6th BEIJING ANALYTIC PHILOSOPHY CONFERENCE

RENMIN UNIVERSITY

1 JUNE 2019

中國人民大學

RENMIN UNIVERSITY OF CHINA

SCHEDULE

1.	"Desire-Satisfactionism and the No-Future-Time-of-Desire View" **Frederick Choo** (Nanyang Technological University)	**Room 500** 9:30 – 10:15 am
	Comments: **Chang Liu** (Renmin University)	
2.	"The Problem of Disappearing Agency and Two Intuitions about Action" **Jingbo Hu** (University of Sheffield)	10:20 – 11:05 am
	Comments: **Ju Chen** (Renmin University)	
3.	"On the Validity of the Consequence Argument" **Liang Jin** (Shandong University, University of Missouri)	11:10 – 11:55 am
	Comments: **Mu Liu** (Renmin University)	
4.	"A Weak Knowledge Condition of Responsibility in Tracing" **Kerong Gao** (Georgia State University)	**Room 600** 10:20 – 11:05 am
	Comments: **Huicheng Ye** (China University of Political Science and Law)	
5.	"On Moral Responsibility for Implicit Biased Behaviors" **Xiaolong Wang** (Rutgers University)	11:10 – 11:55 am
	Comments: **Jie Tian** (Renmin University)	
	***** LUNCH BREAK ***** / ***** GROUP PICTURE *****	12:00 – 1:25 pm
6.	"Agent Regret and Accidental Agency" **Shaun Nichols** (University of Arizona)	**Room 500** 1:30 – 3:00 pm
	Comments: **Ding Lu** (Capital Normal University)	
	***** PAPER AWARDS *****	

Desire-Satisfactionism and the No-Future-Time-of-Desire View

Frederick Choo

Introduction

A prominent theory of well-being is desire-satisfactionism. Roughly, desire-satisfactionism states that a person's well-being increases if one's desires are satisfied, where a desire is satisfied if the desired object obtains.[1]

We have many present-directed desires (desires directed towards a present state of affairs), past-directed desires (desires directed towards a past state of affairs) and future-directed desires (desires directed towards a future state of affairs). Eden Lin notes that this raises two distinct questions that need answering.[2] The first is the benefit question. Let *t* be the time which a person has a desire for *p*, and let *t** be the time which *p* obtains. The benefit question asks, "How must *t* and *t** be related in order for you to *benefit* from the satisfaction of your desire?"[3] In cases involving present-directed desires, *t* and *t** are the same time. In cases involving past-directed desires, *t* is later than *t**. In cases involving future-directed desires, *t* is earlier than *t**. To address the benefit question, a theory must say whether a person can benefit in all of these cases or only in some. The second is the timing question which asks, "At what *time*, if any, do you benefit from the satisfaction of your desire?"[4] To address this question, a theory must say *when* a person benefits in each of these cases (if they do benefit).

One prominent theory is *Time-of-Desire* which states that a person benefits from a satisfied desire at the time the person has the desire. In answering the benefit question, proponents of *Time-of-Desire* usually hold that one can be benefited by past-directed, present-directed and future-directed desires. Call this view *Unrestricted-Time-of-Desire*. In this paper, I argue for a version of *Time-of-Desire* which answers the benefit question differently. On my view, a person can only be benefited by past-directed and present-directed desires. Whenever I am benefited, it is either because I have a present-directed desire towards a present object which obtains presently, or because I have a past-directed desire towards a past object and the object has obtained in the past. Future-directed desires cannot benefit a person, even if these desires are satisfied. Call this view *No-Future-Time-of-Desire*. My goal is to show that *No-Future-Time-of-Desire* is superior to the other existing views such as *Unrestricted-Time-of-Desire*, *Time-of-Object*, *Asymmetrism* and *Concurrentism*.

Current Views on the Benefit Question and Timing Question

Many views address the benefit question and the timing question differently. One view is Chris Heathwood's *Concurrentism*, which states that a person is only benefited if the desire and the desired object obtain at the same time. In other words, only present-directed desires can affect one's well-

[1] This differs from the feeling of having one's desire-satisfied or the feeling of satisfaction.
[2] Lin 2017: 164.
[3] Ibid.
[4] Ibid.

being.[5] In contrast to *Concurrentism*, all other views allow a person to be benefited even in cases involving past-directed desires and/or future-directed desires (i.e. where there is no temporal overlap between the desire and the desired object).

