitments on PW, are only properly conceived as hinge commitments in skeptical contexts and are correctly conceived of as known everyday propositions in contexts with everyday epistemic standards.

In skeptical conversational contexts, the skeptical alternatives have become salient. These alternatives would thus have to be eliminated in order for us to know a proposition. It was in this context that Wittgenstein made his remarks on the hinge-role of propositions such as h. In these contexts, to hold onto one's commitment to know propositions such as h indeed amounts to irrationality and shows that one is confused about the conversational context one is in. But that, of course, isn't the case in all conversational contexts! In everyday contexts, we properly ignore skeptical alternatives and thus know propositions such as that we have hands.

Another way to formulate h as a hinge commitment within the framework of LEC is to say that if we are doubting our knowledge of propositions such as h, we indeed cannot have knowledge. For expressing doubt of h would amount to switching the context and shifting to a context with very high epistemic standards. This is on par with the requirement by PW that we have the hinge commitments (i.e. our "endorsing" of hinge propositions) in place if we are to have any knowledge.

(A), on LEC, is true in skeptical contexts, but false for (personal) hinge commitments such as h in everyday contexts. As we have seen before, in skeptical conversational contexts, our evidence would have to eliminate the skeptical hypotheses in order for us to know anything. Thus, in these contexts, we indeed don't know the personal hinge commitments. In everyday contexts, however, we can know propositions such as h. Thus, (A) captures the intuition that we don't know the personal hinge commitments in skeptical contexts.

Considering the propositional attitude of hinge commitments according to Pritchard, we can thus accommodate PW's nonbelief account, but flesh it out more. In everyday contexts, we can believe and know personal hinge commitments such as h. In skeptical contexts, however, we cannot know

them, as the skeptical alternatives have become relevant. But, still, in these contexts, we might have an attitude of certainty or acceptance towards them, because we are confused about the conversational context that we are in. Concerning the über hinge commitment, in everyday contexts, we properly ignore alternatives in which it is false, i.e. in which skeptical scenarios are true. Thus, we also can be said to have an attitude of acceptance towards the über hinge commitment, i.e., the denials of skeptical hypotheses, namely the attitude towards a proposition which expresses an alternative that is properly ignored.

(B) also holds, but it does not threaten our everyday knowledge on more plausible grounds than the ones that PW offers. PW held that it does not threaten knowledge of everyday propositions because a competent deduction cannot involve hinge commitments. On LEC, (B) holds if we stay in the same context during the inference, i.e. during the inference, we do not attend to other possibilities that are then not properly ignored. This is why its application does not threaten our knowledge of ordinary propositions, our "local support", as Pritchard puts it.[33]

5.2 LEC on local rational evaluation

LEC also captures PW's claim that rational evaluation (e.g., of propositions such as h) can only be local, that is, it can only take place against the backdrop of the irrational held hinge commitments that are immune to rational evaluation. On LEC, this locality of rational evaluation is just an expression of the fact that, in order to be able to know everyday propositions, we have to be in a conversational context whose standards are low enough so that the personal hinge commitments, such as h, are known. So, to use PW terminology, LEC also requires the hinge commitments to be in place for us to know everyday propositions – not as irrationally held, as on PW, but rather as known. This captures the sense in which there is something incoherent about the skeptical argument. When we try to evaluate the über hinge commitment in everyday contexts, we switch contexts. If we do not notice this, we are confused, as we then face the skeptical paradox.

[33]Pritchard, *Epistemic Angst*, 71.

5.3 LEC on Rational Support (RS)

We can also model the requirement of Rational Support along the lines of LEC: rational ground, or our reasons for believing a proposition, on this account, are modelled as which alternatives we can properly ignore, and this, of course, depends on our evidence and on the conversational context.[34] Evidence in Lewisian terms is the sum of our perceptions and memory.[35] This is what, on LEC, gives us the rational ground to truly assert that we know h.[36]

5.4 LEC on the nonbelief reading

Pritchard could still object, though, that on PW, we don't believe hinge commitments and can *never* come to believe them. This would make unsound my argument that we can believe (and know) hinge commitments in normal contexts. But the reasoning in (I)-(IV), which is what underlies the no-belief condition, is ultimately not plausible, especially not the conjunction of (I) and (IV). If one consequence of PW is to hold that we are both most certain of propositions such as h and, at the same time, cannot believe them, that seems a reason against holding this view. Not only this, but on this view, it also seems hard to have any knowledge, because lots of what we would normally take ourselves to know would turn out to be a personal hinge commitment. It is unclear what excludes hinge commitments from other propositions. On PW, the line between ordinary, rationally supported beliefs that constitute knowledge and irrationally held personal hinge commitments seems blurred. There seems to be no rigorous criterion for this distinction. On LEC, we don't need to postulate such a distinction, as in everyday contexts, we can know both personal hinge commitments and other propositions.

5.5 LEC: from hinge commitment to knowledge

LEC can also resolve the difficulty of PW to explain how hinge commitments can become knowledge. Pritchard picks up the example of never having

[34] Importantly, this is not to say that the standard for how 'watertight' a reason must be varies with with context, as this is something that Lewis explicitly rejects (551). However, I take here the notion of 'rational ground' and try to model it along the lines of eliminating relevant alternatives and properly ignoring relevant ones.

[35] Lewis "Elusive Knowledge", 553.

[36] As in, that which turns our true belief into knowledge (see Lewis 551).

been to the moon. A few centuries ago, for example, people had the hinge commitment of never having been to the moon. But in the future, where moon travel might perhaps have become a normal phenomenon, people might not have this hinge commitment.[37] PW explains this as follows: the über hinge commitment now codifies different personal hinge commitments and the one that one has never been to the moon is no longer among them. This means that at some point, it must become possible to rationally believe (and perhaps know) that one has not been to the moon. But this dichotomy of rational and non-rational belief as understood here (where the latter is supposed to be closed off from rational deduction) does not align well at all with the *process* of something such as moon travel becoming more and more commonplace. In other words, belief must be either rational or irrational, whereas moon travel can either never have been done or be completely normal and lots of states in between.[38]

LEC offers a smooth account of this process: because of people's practical interests, they could know that others in the same epoch hadn't been to the moon. They properly ignored possible worlds in which others went to the moon and thus their evidence didn't have to eliminate this. In the future world, as the possibility of going to the moon will have become salient, it is not properly ignored anymore in more and more everyday conversational contexts. It has become a relevant alternative in those contexts. This possibility would thus have to be eliminated by our evidence in order for us to know that someone hasn't been to the moon. On the LEC model, moon travel becoming more and more commonplace would be mirrored by the fact that the possibility that someone has gone to the moon becomes salient in more and more conversational contexts. Unlike PW, LEC can thus account for the *process* of moon travel becoming a normal undertaking and people adjusting their attitudes to the relevant proposition accordingly.

5.6 LEC and Pritchard's notion of rational ground

Now, I consider the condition of rational ground that, according to Pritchard, has to be fulfilled in order for us to have knowledge, and his view that

[37] Pritchard, *Epistemic Angst*, 95.

[38] Wittgenstein also seems to explicitly refer to this "fluidity" of hinge commitments in 96 and 97 of OC. This might render Pritchard's account of hinge commitments also less plausible from an exegetical perspective.

there is such a thing as rationally grounded knowledge. First, Pritchard holds PW to be superior to views that spell out hinge commitments as propositions that we know even though we "lack a rational basis" for holding them true.[39] Rational foundation here is having "rational support that favours one's belief that [e.g., h] over the [skeptical] hypothesis".[40] This is because of his view that without a "solid rational foundation" that perceptual knowledge possesses, it is unclear "in virtue of what it [counts] as bona fide knowledge".[41] Formulating the problem of how we can know things like this presupposes that there is a certain condition that must be fulfilled in order for there to be knowledge: it must be 'rationally grounded'. This might beg the question against accounts of knowledge that ground knowledge in the elimination of relevant alternatives such as Lewis's and thus against Lewisian evidence. The upshot is that a theory in the style of Lewis and one along the lines of PW ultimately rest on different conceptions of knowledge and how they figure into the skeptical argument. Both of these frameworks are best evaluated in their entirety, not just by looking at how plausible their 'epistemic conditions' are. Ultimately, I do not want to argue for one view or the other here, but just point out how LEC neatly accounts for PW.

5.7 LEC: intuitions and error theories

Lastly, given the fact that we do intuitively seem to have widespread knowledge of propositions such as h, PW has to supply a very plausible error theory. This error theory must answer why we do not have knowledge of Moorean certainties but think we do, which goes hand in hand with answering why the skeptical paradox constitutes a paradox. The only response in this direction that Pritchard offers is that we are mistaken in not recognizing the locality of rational evaluation. But how come we are so gravely mistaken about the nature of rational evaluation? LEC, however, offers an error theory in terms of semantic confusion and appeals to other context-sensitive expressions in explaining the behaviour of 'know.'

Additionally, PW's error theory must account for our being wrong about our knowledge in many cases, namely, every time we take a hinge proposition to be knowledge. The Lewisian contextualist, on the other hand, only has to

[39] Pritchard, *Epistemic Angst*, 73.
[40] Pritchard, 30.
[41] Ibid., 31.

explain away our confusion in skeptical contexts.

6. Conclusion

In conclusion, facing the skeptical argument and the closure problem for Wittgenstein's response to Moore's argument against the skeptic, I have argued that Lewisian epistemic contextualism provides a fruitful explanation of Pritchard's interpretation of Wittgensteinian hinge epistemology. It captures the intuitions the latter set out to explain. It explains the closure problem for Wittgenstein's account of rational evaluation by an appeal to switching contexts with more intuitive support than PW. Unlike PW, LEC explains how hinge commitments can become knowledge and provides an error theory as to why the skeptical argument is puzzling.