A second view is *Time-of-Desire*, which states that a person is benefited by a satisfied desire at the time one has the desire.[6] As mentioned, proponents of this view usually hold to *Unrestricted-Time-of-Desire*, allowing a person to be benefited by past-directed, present-directed and future-directed desires.[7] A third view is *Time-of-Object*, which states that a person is benefited at the time the desired object obtains.[8]

A fourth view is Eden Lin's *Asymmetrism*.[9] *Asymmetrism* holds *Time of Desire* is true of past-directed desires while *Time-of-Object* is true of future-directed desires. In other words, in cases involving past-directed desires, a person is benefited when one has the desire, but in cases involving future-directed desires, a person is benefited when the desired object obtains.

Intuitive Cases

In this section, I offer two intuitive cases in support of *No-Future-Time-of-Desire*. First, consider a case involving past-directed desires.

> *Drunk*. Suppose that I was drunk last night and did not make a fool of myself. While drunk however, I did not care about whether I would make a fool of myself or not. The next day, I wake up sober but I do not remember what happened the night before except that I was drunk. I now desire not to have made a fool of myself the night before.[10]

Intuitively, my well-being has increased. It seems better for me that I did not make a fool of myself last night. When does my well-being increase? Intuitively, I am benefited the next day when I was sober and desired not to have made a fool of myself the night before. It seems implausible to say that my well-being increases last night, at the time when I did not even have the relevant desire yet.[11]
Of all the views, only *No-Future-Time-of-Desire*, *Unrestricted-Time-of-Desire* and *Asymmetrism* accommodate this intuition. All three views hold that in cases involving past-directed desires, a person benefits at the time of desire. *Time-of-Object* gives the wrong verdict as it entails that I benefited last night when I did not care whether I would make a fool of myself. *Concurrentism* gives the wrong verdict since it entails I do not benefit at all.

---

[5] See Heathwood 2005. In regards to the timing question, a person is benefited at that time which one has the desire and the desired object obtains.
[6] Dorsey 2013, Bruckner 2013.
[7] To note, although Bruckner (2013) merely defends *Time of desire* in regards to future-directed desires, Bruckner has told me that he is also inclined to hold *Time of Desire* in regards to past-directed and present-directed desires as well.
[8] This does not seem to be defended by anyone though it is often discussed in the literature.
[9] Lin 2017.
[10] Purves 2017: 802-803 and see also Lin 2013: 161, 167.
[11] This intuition seems shared by a number of philosophers. See for example Sarch 2013: 232-233, Lin 2017: 167 and Purves 2017: 803.

One might object here that the intuition that I benefit is not due to a past object obtaining but due to a present object.[12] After all, it seems plausible that I desire not to have made a fool of myself because I desire to not face the consequences. So the desire not to have made a fool of myself last night is merely an instrumental desire. What I intrinsically desire is to not face the consequences. It is this present-directed desire which is being satisfied which increases my well-being.

To address this, we can add further details into *Drunk*. Suppose I was hanging out with a new group of people I never met, and after that night, I would fly away and never see them again. I know that the group would not try to harm me by posting videos of me, or making fun of me, and so forth. So regardless of whether I made a fool of myself or not, there would be no bad consequences. Despite knowing this, I still desire not to have made a fool of myself last night. I have such a desire intrinsically. With these details, it still seems better for me if I did not make a fool of myself last night.
Next, consider a case involving future-directed desires.

> *Fireman*. Suppose when I was young, I desired to be a fireman.[13] However, as I grew older, I completely lose the desire to be a fireman. Suppose that when I am 40, I end up being a fireman for the next 10 years although I have already long lost the desire to be a fireman.

Intuitively, my well-being has not increased. It seems implausible that I benefited when I was young because it is years before the desired object obtains and hence my desire is left unsatisfied then. It also seems implausible to say that I benefited when I was 40 since I did not have the relevant desire then. How can it be good for me to get what I no longer want?[14]

Of all the views, only *No-Future-Time-of-Desire* and *Concurrentism* accommodates this intuition. Both these theories hold that in cases involving future-directed desires, a person does not benefit. *Unrestricted-Time-of-Desire* gives the wrong verdict since it entails that I benefited years before the desired object obtains. *Time-of-Object* and *Asymmetrism* gives the wrong verdict since it says that I benefit at 40 years old, when I get what I no longer want.