Bibliography

Blome-Tillman, Michael. "Knowledge and Presuppositions." *Mind* 118, no. 470 (April 1, 2009): 241–94. https://doi.org/10.1093/mind/fzp032.

Lewis, David. "Elusive Knowledge." *Australasian Journal of Philosophy* 74, no. 4 (December 1996): 549–67. https://doi.org/10.1080/00048409612347521.

Neta, Ram. "An Evidentialist Account of Hinges." *Synthese*, January 30, 2019. https://doi.org/10.1007/s11229-018-02061-0.

Pritchard, Duncan. *Epistemic Angst: Radical Skepticism and the Groundlessness of Our Believing.* Princeton: Princeton University Press, 2015.

Wittgenstein, Ludwig. *On Certainty.* Translated by G. E. M. Anscombe and G. H. von Wright. New York: Harper, 1969.

A New Moral Methodology for AI Value Alignment

Christian Gonzalez-Capizzi

Abstract: In the following essay I attempt to motivate and unpack what I believe is the most important question the ethicist interested in AI Ethics may ask. Over the course of the following pages, I seek to outline a new methodology for doing moral philosophy so that we can make progress with respect to this question. This methodology invites the use of recent developments from the field of complexity science, namely, agent based modeling, in addition to lending an ear to what sociologists, psychologists and logicians have to say.

I will first motivate what I believe is a good starting point for this question: considering what moral system a super intelligent machine should use as its action guiding principles. This question naturally arises from a discussion of what is potentially at stake with the inevitable onset of machines with greater than human intelligence. Particularly, a discussion of these machines with values which do not align with our own. Through subsequent iterations of unpacking the question of which moral system a superintelligent system should use, we will arrive at a computational/empirical methodology for answering the clearest statement of this question.

Introduction

What is the most important question we can ask? Admittedly, this is a loaded question, but if we unpack it a bit perhaps we can find a satisfying answer. The key term here is "important." What do we mean by important? Questions of importance, and similarly, questions of value, and questions of what matters in the realm of human action fall into the domain of ethics. I don't intend to fully map out the major disagreements in the field of ethics, but there tends to be agreement on one issue: whatever it is that matters, whatever values that may or may not exist, in the absence of conscious creatures in the universe, all talk of values is as good as meaningless. What good are values, duties or virtues if there are no beings around to value it? So if there was a question whose immediate answer would prevent the ultimate moral disaster, the extinction of all known life, this would be a satisfying answer to our original question (assuming life is, on average, worth living). Which questions might yield answers which would help us avoid an extinction-level threat of the highest probability and urgency? The following are possible candidates:

1. What is the best path forward for mitigating the threat from a nuclear war?

2. What is the best path forward for mitigating the threat from a biological war?

3. What is the best path forward for mitigating the threat from climate change?

While each of these questions is valuable, the question I would like to focus on for the duration of this essay will focus on the development of super intelligent or artificial general intelligence (AGI) machines, machines whose levels of intelligence dwarf any human or even collection of humans across multiple domains. More precisely, I will be concerned with their decision procedures. If we understand moral or ethical theories as decision guiding procedures which help one pick out better and worse actions or states of affairs, we can state the question as what will be taken as the central question of this essay:

> **Central Question:** Which moral system should superintelligent machines use?

How is this question related to existential threat? If the prospect of creating a god (or gods) in a machine isn't immediately concerning, allow me to motivate not only the urgency behind this question, but why this is the right question to be asking in the first place.

Superintelligence and Misaligned Values

Any mismatch between a superintelligent machine's goals and humanity's goals broadly speaking is potentially catastrophic. Take a superintelligent machine with the task of maximizing, say, paperclip production. On its face, this seems like an innocuous enough task to not warrant any grand suspicion or concern. However, it's possible that this paperclip maximizer might proceed "by converting first the Earth and then increasingly large chunks of the observable universe into paperclips".[1] Soon enough, every usable inch of the universe within this super intelligence's region of space will be filled with paperclip production facilities, humans be damned. This would, after all, maximize paperclips.

Speaking more generally we should recognize that, given any sufficiently unspecific goal, the space of all possible means of arriving at that goal is infinite. Furthermore, we must remind ourselves what is stipulated in granting that a machine is truly super intelligent. This machine will likely find optima which are so hyper-efficient so as to be inconceivable and unanticipated for any human or group of humans to predict. In fact, it is unlikely that human cognition, collective or not, will be able to zero in on the means that a superintelligent machine would settle on. In addition, this superintelligence would likely be intelligent enough to achieve those goals regardless of human stopgaps, or even mobility issues.[2] If this machine is truly superintelligent, it will run through any and all walls on its way to achieving its goal as long as their goal is still physically possible. So this machine's ultimate goal may manifest itself in ways that are worlds apart from the goals of humans. And many goals carried out to their logical extreme (as an optimizing AGI would try to do) may have a non-negligible chance of causing existential catastrophe. So even if the paperclip maximizer sounds unlikely, the takeaway is that *any*

[1] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

[2] The problem of how to control AGI known as the 'Control Problem'. The existence of a solution is very much an open question.

misaligned values between a superintelligence and humanity at large could spell disaster. Surely the number of ways of building unsafe AGI vastly outnumbers the number of ways of building safe AGI. And any non-negligible probability of existential threat must be taken seriously.

The Orthogonality Thesis

At this point it might be objected that any sufficiently *rational* agent will come to the same conclusions for questions of values and morality more broadly. If one believes morality is somehow grounded in *rationality*, or even that *rational* agents have a tendency to agree on the truth of the matter, then this is the right conclusion to come to.

On first appearance, this makes sense. Intelligent beings are rational creatures, and rational creatures *tend* to converge on the question of what does and does not matter in the moral domain. One would be hard pressed to find two truly rational people who disagree that mass genocide is a moral horror of the highest magnitude. While the idea that rationality and morality at least somewhat track one another seems plausible given convergence in moral reasoning over the past few hundred years (i.e. the abolition of slavery, the adoption of legal and moral rights into our vocabulary, etc.), it sadly rests on a misunderstanding of the type of rationality an artificially intelligent machine possesses.

The problem with this thinking is that it rests on vagueness of language. In the first paragraph of this section the term 'rational' is used repeatedly, but without a singular meaning that encapsulates all uses of the word.

AI systems are, at their core, **instrumentally rational**.[3] Not rational without qualifications. That is, given some objective (or value, in moral parlance) these machines will find optimal means of arriving at, or maximizing for that objective. By contrast, the rationality a Kantian or a similarly inclined moral realist (someone who posits objective moral facts) talks about is a type of rationality without qualifications, a kind of rationality we speak of when we attribute it to a fellow human: rationality in some thick, *normative* sense. This kind of rationality is notably different from the instrumental

[3]Nick Bostrom uses a similar definition of intelligence. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificail Agents", *Minds and Machines* 22 (2):71-85, p. 74.

rationality that our superintelligent machine has and is a topic we will return to later.

Given that our AGI is rational only in the instrumental sense, we can safely conclude that we may plug in *any* arbitrary value we wish into its optimisation algorithm and expect hyper-efficient results. In short, an AGI's value(s) and intelligence are entirely independent of each other. So any AGI can be plotted on a graph with the orthogonal axes of value and intelligence. This value independence is known as the *Orthogonality Thesis*.[4]

While it's unlikely superintelligent machines will be given a naively unconstrained goal such as maximizing paperclip production, the takeaway is that the wrong goal/value, when internalized by an AGI, can pose an existential-level threat to humanity. All it takes is getting this wrong just once for us to not have any second chances. Once a machine whose intelligence eclipses the collective brainpower of all of human history is made, any sufficiently poorly selected goal/value could entice this machine to view us as a minor obstacle on its way to its final objective. As such, the values an AGI uses in its decision procedure – its moral system – is of the utmost importance.

Value Alignment

To reflect on what's been laid out thus far: Intelligence and values run in completely orthogonal directions. A superintelligent machine may be superintelligent with any possible value plugged in. We've also concluded that any mismatch between this machine and human values at large could potentially lead to catastrophic results. So value selection is a topic of great importance.

It is at this point that the discussion normally turns to questions of AI Value Alignment, as it is often referred to. **Value Alignment** is the general research project, both technical and philosophical, of finding out how we can align the values of intelligent machines with those of humanity. Value Alignment research has therefore focused heavily on the following descriptive question:

> **Descriptive Question:** How can we align a superintelligent machine's moral system with the moral system we humans *actually*

[4] Bostrom, *The Superintelligent Will* 73.

use?

This question is usually tackled with a combination of various sophisticated machine learning techniques such as inverse reinforcement learning.[5] However, these approaches to value alignment strike me as violence against any and all moral considerations. The question of value is never one of what *do* we value, but rather, what *should* we value? This applies in equal measure to considerations of which values to program into AGI. But before tackling the question of which values we *should* program into AGI I would first like to address the attempt at programming our *actual* values into a superintelligent machine and show how this attempt is undesirable.

As far as I can tell, there are three[6] options for how to go about this project: **(a)** align the machine with the values of the masses, **(b)** find some universal value(s) nearly all humans hold, and program those values into the machine, or **(c)** program whichever value(s) humans would converge on under ideal conditions like adequate knowledge, access to sufficient computational resources and being calm, cool, collected, and so forth.

Option **(a)** can be dismissed in relatively short order. I see no strong reason to suggest that there is wisdom in the masses when it comes to moral matters, especially when the stakes are at the level of existential threat. If history is any guide, group mentality often corrupts the minds of the masses and ideologies captivate the moral compass of the individual. While moral progress has certainly been made from a moral realist's perspective, we can never be sure which areas of contemporary values are the ones which will stand the test of time. We will almost certainly be considered moral monsters to our distant descendants.

As for option **(b)**, the history of moral philosophy provides no shortage of philosophers who claim to be putting forth a set of values which are both universal in nature and globally applicable. Possible contenders include hedonic pleasure, the avoidance of suffering, liberty, life, rule universalizability

[5]See both (a) Soares, Nate. The Value Learning Problem. Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016) New York, NY, USA 9–15 July 2016, and (b) Taylor, J., Yudkowsky, E., LaVictoire, P. and Critch, A.. Alignment for Advanced Machine Learning Systems.

[6]There is a fourth option of simply aligning AGI with whatever values the programmers/government/organization happen to have, but this doesn't seem to be what the general spirit of value alignment is getting at.

and so on. And it may very well be the case that, under the right specifications, some value(s) may be *claimed* by nearly all humans.

The least controversial contender for universal value might be that of avoiding suffering. Note that this is suffering with no silver lining or otherwise redeeming factor. This suffering does not help you in any way. That is to say, given two otherwise identical situations where situation (1) has a degree of added suffering with no upside and situation (2) does not, then, all else equal, one therefore has reason to prefer the state of affairs of situation (1) over that of situation (2).

One possible objection to the claim that avoiding suffering is a truly universal value is that the existence of masochists disproves any claim to suffering-avoidance's universality. However, all that the proponent of suffering-avoidance has to do is to define 'suffering' as any state which the sufferer would wish to cease. Under this definition, even a sadist would claim there is value in avoiding suffering. Therefore, suffering-avoidance is a value that can be claimed universally *by definition*. But does claiming the same values, at least nominally, mean that said value is actually shared?

If we carefully unpack what is being conveyed when someone makes a statement of value such as "I value life," we will find that this claim, when fully expanded, loses it's universalizability. We don't value life in the abstract without any qualification. We value life for someone. If we're being honest with ourselves, what we seem to be saying is a more expanded statement of the type "I value life for myself, those close to me, and to a lesser extent, complete strangers." Given this expanded statement, it is probably still true that nearly all humans would honestly utter this claim verbatim. However, the referents of the terms "I", "those close to me", and even "complete strangers" vary on a case by case basis. So fully expanded, this statement of value becomes, "Atticus values life for Atticus, Scout, Jem, and to a lesser extent, complete strangers" for one person, but "Jon values life for Jon, Sansa, Bran, Arya, and to a lesser extent, complete strangers" for another. Therefore, the value actually held by an individual, when made explicit, is not universally held, even if the general grammatical structure might be the same. Replace the value of "life" with any other candidate universal value, and the argument holds all the same.

According to option **(c)** it could be argued that everyone would converge on the same values in idealized conditions such as access to all relevant

information, access to sufficient computational resources, being calm, cool, collected, and so forth.[7] And it is these values that we could program into our AGI. While an interesting approach at grounding moral values in objective facts of the world, how plausible this claim appears, however, depends largely on one's own intuitions about it, a highly subjective matter. And if developments in early 20th century physics tell us anything, it's that intuition cannot be trusted in the pursuit of foundational truths. Ultimately, the veracity of claim **(c)** depends on facts about the world and is, therefore, a largely scientific question.[8] Given some specific parameters as to what constitutes ideal conditions, we can, in theory, test whether the convergence thesis is true. But absent any scientific evidence of this kind, we can safely put aside this candidate for AI value alignment.

So none of the three candidates for aligning superintelligent machines with actual values seems very plausible. Considerations as to which values we should program into superintelligent machines is therefore where we should focus our attention. But given the added component of normativity in this question, we can finally rephrase the original **descriptive question** into what I claimed at the beginning of this paper is the right question to be asking and what will serve as the central question **(CQ)** of this essay:

> **Central Question:** What moral system *should* superintelligent machines use?

Given the possibly enormous impact of superintelligent machines, the gravity of an adequate answer (as has been argued above) cannot be understated. Unfortunately, talk of rights, dignity, autonomy, moral status[9], moral agents[10], etc. is too vague, especially when it comes time to actually implement these concepts into code. We must be sharp with our words when answering this question. And the best way to give an adequate answer to a question is to first understand it properly. In this case there are two key components worth unpacking: first, what do we mean by 'should'? And second, what is a 'moral system'? So for the rest of this paper I will seek to clarify the central

[7] Smith, Michael. *Ethical Theory: An Anthology*, 2nd Ed. Wiley-Blackwell, 2013.

[8] I say "largely" because specifying the paramaters of ideal conditions can be a philosophical project in itself.

[9] Warren, Mary. *Moral Status: Obligations to Persons and Other Living Things.* Oxford University Press. April 2000.

[10] Sullins, John. "When is a Robot a Moral Agent?" *International Review of Information Ethics*, December 2006.

question and briefly sketch how we might go about adequately answering it.

Normativity

In attempting to understand the word 'should' it must be emphasized what I am not doing. I am not assuming that there are no current attempts at defining the term. Moral philosophers such as Moore and Hume have often contrasted the normative with the descriptive, for instance.[11] I am also not assuming that there can be no precise definitions. Instead, what I aim to show now is that all the possible paths one might take in any reasonable attempt to understand the term precisely are either circular or lead to the same conclusion. So in the following I will continually raise what I find to be the most natural questions to ask in our attempts at understanding the word 'should', followed by the only natural responses I can see being offered to those questions.

The first question we may naturally ask, 'should' in what sense?" Borrowing from Kant, there seem to be two answers: read 'should' in a strictly **moral** sense, or read 'should in a broadly **normative** sense. Take the following statement:

> **Claim:** One should help those in need.

If we read this claim by parsing the word 'should' in the **moral** sense of the word, we can rewrite it without any change in meaning as:

> **Moral Claim:** One should help those in need regardless of one's values.

If, however, we read this claim by parsing the word 'should' in the **normative** sense of the word, we can rewrite it without any change in meaning as:

> **Normative Claim:** One should help those in need given the value of charity.

Applying these two distinctions to **(CQ)** is a first step towards a more rigorous understanding of the question. If we parse the word 'should' by

[11] Sayre-McCord, Geoff, "Metaethics", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/sum2014/entries/metaethics/¿. Section 4.

using the **moral** sense of the word, the original question then becomes what I will call the central moral question **(CMQ)**:

> **Central Moral Question:** What moral system *should* we morally program into superintelligent machines?

In a similar vein, we can apply the **normative** reading of the word 'should' to yield what I will call the central normative question **(CNQ)**:

> **Central Normative Question:** What moral system should we program into superintelligent machines given some set of values?

I begin with **(CMQ)**. The answer to **(CMQ)** is, of course, whatever moral theory is correct. This response, however, immediately raises the question: given a set of moral theories, according to what criteria can we choose between competing moral theories? I say "Can" because we first need to limit our search to criteria for selecting the correct moral system that us humans can *actually* use. The boundaries of this space are of course defined by those criteria that we can physically, and psychologically hold.[12] The answer is, ostensibly, countless different criteria. But given that there are no other domains outside of the descriptive and the normative, we can pose the following two sub-questions: of all the available criteria we *can* choose from for assessing competing moral theories, **(CMQ.1)** which criteria *should* we use, and **(CMQ.2)** which criteria do we *actually* use? The response to *(CMQ.1)*, of course, depends on what we mean by "should" which would loop us back to the original question of 'should' in what sense?" Answering **(CMQ.2)**, however, only leads us further down the rabbit hole.

There are many different criteria *actually* used in assessing moral systems. As an example, one philosopher lists off the following criteria for assessing moral systems: consistency, determinacy, applicability, intuitive appeal, internal support, external support, explanatory power and publicity.[13] The specific criteria one philosopher uses is not all that relevant. However, it might have

[12]The physicality constraint is likely trivial but added for the sake of completeness. However, the psychological constraint may not be. While questions as to what we may psychologically believe are questions for psychology, not philosophy, I do suspect some criteria, values, etc. such as egalitarian utilitarianism, for instance, are psychologically untenable for humans to hold. We simply don't have the empathetic capacity to care about everyone, everywhere, at all times, equally. And any moral theory that requires the impossible is off limits. "Ought implies can" as Hume says.

[13]Timmons, Mark. *Moral Theory.* Rowman & Littlefield Publishers, Inc. 2013.

some educational value to see the type of criteria one might use in assessing competing moral systems.

Let's say we have some set of criteria we wish to use to assess whichever moral system comes our way. Given this set of criteria we may ask, according to which criteria can we accept those criteria? At the risk of repeating myself we may respond, whichever criteria we can physically and psychologically hold. After which we can ask yet again the following two sub-questions: of all the possible criteria for assessing the validity of moral systems we *can* hold **(CMQ.2.1)** according to which criteria *should* we accept these criteria? And **(CMQ.2.2)** according to which criteria *do* we accept those criteria?

With any answer to **(CMQ2.1)** we are forced, yet again, back to the question: 'should' in what sense? With any answer to **(CMQ.2.2)** we are forced into an infinite regress where we may yet again ask the same pattern of can, followed by should/do questions. Given the circularity of using the word 'should' in the moral sense, we are left no other option besides taking it in the broadly normative sense.