These two cases support only *No-Future-Time-of-Desire*. However, *No-Future-Time-of-Desire* enjoys more than intuitive support, the next two sections covers two principles that support *No-Future-Time-of-Desire*.

The Synchronic Resonance Constraint

A major motivation for desire-satisfactionism is that it accommodates the resonance requirement, which states that in order for something to be good for you, it must resonate with you.[15] After all, how can something be good *for me* if I do not have some sort of favorable attitude towards it at all? As Peter Railton says, "What is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive ... It would be an intolerably alienated conception of someone's

[12] I thank an anonymous reviewer for raising this concern.
[13] To note, the desire here is not to be read as a conditional desire. It is not that I desire to become a fireman given that this is what I will desire in the future. Rather, I desire to become a fireman regardless of whatever I desire in the future.
[14] This intuition seems shared by a number of philosophers. See for example Sumner 1999: 126-127, Bykvist 2003, Heathwood 2005: 490, Bradley 2009: 20-22 and Bruckner 2013.
[15] This requirement is also often referred to as "non-alientation" and "internalism." See for example Rosati 1996: 297-326, Railton 2003: 47-48 and Dorsey 2017: 685-708.

good to imagine that it might fail in any such way to engage him."[16] Desire-satisfactionism accommodates the resonance requirement by holding that what is good for a person is determined by a person's desires.

Dale Dorsey proposes that desire-satisfactionists further accept what Lin calls the:

> "*Synchronic Resonance Constraint*: You do not benefit from a particular event, *e*, at time *t* unless, at *t*, you have a favorable attitude toward *e*."[17]

Given that desire-satisfactionists accept the resonance requirement, it would be strange to say that a person is benefited at a time at which the desired object does not resonate with the person, rather than the time at which the desired object resonates with the person. This gives us reason to accept the *Synchronic Resonance Constraint*. As Dorsey says, if one does not accept the *Synchronic Resonance Constraint*, then this "would leave it open that I can lack a desire for ϕ at a particular time, but that ϕ is nevertheless good for me at that time."[18] This would allow "welfare goods that alienate the person whose benefits they are."[19]

Lin objects to Dorsey's *Synchronic Resonance Constraint*, pointing out that "there is another similar principle that Time of Desire fails to accommodate:

> *Synchronic Resonance Constraint**: You do not benefit at any time from a particular event, *e*, unless you have a favorable attitude toward *e* at the time at which *e* occurs."[20]

Just as proponents of *Time of Desire* would plausibly think that desire-satisfactionists need not accept the *Synchronic Resonance Constraint**, Lin thinks that desire-satisfactionists need not accept the *Synchronic Resonance Constraint* as well.

Lin's objection however can be addressed. First, Lin does not show what is wrong with the *Synchronic Resonance Constraint* or the *Synchronic Resonance Constraint**. He merely points out that proponents of *Time of Desire* would likely reject the latter while he rejects both. This in no way shows what is wrong with any of the principles. Second, what motivates the *Synchronic Resonance Constraint* does not also motivate the *Synchronic Resonance Constraint**. Given the resonance requirement, it would be more plausible to suppose that one benefits at the time of resonance rather than at the time the person is alienated. After all, it seems odd to allow objects to be good for me at times I do not desire it. Consider *Fireman* again. How can being a fireman be good for me at the time I have no desire for it? I am alienated from it. Notice that the focus of the resonance requirement is on *resonance*. In contrast, there isn't a similar argument from the resonance requirement to the *Synchronic Resonance Constraint**. To argue from the resonance requirement to the *Synchronic Resonance Constraint**, we would have to say that the focus of the resonance requirement is on both *resonance* and *the desired object*. But the resonance requirement by itself says nothing about the desired object. Therefore, we have an argument from the resonance requirement to the *Synchronic Resonance Constraint* but not to the *Synchronic Resonance Constraint**.

---

[16] Railton 2003: 47.
[17] Lin 2017: 179. See also Dorsey 2013: 156-157.
[18] Dorsey 2013: 156.
[19] Ibid.
[20] Lin 2017: 179.

No-Future-Time-of-Desire, Unrestricted-Time-of-Desire and *Concurrentism* can clearly accommodate the *Synchronic Resonance Constraint* since these views always locate the benefit at the time the person has the relevant desire. *Time-of-Object and Asymmetrism* would violate this constraint since in the case of future-directed desires it allows a person to be benefited at a time the person lacks the relevant desire.