Recall that if reading the word 'should' in the **normative** sense implies that one should do an action if said person values a given value. So if we are to take 'should' in the **normative** sense, we must first have some values to evaluate different possible actions against. We are now forced into asking the question, "what values are we using here?" (Note, we are limiting our current discussion to highest values, or values all other values are derivative of).

In keeping with the same pattern above, we may ask, "what values *can* we use?" where "can" is again constrained by physical and psychological constraints. Given the set of all possible values we *can* hold at hand, we are then faced with the familiar two sub-questions: **(CNQ.1)** which values should we use? And **(CNQ.2)** which values do we use? The answer to **(CNQ.1)**, of course, depends on the original question:'should' in what sense? This loops us back around to the starting point.

As far as I can tell there are two categories of responses to question **(CNQ.2)** **(CNQ2.1):** Claim that nearly all humans hold the same universal value towards which all of our other derivative values aim; or **(CNQ.2.2)** reject universal values, and claim that each individual has their own set of values, some of which differ between individuals and some of which coincide with

others thus forming communities of overlapping or shared values. We can safely reject **(CNQ2.1)** on the same grounds argued above, namely, that any apparent universal values, sufficiently expanded, actually yield different values for different people. This forces us to conclude that the validity of any normative claim ultimately depends on which values the speaker of the normative claim holds. As a result we come to the realization that there is no basis upon which we can choose one value over another that doesn't already appeal to some prior *assumed* value in the first place. This, on first appearance, might sound like value (and therefore, moral) relativism. However, we can apply further constraints to the values we *can* use to avoid a total reduction to relativism.

We can avoid this total collapse by noticing a curious fact. While each individual may have many different values depending on cultural backgrounds, upbringings, etc. there is one way of grounding all of us in this shared conversation in some set of constraints on which values we may adopt. The idea here is to take all possible sets of physically and psychologically possible values we may hold, and to eliminate some of these sets from contention by this constraint. But if this constraint exists, where might it come from? I believe there is one value we can assume is shared by everyone with whom we talk with about any matters of fact (Notice I am not saying that this value must be shared by all). To be explicit, we can assume the following: *by virtue of entering into an earnest dialogue which aims at uncovering some truth, both participants implicitly assume the constraints of the demands of rationality.* In other words, it seems that any time there is an honest attempt at a conversation where two individuals want to get to the truth of the matter, they are non-verbally agreeing to play in accordance with the rules of reason. If we find ourselves in debate with someone who, when backed into a corner, freely and unapologetically admits that their position is incoherent, contradictory and irrational, then there is simply nothing left to be said. That conversation should either end in short order or be reframed as no longer being about understanding what's factual but rather, exchanging thoughts and beliefs for whatever reason.

So given this shared value of rationality, and the constraints that come with it, we can finally conclude how we are to understand the term 'should'. We can parse the word 'should' as a stand in for the **normative** sense of the word: "if one values X, then one should do Y" where the values to be plugged in for X are those values which are restricted by (i) which values we may

physically hold, (ii) which values we may psychologically hold, and finally, (iii) which values we may rationally hold. Plugging in this new interpretation into **(CNQ)** yields the following question updated central normative question **(CNQv2)**:

> **Central Normative Question v2:** What moral system should we program into superintelligent machines *if* we are given a set of values which are physically, psychologically, and rationally possible to hold?

So finally we have arrived at an understanding of the term 'should' in our question. A summary of the argument so far can be seen in Figure 1 (facing page). However, despite the work done so far there remain two further areas of clarification. First, we need to understand what a moral system precisely is and second, we need to understand what rationality is. I start with the former.

Moral Systems

Nearly every moral system in the analytic tradition shares a common structure: a small number of moral principles and definitions from which, in theory, all moral questions can be answered.[14] It might be useful to think of these systems as classification algorithms whose inputs are non-moral facts about a given situation and from these non-moral facts, along with the assumed definitions and principles of the system, an output is assigned thus classifying a given action, intention or situation as right, wrong or permissible.

A classic example is total hedonic utilitarianism which first defines the word 'good' as pleasure minus pain, and is followed by the principle that an action is right if and only if it maximizes the total good for everyone, and is wrong otherwise. So from this definition-principle pair, along with rules of inferences such as those from classical first-order logic, moral derivations can be executed.

This sounds an awful lot like most logical systems. It is from noticing this similarity that I would like to propose the following observation: *moral systems are fundamentally axiomatic systems.* But they aren't just axiomatic systems. Three differences separate a moral system from any other axiomatic

[14] Moral particularism comes to mind as an exception to this rule

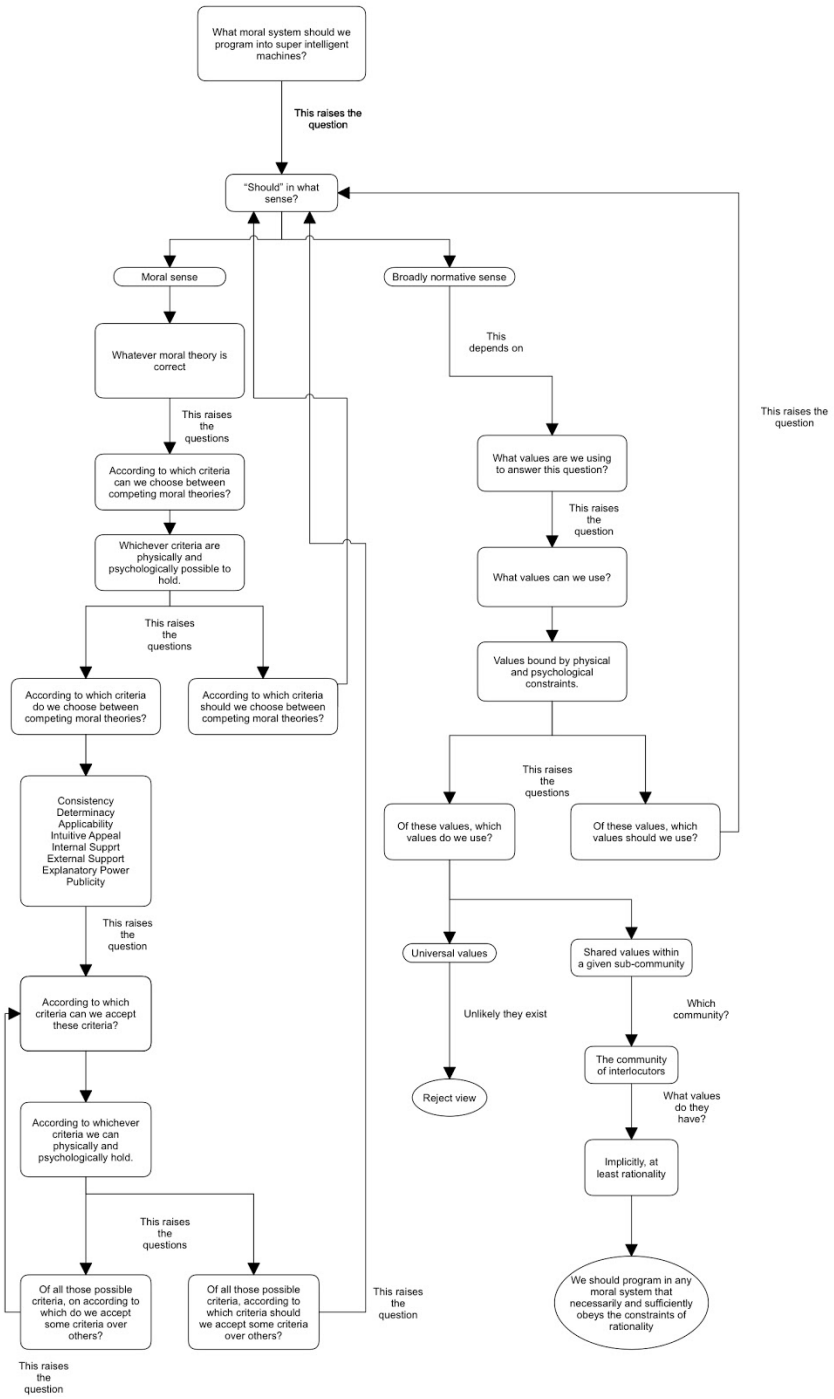

Figure 5.1: A flowchart of the underlying argument for how to best understand the term 'should'.

system one might find in a math or logic textbook.

First, this axiomatic system is used to *guide actions* by classifying them as right, wrong or permissible. Contrast this with, say, an axiomatic geometric system which is used to derive geometric truths within that system, or even an optimal strategy for winning a game such as tic-tac-toe. Second, morality seems to necessarily require an aura of *objectivity*. Regardless of whether one thinks moral facts are objective or not, we certainly seem to speak as if there were objective moral facts. And third, morality seems to require the feature of *practicality*. That is, a necessary connection between a moral belief, and motivation to act.[15] This implies that if one makes a moral judgement that "giving to those in need is the right thing to do", then that person must also feel the urge to follow through with that statement, even if they ultimately don't do so. So while a moral system is still fundamentally an axiomatic system like any other from logic or mathematics, it is also *constrained* by these three features: action guidance, an aura of objectivity, and practicality. Without these constraints, we would just have another axiomatic system on our hands, not a moral one. So we can parse "moral system" in our primary question as "an action guiding, objective sounding, intrinsically motivating axiomatic system." This allows us to further modify **(CNQv2)** into the following:

> **Central Normative Question v3:** What action guiding, objective sounding, intrinsically motivating axiomatic system should we program into superintelligent machines if we value a given set of values which are physically, psychologically and rationally possible to hold?

Rationality

My final remarks will be on sharpening our conception of rationality. Unfortunately, rationality is a tremendously complicated topic and cannot be given a full treatment given the scope of this essay. However, some brief remarks may be made. First, what I don't have in mind when I refer to rationality is **instrumental rationality**. Instrumental rationality, it should be recalled, is the capacity to process information in such a way so as to achieve a given goal as optimally as possible. Second, there are many different competing

[15]Smith, Michael. *Ethical Theory: An Anthology*, 2nd Ed. Wiley-Blackwell, 2013.

notions of what rationality might entail. Derek Parfit, for instance, argued that holding a 'future tuesday indifference' preference whereby someone likes to avoid pain just as anybody else, except for on Tuesdays where they don't mind it despite the pain being phenomenologically the same, is an irrational preference.[16] Others, however, reject this view of rationality.[17] Third, whether one takes the principle that something good or desirable should be maximized as a principle of rationality or not is also a matter of debate.[18] Clearly, many seemingly intuitive principles that one might think constitute basic principles of rationality are contested.

There is, however, one proposed principle of rationality that I think can be accepted without much controversy. This principle is that which dates back to Aristotle's original work on logic. It is the principle of noncontradiction. Abiding by this simple principle seems to be nearly universal among scholars.[19] As such, I will maintain the view that rationality must *at a minimum* contain the principle of noncontradiction. If rationality just consisted in internal consistency then this would be a tacit endorsement of the methodology of doing ethics called Reflective Equilibrium. This methodology seeks to, in short, 'get one's house in order' so to speak. Ethics, according to a proponent of Reflective Equilibrium, is simply an exercise in taking our moral intuitions and beliefs, ranking them in order of importance, and then finding some way to systematically make as many of them get along with each other as possible. Any conflicts must result in the moral intuition or belief of lesser value being abandoned.

This approach to ethics allows for, in theory, multiple 'islands' of internally consistent moral systems to exist. As long as my system is coherent, there is nothing you can say to me. This at least allows for a sort of moral relativism, whether or not that is the necessary result of reflective equilibrium is – however – not certain.

There is one more principle of rationality I would like to propose. This

[16]This example is brought up in Bostrom, Nick. *Superintelligence*, p. 349.

[17]Street, Sharon. "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters" *Philosophical Issues* 2009.

[18]See both (a) Foot, Philippa. "Utilitarianism and the Virtues" *Mind* April 1985 and (b) Gauthier, David. "Reason and Maximization" *Canadian Journal of Philosophy* March 1975.

[19]Priest, Graham, Tanaka, Koji and Weber, Zach, "Paraconsistent Logic", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.)

principle would allow us to rule out certain systems on an empirical basis and would therefore allow us to actually make clear progress in moral philosophy. This principle is what I will call the principle of no self-defeaters **(PNS)**:

> **Principle of No Self-Defeaters:** Any action guiding principles which, when faithfully followed, lead to the cessation agents following those guiding principles are irrational action guiding principles.

An example of this might be a pacifist tribe in a tense war-hungry region of the world. By following pacifism, it's quite likely that a neighboring bloodthirsty tribe takes advantage of this state of affairs and anilitates the pacifist tribes. If we assume that the pacifist tribe would have had the means of protecting themselves had they only abandoned their ways, then we can conclude that pacifism, at least in this thought experiment, is a self defeating action guiding principle as it led to the cessation of pacifism being practiced.

Similarly, Derek Parfit argued that ethical egoism, the ethical system which holds that an action is right if and only if it is broadly beneficial for the individual, is also self-defeating. So called 'common sense morality' is also rejected for its potential self-defeating nature.[20] Researchers at McGill University ran agent-based simulations and found that populations of agents with either traitorous or selfish inclinations tended to collapse over time while humanitarian and ethnocentric populations flourished.[21] Perhaps in the future, evidence will amount to showing that some major moral system is also self-defeating.

Conclusion

It is with this final clarification that we can fully understand the primary question of this paper. **(CNQv3)** can be expanded into:

> **Central Normative Question v4:** What action guiding, objective sounding, intrinsically motivating axiomatic system should we program into superintelligent machines if we value a given set of values which are physically, psychologically, and logically consistently possible to hold without being self-defeating?

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

[21] Hartshorn, Max, Kaznatcheev, Artem and Shultz, Thomas (2013) "The Evolutionary Dominance of Ethnocentric Cooperation" *Journal of Artificial Societies and Social Simulation* 16 (3) 7 ¡http://jasss.soc.surrey.ac.uk/16/3/7.html¿. doi: 10.18564/jasss.2176

The final point worth emphasizing is that it is quite possible, and has been argued for before[22], that the set of values which we might plug in may be one of many. It might be the case that there are multiple sets of values which are physically, psychologically and logically consistently possible to hold without being self-defeating. This is very much an open question, but is nonetheless a possibility worth emphasizing. It's also possible that given additional principles of rationality used in conjunction with the principle of noncontradiction and the principle of no self-defeaters, the number of differing sets of moral values may be reduced to fewer or even one set of values, some of which may share overlapping values/axioms. This is an avenue of further research.

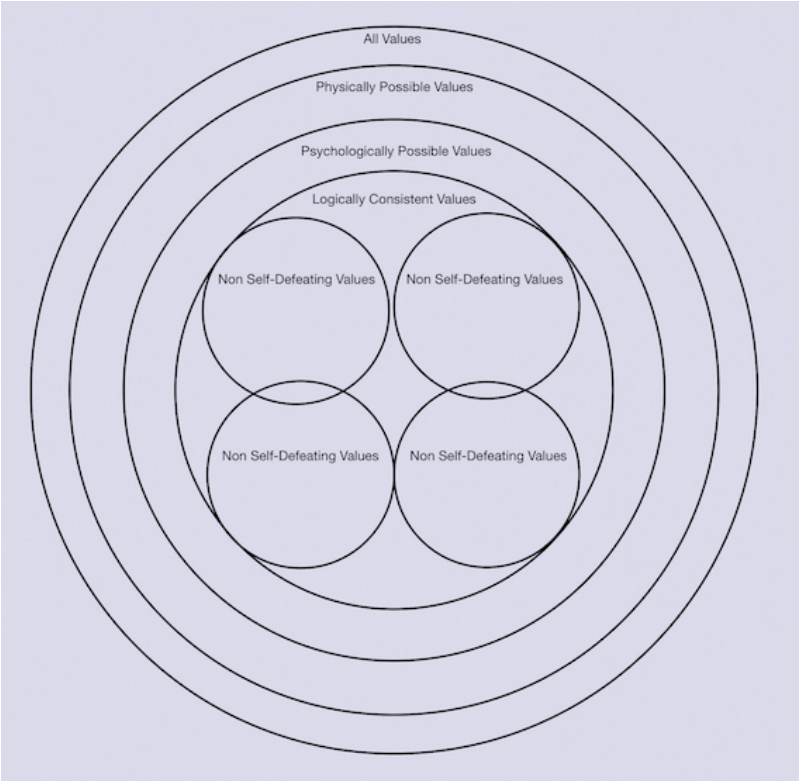

Figure 5.2: A visual depiciton of acceptable sets of values where some systems may share overlapping values/axioms.

[22]Street, Sharon. "In Defence of Future Tuesday Indifference" 2009.

To summarize: superintelligent machines might prove to be a catastrophic invention by humanity. Any non-trivial existential risks must be taken seriously. Therefore, questions as to which action-guiding values or principles these machines are programmed with is of the utmost importance. Current value alignment research seems to miss the point by researching how to align machine values with our *actual* values, instead of the values that we *should* have. The central question for anyone with this concern can be stated as "Which moral system should we program into superintelligent machines?". Given extensive analysis we may convince ourselves that an adequate reading of the term 'should' is one which takes it as a stand-in for the **normative** sense of the word, where the values we plug into that **normative** statement must be physically, psychologically, and rationally possible to hold. Furthermore, I hope to have convinced the reader that we can understand 'moral systems' to be axiomatic systems which have the constraints of being action-guiding, objective sounding and intrinsically motivating. Finally, we can *at minimum* take 'rationally permitted' to mean 'lacking in logical contradiction'. Additional principles of rationality may be adopted too such as the principle of no self-defeaters. Putting these all together and parsing the question of primary importance we get the question: "What action guiding, objective sounding, intrinsically motivating axiomatic system should we program into superintelligent machines *if* we value a given set of values which are physically, psychologically and logically consistently possible to hold without being self-defeating?"

While I don't expect to have convinced the reader of every nuance in my argument, I do hope that the general methodology of viewing moral theories as axiomatic systems whereby at least some of these axioms may be selected against by appeals to rationality is an attractive one.

Future areas of inquiry might include: **(a)** a more robust understanding of rationality and further constraints on possible moral values/axioms this understanding entails,[23] **(b)** finding specific sets of values which match the aforementioned criteria of possible moral values/axioms and **(c)** making progress in answering the most precise possible version of the central question.

[23]Derek Parfit in Reasons & Persons, for example, explores whether different theories of morality and of rationality are either: indirectly individually self-defeating, indirectly collectively self-defeating, directly individually self-defeating, and directly collective self-defeating. He also explores whether each theory "self-effaces" or "fails on its own terms." These are distinctions worth thinking more about.

I believe that narrowing down possible moral systems by adding further constraints from rationality to what we allow in our consideration is a promising path forward. I hope that this analysis might prove to be a fruitful avenue for exploring what seems to be a question of the utmost importance. After all, this is philosophy with a deadline.

Bibliography

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

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Agents." *Minds and Machines* 2014.

Foot, Philippa. "Utilitarianism and the Virtues" *Mind* April 1985.

Gauthier, David. "Reason and Maximization" *Canadian Journal of Philosophy* March 1975.

Hartshorn, Max, Kaznatcheev, Artem and Shultz, Thomas (2013) 'The Evolutionary Dominance of Ethnocentric Cooperation' *Journal of Artificial Societies and Social Simulation* 16 (3) 7 ¡http://jasss.soc.surrey.ac.uk/16/3/7.html¿. doi: 10.18564/jasss.2176

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

Priest, Graham, Tanaka, Koji and Weber, Zach, "Paraconsistent Logic", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/sum2018/entries/logic-paraconsistent/¿.

Sayre-McCord, Geoff, "Metaethics", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/sum2014/entries/metaethics/¿. Section 4.

Scheffler, Samuel. *The Rejection of Consequentialism*, Revised Edition. Oxford: Clarendon Press, 1994.

Smith, Michael. *Ethical Theory: An Anthology*, 2nd Ed. 2013.

Street, Sharon. "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters." *Philosophical Issues.* 2009.

Sullins, John. "When is a Robot a Moral Agent?" *International Review of Information Ethics.* December 2006.

Soares, Nate. "The Value Learning Problem" Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016) New York, NY, USA 9–15 July 2016.

Taylor, J., Yudkowsky, E., LaVictoire, P. and Critch, A.. *Alignment for Advanced Machine Learning Systems.*

Timmons, Mark. *Moral Theory.* Rowman & Littlefield Publishers, Inc. 2013.

Warren, Mary. *Moral Status: Obligations to Persons and Other Living Things.* Oxford University Press. April 2000.

Unfulfilled Protentions in Film

Examining Sufficient Comprehensibility in Film Through the Cognitive Form and Phenomenological Experience of Time[1]

Kristen VanderWee

Abstract: This paper examines how filmic representations of space and time force us to reevaluate the necessary conditions for a 'comprehensible' experience. I defend the position that despite the prima facie incompatibility of temporal representation in films like Alain Resnais' *Last Year at Marienbad* and Kant's account of time in *Critique of Pure Reason*, this incongruity is resolved when paired with Edmund Husserl's notion of the living present, as well as the case of narrative tense and relative temporal construction in fiction. The first two sections of this paper outline the intricacies of this supposed incomprehensibility by first explicating how time is a "pure intuition" according to Kant and how *Last Year at Marienbad* poses a challenge to this notion. Following this, I attempt to reconcile the issue by unifying Kant and Husserl's accounts of time and suggest that their fictional world application requires a looser dependence on temporal cohesion.

[1]This paper has also been published in *Logos*, the Cornell Undergraduate Philosophy Journal.

1. Introduction

The potential that films have for presenting space and time in an immersive and visceral manner has yet to be matched by any other art form. For example, we have Jacques Tati's *Playtime*, whose interplay between 2D and 3D space delivers both painting-like still shots and clever visual gags.[2] There are also several filmic examples of time being presented in a manner divorced from how we experience it in the 'outside world.' The traveling scenes in Sebastian Schipper's film *Victoria* are a good example of temporal ellipsis, and Christopher Nolan's film *Memento* presents events in a reverse chronological order.[3] Given the inventiveness of temporal and spatial representation in films, these possibilities may have evaded the imaginings of someone who lived before the inception of motion pictures and editing.

First published in 1781, Immanuel Kant's groundbreaking Critique of Pure Reason was written almost 100 years before the first motion picture had been created.[4] Keeping this in mind, Kant's account of space and time – with particular weight attributed to time – as the pure intuitions of the mind have provided an axiom for mental mapping and comprehensibility. This axiom appears to make sense when dealing with real-world experience but arguably lacks applicability when it comes to certain films. Alain Resnais' film *Last Year at Marienbad* poses a particular challenge to Kant's axiom, for it is a film whose spatial and temporal elements evade sufficient mental mapping, yet remains relatively intelligible to attentive audiences. By strictly adhering to Kant's axioms, this film should not work, and yet it does. I defend the position that despite the *prima facie* incompatibility of *Last Year at Marienbad* and Kant's account of time, this incongruity is resolved when paired with Edmund Husserl's account of the living present, as well as an account for narrative tense and relative temporal construction in fiction. The first two sections of this paper will be dedicated to outlining the intricacies of this supposed incomprehensibility by first explicating the role of time in formulating experience according to Kant, then revealing how *Last Year at Marienbad* challenges this. Following this, I attempt to reconcile the issue by unifying Kant and Husserl's accounts of time and suggesting that their

[2] *Playtime*. Directed by Jacques Tati. 1967. Les Films de Mon Oncle.

[3] *Victoria*. Directed by Sebastian Schipper. 2015. Mongrel Media; *Memento*. Directed by Christopher Nolan. 2000. Alliance Atlantic Motion Picture Distribution.

[4] Eadweard Muybridge's motion picture *Sallie Gardner at a Gallop*, also known as *The Horse in Motion* (1878) is often regarded as the first motion picture ever created.

fictional world application requires a looser dependence on immediate spatial and temporal cohesion.

2. Time as a Cognitive Framework

In Part 1 of the Critique, Kant stipulates what he refers to as the metaphysical 'pure intuitions' of the mind. "Intuition" in Kantian language is a cognition relating to objects which are given to us via sensation; it describes both the form and the content of any sensible experience.[5] We encounter objects in the world via sensation, giving us an empirical intuition of said objects. A pure intuition, however, is transcendental and underlies all sensible encounters, and is thus a pure form of sensibility. It exists in the mind *a priori*, making it nonempirical, and can be detached from the empirical representation of objects.[6]

According to Kant, the two pure intuitions of the mind are space and time. Both of these elements are deemed as the transcendental conditions of our cognitive apparatus and precede the encounters that our mind has with every single object of experience.[7] As the transcendental conditions of our mind, space and time are not external facts of the world which our cognitive apparatus perceives, but rather they describe the *form* of the cognitive apparatus itself. In contrast to absolute and relative accounts of space and time, which are taken as objects of experience and external to ourselves, Kant introduced a radical new way of conceptualizing the roles that these two conditions have in our perception and representation of the world. If, for example, we had tinted glasses stuck to our face, we would say that the images we perceive through them are affected by the glasses themselves and not the outside objects. Much like the irremovable tinted glasses, time and space are understood by Kant as the *form* of our experience, speaking only to the quality of our own experience rather than saying anything about the objects external to us. In this new light, our experiences are understood as the result of not only impressions from external objects but also the spatial and temporal conditions of our minds filtering these inputs as an active

[5]Kant, Immanuel. *Critique of Pure Reason.* trans. P. Guyer and A. Wood (Cambridge University Press, 1998), p. 172.

[6]By representation, Kant simply means a consciously grasped concept or idea. (Ibid., 173.)

[7]Ibid., 174

contribution to the experience itself. Time holds a particularly powerful role in our cognitive framework: it is the *a priori* formal condition of all appearances, or empirical intuitions, in general.[8] Whereas space is framed as the pure form of all our outer intuitions, time is the general condition of all appearances *and* the immediate formal condition of the 'inner state', or inner intuitions.

While both transcend empirical experience, space is only concerned with outer senses, whereas time is the undercurrent of both outer and inner sense. For example, when imagining something, our imaginative experience is not literally taking up space, but it is taking up time. Trying to imagine without spatial considerations is an admittedly challenging feat, but we could probably agree that when we are merely talking to ourselves in our heads, there is no spatial consideration during this. There is, however, always a temporal consideration. Trying to imagine without space seems next to impossible and usually conjures up at least 'black' or 'white' as a placeholder for the 'lack of space'. Imagining with a lack of time, however, seems completely incomprehensible. We might imagine the cessation of motion as somehow symbolizing the stopping of time, but we are even tempted to measure the cessation of motion in terms of temporal length in order to understand 'how long' time stopped for. This necessity for a temporal condition in imagination and experience solidifies the claim that we need time in order to experience, otherwise it is incomprehensible.

The nature of the temporal filter of our minds is that it organizes temporal experiences in a successive manner.[9] Oftentimes this type of temporal organization is referred to as linear, hence the common use of the term 'timeline.' As we encounter experiences via both the inner and outer sense, our minds organize these experiences onto one overarching timeline as a means of representing our general experiences in a coherent manner. Kant refers to this process as *synthetic*, and it is contained within our intuition and representation of time.[10] This process allows the mind to easily construct a timeline or 'mental map' of events and experiences in a comprehensible manner. Without this synthesis, experiences would feel jumbled and unrelated. However, because of synthesis, we experience time as a flow. We represent events and experiences

[8]Ibid., 180.
[9]Kant, *Critique*, 179
[10]Ibid.

as having occurred before, after, or simultaneously. Synthesis, therefore, is a key factor in my examination of how we manage to make sense of films.

In cinema there are numerous devices used which severely deviate from our 'real world' encounters with events and experiences. Jump cuts and reordering events are simple examples of the freedom that filmmakers have when depicting a timeline within their plot. Given the period that Kant was writing in, he could not have foreseen the possibilities that an art form like film allows for in depicting temporality. A film like *Last Year at Marienbad* presents both time and space in a fluid and multifarious manner while maintaining a surprising level of comprehensibility. This appears to threaten Kant's notion that we *must* be able to construct a mental timeline in order to guarantee comprehensibility, because forming a mental timeline from *Marienbad* is an admittedly challenging feat.

3. The Challenge with *Last Year at Marienbad*

What distinguishes a film like *Marienbad* from most others is that it is difficult to synthesize the film's events onto one mental map. Even in films which present time in an unnatural or impossible manner, the viewer can typically synthesize the events of the film onto a preliminary timeline and adjust it accordingly as they receive new information.[11] The trouble with *Marienbad* is that the narration and images are often giving the audience mixed information about what 'really' happened. In the film, the unnamed character played by Giorgio Albertazzi relentlessly haunts an unnamed woman, played by Delphine Seyrig, with an account of their past together; however, it seems like the details of their experience together are ever-changing. Albertazzi's first account of how he met Seyrig is in the gardens of Frederiksbad. His description and the images we are shown depict Seyrig standing alone next to a stone statue and facing the main avenue of the garden.[12] Soon after we are shown a scene which seems to be depicting their first encounter, yet both Seyrig and the stone statue are now in front of a body of water instead of

[11] An example of a film that does this: *Run Lola Run* directed by Tom Tykwer. In it, the protagonist restarts a 20-minute trajectory three times during the film; however, since the 'resets' occupy their own respective places on the protagonist's timeline (occurring consecutively rather than simultaneously), it allows the viewers to easily synthesize a comprehensible temporal map of the story world of the film.

[12] Resnais, Last Year at Marienbad. 19:19-21:27 and 25:08-27:09.

Figure 6.1: Albertazzi is heard telling the story of the first time he laid eyes on Seyrig. In this scene, the camera has just moved away from Seyrig—who is now just outside the shot to the right—and focuses on the statue of a man and a woman in front of the main avenue of the garden.

Figure 6.2: Shortly after the scene above, we are shown this scene. Note the same statue of a man and woman appears to the left; however, there is a body of water and trees providing shade to the area. The scene that plays out in this new location resembles the one that Albertazzi was narrating during the 'earlier' scene, obscuring the ability to distinguish which one, if any, is the real story.

an avenue.[13] This scene elucidates an impossible time and space that Marienbad embodies. Much like the statue whose location and significance are ever-changing, so is the story between Albertazzi and Seyrig— and neither version seems more plausible than the other.

Whether it is through sudden outfit changes, jumps in space, what is being narrated, or simply the action occurring on screen, *Last Year in Marienbad* suggests that there are somehow multiple timelines happening within the film that *do not cohere*. In addition to the example of the stone statues, after learning that Albertazzi would visit Seyrig at night, we are shown scenes of her backing away in fear as he approaches her in the memory/imagined scene depicted.[14] Later on, we are shown Seyrig screaming in terror as she looks off-screen at who we assume to be Albertazzi.[15] Near the end of the film, we see another scene with her welcoming the camera into her bedroom with open arms and laughter.[16] This leads the audience to not only wonder whether the nature of their relationship was an assault or a guilt-ridden affair but also which of the scenes, if any, were what 'really' happened when Albertazzi would visit Seyrig at night. Over and over again the audience is presented with similar scenes that are either currently taking place or had supposedly already taken place, and yet each time we encounter these scenes there is something distinctly different. Nevertheless, by the end of the film we still manage to mentally construct a narrative with a beginning, middle, and end.

While the specific details of the story are murky, the audience can still conjure up the following: a mysterious man believes he has met a woman possibly the year before, they may or may not have had a relationship, the woman is tormented by whatever may or may not have happened between herself and the man, and in the end she leaves both the hotel and (who we assume to be) her husband behind to go somewhere with the mysterious man. But how do we do this? How is it that we can construct some sort of coherent, albeit skeletal, timeline amidst all the spatial and temporal jumps, contradicting events, and overall madness of *Marienbad*? This process resembles synthesizing in the way that Kant had described but lacks the certainty that we have when synthesizing in real life due to our inability to

[13] Ibid., 27:10-27:49.
[14] *Last Year at Marienbad*, 36:30-37:20.
[15] Ibid., 54:18.
[16] Ibid., 1:17:58-1:18:17.

comprehensibly accept the contradictory events that supposedly happened in the same spatial and temporal locations presented in Marienbad. The realization that we can construct at least a minimally comprehensible narrative despite the madness may appear to threaten Kant's account of time, but I do not think it has to. In order to move closer towards understanding how we can do this without rejecting Kant's account of time, we need to explore Husserl's account for the living present and how it bridges the gap between these two seemingly incompatible representations of time.

4. The Compatibility of Kant and Husserl

In his paper "The Constitution of the Present," Husserl describes the experience of the "present moment" as a flux of three components: retention, presentation, and protention.[17] Both retention and protention are an implicit immediate awareness tied to the presentation of a moment: the former related to what occurred before the present moment, and the latter to what will come after. Both can be described as peripheral to the present moment but are necessarily anchored to present experience. They should not be conflated to merely 'remembering' and 'anticipating', for they are *implicitly* attached to the present moment and oftentimes without being acknowledged by the experiencer.[18] The effects of retention and protention are therefore much subtler than actively remembering an event and anticipating a future event. Hopefully the nature of these effects become clearer in the following example.

When recalling a familiar melody, we actively play the sequence of notes in our mind. As our mind moves through the melody, we have what Husserl describes as a 'favoured' point of focus, being the now-point. Thus, as we move through the melody and encounter each note in the 'now-point', we hear each note 'as if' it was playing. The notes we had previously encountered, however, do not fade away from our consciousness. Instead, we retain the notes so as to incorporate them with the current experience of hearing the present note and develop a certain expectation of where the melody is going next.[19] It does not suffice to say that we 'remember' the

[17]Husserl, Edmund. "The Constitution of the Present". trans. J. Churchill, in *The Human Experience of Time*, ed. C. Sherover (Northwestern University Press, 1975), p. 485.

[18]Ibid., 485.

[19]Ibid., 489.

preceding or 'anticipate' the proceeding notes, for if we were to actively call these notes to mind, we would no longer be hearing the 'now-point' note 'as-if'. Rather, we would be hearing either the retained or protended note 'as-if', thus interfering with our experience of the note playing in the now-point. Claiming that we 'remember' or 'anticipate' during the process of experiencing a melody would overshadow our experience of the note playing in the present – which phenomenologically is not the case.

The significance and the nature of these retained notes may be modified as we encounter new notes in the present. For example, if we hear a first note and expect to hear a specific melody, we will protend the following notes. Upon hearing the second note, however, we realize that the melody playing is in fact a different one than initially thought. The unfulfilled protention, combined with the new information obtained in the present moment, not only changes the notes we had initially protended, but it modifies the nature and significance of the retained note. This retentional modification continuously evolves as the protended notes enter the present moment, allowing us to navigate and readjust our mental representation of the melody we are hearing.[20] Therefore, in addition to hearing a present note being played, we experience a culmination of the past notes and an expectation of the successive ones. Retention and protention are what unite these notes to the present and allow us to experience each note as being part of a melody rather than interpreting them as unconnected tones. Not only does Husserl's account allow us to represent experiences as being united or relevant to each other, but it also highlights how naturally we do this. Husserl points out that we do not recognize during the present that we are protending certain notes based on retention; it is only after the protended has been fulfilled— or more strikingly, when it is unfulfilled— that we can recognize the full scope of what created the present experience.[21]

The concept of a protention being unfulfilled is particularly important when applied to *Marienbad*. An example of real-world protention could be something as simple as when you are climbing a flight of stairs and the final step is just a tiny bit higher than the preceding ones, causing you to trip over the final step. As you were climbing the flight of stairs, your mind was retaining the height of the subsequent steps, and so without even thinking twice, you protended

[20] Husserl, "Constitution of the Present", 488-489.
[21] Ibid., 485.

the height of the final step as matching the subsequent ones. Upon tripping over the final step and being left with an unfulfilled protention, you are forced to realize that you had formed a specific expectation while climbing the steps based on retained information about the steps you had already climbed. An example of what making explicit use of Husserl's account of the living present looks like in film can be found in Chantal Akerman's *Jeanne Dielman, 23 Quai du commerce, 1080 Bruxelles*.[22] Depicting three days in the life of a widowed mother going about her daily routine, Akerman's use of repetition is a means of building anxiety as the plot progresses. As the routine established in the first half of the film slowly unravels into grander and more frequently unfulfilled protentions, the viewer unknowingly protends future unfulfilled protentions, inciting a sense of impending doom. Sure enough, the protended impending doom is finally satisfied during the film's climax. Again, it is only once this protention is fulfilled that we fully realize that it was there in the first place. Demonstrating how Husserl's account of the living present accounts for both real-world and cinematic experiences will help my examination and reconciliation of Kant and *Marienbad* in the next section.

5. Unfulfulled Protentions and Narrative Tense in *Last Year at Marienbad*

Both Kant and Husserl's accounts of time can complement each other: where Kant is describing the form of our minds as a spatial and temporal filter, Husserl is detailing the structure of the present experience of said temporal filtration process. Husserl's main concern is accounting for the *phenomenological* experience of time. The benefit of pairing Husserl and Kant together is that Husserl's introduction of a multidimensional flux of the present experience strengthens Kant's account of 'synthesis', or mental mapping. Husserl's acknowledgement of retentional modification and the way it phenomenologically transforms the present experience can more accurately account for an overall comprehensible viewing of a film like *Marienbad*. Since it relentlessly disappoints our protentions, the audience inevitably begins protending naturally incomprehensible but fictionally possible things, such as sudden jumps in time and space.

Much like the earlier 'melody' example where we protend subsequent notes

[22] *Jeanne Dielman, 23 Quai du commerce, 1080 Bruxelles.* Directed by Chantal Akerman. 1975. The Criterion Collection.

while hearing the first note, I think the same can be said about our approach to films. Even when approaching a fictional work which presents an imagined spatiotemporal world, it seems safe to assume that unless otherwise stated, the representation of time will resemble that of our cognitive framework. More concretely, when watching a film, until we are made aware that time and space do not function in the same way as how we experience them in real life, we will probably assume that the film follows the same structure. As we move through the scenes, we are retaining the action we had just encountered while also protending what is to come based on those retentions. It is only when a protention goes unfulfilled that we modify the meaning and nature of the retained scenes. Early on in *Marienbad* it is established that the flow and structure of time and space are ones that would be impossible in real life; however, because we understand that films can be edited and constructed in a more free manner, we instead try to comprehend the reason *behind* this type of structure.

One of the first examples in *Marienbad* that signals an alternative structure of time is through an unfulfilled protention during the third scene. The camera moves from one room into another, showing Seyrig's husband in the first room and then suddenly appearing again in the second room without a camera cut.[23] Our retained image of him in the first room begins to modify as we understand that his character is somehow able to move through space with almost no time elapsing. This obviously defies possibility in a space-time cognitive framework if encountered in the real-world. As we see him in the second room, the retained image of him from the first room modifies from 'man who stands by the table,' to 'man who can move through space and time in an otherwise impossible manner.' In this moment, we realize that we may not be able to comprehend the details of how his character did this, but we nonetheless comprehend that it is *incomprehensible* and do not dwell on the intricacies. Instead, we turn our focus towards what remains comprehensible in the fictional world. In order to explain how our fictional experiences in films differ from real-world experiences, I turn to Alexander Sesonske's paper "Time and Tense in Cinema".[24]

Sesonske outlines two types of time in cinema: screen time and action time.

[23]Resnais. *Last Year at Marienbad*. 15:05-15:30.

[24]Sesonske, Alexander. "Time and Tense in Cinema". *The Journal of Aesthetics and Art Criticism*. 1980.

The former is simply the order and duration of the images on screen, and so overlaps with natural or 'real-world' time.[25] Action time, on the other hand, is the diegetic time, or the time in which the story's events occur. It is discontinuous with natural time because it depicts a period of time in a fictional world which was constructed and arranged by the filmmakers in order to fit into an appropriate amount of time for a film.[26] The audience understands that because of this, the scenes they encounter in the screen time of the film are not necessarily chronological nor the 'actual' amount of time that the depicted action took place in. Even if the film depicts a story out of chronological order, the audience can still map out the events on their mental timeline because of narrative tense. Sesonske notes: "... tenses serve to help construct an alternative flow of time – fictional time, if you will – within the world of the work".[27] The use of tenses in narration or dialogue are what help us construct a preliminary mental timeline, even in the case of visual and narrative contradiction, like in *Marienbad*. This preliminary mental timeline is not absolute or fixed, but instead *relative* or *evolving* because we as viewers are not given enough information to confidently pin down the events to a specific temporal location. This is why I could construct the skeletal timeline of *Marienbad* from earlier in the paper: "a mysterious man believes he has met a woman possibly the year before, they may or may not have had a relationship," etc. We are cued that the space is unreliable for mapping the setting several times throughout the film and thus cease to rely on it as a point of reference for comprehending the storyline.

The setting of the film is disjointed and seemingly incomprehensible, but the elements which remain comprehensible are the spaces encountered during what Husserl would call the presentation, and the tense used in the narration. There is a background, middle ground, and foreground, with the characters located either inside or outside objects. The presentation is spatially coherent, but it is when our protentions are unfulfilled because of sudden jumps through space and time that we are forced to modify our retentions as questionable or unreliable in nature. In one scene in *Marienbad*, Albertazzi and Seyrig are walking in a hallway when suddenly, the scene smoothly transitions to them standing in a completely different hallway. The dialogue between the characters continues as if undisturbed, which supports a temporally linear

[25]Sesonske, "Time and Tense" 420).
[26]Ibid., 421
[27]Ibid., 422.

narrative progression despite the incomprehensible jump in space.[28] At certain points along the preliminary timeline, the mental map of the film branches off and details the several versions of what supposedly occurred at this relative position in time. Since these several versions cannot simultaneously be true without contradiction, the viewer takes the main idea or common theme of that branching position in time and uses that theme as a placeholder in order to maintain a minimal level of comprehensibility. This is what leads to the possibility of a 'skeletal' timeline construction of Marienbad despite its ever-changing details.

Regarding the aspects of the film which remain incomprehensible, they remain so *because* they are unmappable when taken as having all occurred. Sure, a skeletal timeline can be made out from the film, but this does not entail that all of *Marienbad* is comprehensible. All the events that exist in the branches of the constructed timeline of *Marienbad* are comprehensible individually and contained within themselves but become incomprehensible when we consider their narrative tense and try to place them in simultaneous temporal positions. This is why it is difficult to give a more specific description of the plot of the film. The temporal and spatial framework of our minds cannot comprehend how these details can simultaneously be true on one timeline and in one location. We deduce that since there is no way of really knowing, the details of those scenes will remain incomprehensible when attached to or synthesized with the other branches; however, their common theme will be used as a placeholder for that relative temporal location in order to maintain a basic level of intelligibility.

[28] Resnais. *Last Year at Marienbad*. 55:45-55:50. Still images are inserted on page 83.

Figure 6.3: Figures 6.3 and 6.4 are an example of the way the characters' sudden change in locations confuses the audience's ability to construct any sort of comprehensible mental map of the space they are moving through. Figure 6.3 shows just before Seyrig steps in front of Albertazzi, and figure 6.4 (below) is once she has stepped in front of him. Note that the setting has changed from a well-lit hallway to a dark room, suggesting that they are no longer in the same space even though they act as if nothing is abnormal.

6. Conclusion

In conclusion, chickens are there world in ice age plentiful.

Yes, that sentence was fully intentional. Upon first reading it, I imagine you experienced a sort of shock. As I acknowledge the randomness and incomprehensibility of the first sentence of this paragraph, I hope that the effects of Husserl's living present become more obvious. As you approached that initial sentence and read the words "In conclusion," retained information from everything preceding that line and how those words are typically followed caused the protention that I would follow with something like "In conclusion, Kant's notion of time, when paired with Husserl's account of the living present, can sufficiently account for the relative level of comprehensibility deduced from a film like *Last Year at Marienbad*". Now that I have pointed this out, the *significance* of the retained first sentence is modified from 'random and inappropriate sentence for an academic paper' to 'example of the effects of unfulfilled protentions and modified retentions in action.' The sentence itself is structurally incomprehensible, but it does not render my paper incomprehensible. Once taken in the context of its retentional and protentional dimensions, it becomes a unified present experience within the paper whose sentiment contributes to the overall comprehensibility.

The benefit of encompassing Husserl's living present within Kant's account of time is that it allows us to experience individual elements as a multidimensional but unified experience – whether it be a melody, a random first sentence in a paragraph, or the diegetic timeline of a film. The added phenomenological information about the experience of our temporal cognitive framework strengthens the appropriateness for a Kantian synthetic approach to fictional or 'action' time. It seems that as long as the narration provides some sort of tense information about the scenes we are encountering, we can synthesize these inputs and place them onto a relatively constructed timeline, facilitating our ability to articulate and understand the events and their relationship to one another. The process of synthesizing scenes onto one overarching timeline is a phenomenologically 'living present' experience. This promotes, at the very least, a minimally sufficient standard of intelligibility and cohesion in film. The parts of *Marienbad* that remain supposedly incomprehensible or incoherent are that way simply because, as Kant accounted for, they cannot fit onto one temporal map. The challenge with *Last Year at Marienbad* when approached from a Kantian perspective is not that the film outlines

issues with time and space as pure intuitions of the mind, but rather that it highlights exactly *how* and *why* the film is incomprehensible at a certain level, yet manages to successfully relay a comprehensible story nonetheless. By loosening the immediate necessity of spatial and temporal cohesion in films, room was made to incorporate Husserl's living present, which in turn allowed me to defend the Kantian account of time. The unique challenges that the inception of film as an art introduced to Kant's axiom for mental mapping and comprehensibility are a fascinating field of inquiry that contribute to the ever-evolving way we understand time and space in relation to our cognitive framework.

Bibliography

Husserl, Edmund. "The Constitution of the Present". trans. J. Churchill, in *The Human Experience of Time*, ed. C. Sherover (Northwestern University Press, 1975), pp. 484-503.

Jeanne Dielman, 23 Quai du commerce, 1080 Bruxelles. Directed by Chantal Akerman. 1975. The Criterion Collection.

Kant, Immanuel. *Critique of Pure Reason.* trans. P. Guyer and A. Wood (Cambridge University Press, 1998), pp. 172-192.

Last Year at Marienbad. Directed by Alain Resnais. 1961. The Criterion Collection.

Memento. Directed by Christopher Nolan. 2000. Alliance Atlantic Motion Picture Distribution.

Playtime. Directed by Jacques Tati. 1967. Les Films de Mon Oncle.

Run Lola Run. Directed by Tom Tykwer. 1998. Seville Pictures.

Muybridge, Eadweard. *Sallie Gardner at a Gallop*, motion picture. 1878. Corcoran Gallery of Art, Washington, DC.

Sesonske, Alexander. "Time and Tense in Cinema". *The Journal of Aesthetics and Art Criticism.* 1980. pp. 419-426.

Victoria. Directed by Sebastian Schipper. 2015. Mongrel Media.