The All-Conditions-Met Principle

Lin has recently argued for what I call the

> *All-Conditions-Met Principle*: "You do not receive a particular benefit at *t* unless, at *t*, all of the necessary conditions on your receiving that benefit have been met."[21]

Lin says that a "condition has been met at *t* just if it either *is* met at *t* or *was* met at some time prior to *t*."[22] The *All-Conditions-Met Principle* prevents a person from benefiting until the earliest time that all the necessary conditions for receiving the benefit have been met. This means I cannot benefit at a time which I have the desire but the desired object has not yet obtained; and I cannot benefit at a time which the desired object has obtained but I have not desired it yet. Lin has provided various arguments for the *All-Conditions-Met Principle* and it seems right to me that any theory has to meet this principle.[23]
Both *Time-of-Object* and *Unrestricted-Time-of-Desire* cannot accommodate the *All-Conditions-Met Principle*. In the case of past-directed desires, *Time-of-Object* entails that you benefit at a time which the desired object obtains but you have not had the relevant desire yet. In the case of future-directed desires, *Unrestricted-Time-of-Desire* entails that you benefit at a time which you have the desire, but the desired object has not yet obtained.

A proponent of *Unrestricted-Time-of-Desire* may try to accommodate the *All-Conditions-Met Principle* by modifying what counts as the necessary condition.[24] The necessary condition to be benefited is that one's desire is satisfied. If our desires can be satisfied at a time at which the object has not yet obtained, then we can benefit at a time at which the object has not yet obtained. One might wonder how our desires could be satisfied at a time at which the desired object has not yet obtained. A proponent of *Unrestricted-Time-of-Desire* may say that truths about the future can make it the case that my desire is satisfied now.[25] For example, suppose that now in 2019, I desire to be married in 2020; and I will get married in 2020. We can say that it is true now in 2019 that I will get married in 2020. Since it is true *now* in 2019 that I will get married in 2020, and I desire *now* in 2019 to get married in 2020, my desire is satisfied *now* in 2019. One might find this strange, but there are many cases where future states of affairs determine the status of past states of affairs. Donald Bruckner, for example, points out, "A shooting acquires the status of a killing only if the victim dies as a result of the gunshot, which may be some time later."[26] Another example is the truth value of propositions that reference the future. If the weather forecaster says it will rain tomorrow, whether what he says is true or false now depends on what happens the next day.[27] Similarly, we can say that my desire is satisfied now as long as it is true that the desired object obtains at a future time. Therefore, we can say that even though the desired object only obtains at a future time, my desire is satisfied in the present and so I benefit in the present.

---

[21] Lin 2017: 169. Lin calls this the First Principle.
[22] Lin 2017: 169.
[23] Lin 2017: 169-171. See also Purves (2017: 809) who agrees that this principle is attractive.
[24] I thank an anonymous reviewer for a discussion on this strategy.
[25] See Bradley 2009: 23-24, Bruckner 2013: 18.
[26] Bruckner 2013: 25-26.
[27] Bruckner 2013: 18.

This strategy however will not do. First, the two necessary conditions for having a desire-satisfied at a certain time seems to be (a) that the person has had the desire (either at that time or prior to that time) and (b) that the desired object has obtained (either at that time or prior to that time). The conditions for having a desire-satisfied at a certain time is not about whether *it is true* that the event would obtain at a future time (or about whether *it is true* that one would have the desire at a future time). So even if it is true now in 2019 that the desired event would obtain, the point is that one has the desire at 2019 but the desired event has not yet obtained at 2019. Therefore, we should not say that my desire is satisfied in 2019, at a time which the event has yet to obtain. I have not gotten what I wanted in 2019. This is why we should say that in 2019, my desire has not yet been satisfied though it will be satisfied in 2020. The 'it is true that' talk is simply irrelevant to having a desired satisfied at a certain time.[28]

Second, we should reject the 'it is true that' talk because it would result in implausible consequences. If we allow the 'it is true that' talk in the case of future objects obtaining, then it seems reasonable to extend the 'it is true that' talk to our future desires as well. If so, then we can count a desire as satisfied at the time which it is true that a person would have the desire and which it is true that the event would obtain. For example, if in 2018 it is true that I will form the desire in 2019 to be married in 2020, and if in 2018 it is true that I will be married in 2020, then my desire is satisfied in 2018. This is implausible. How can my desire be satisfied in 2018 prior to me having the desire and prior to the event obtaining? This is not the only implausible consequence. Consider a case of past-directed desires. Suppose my parents forced me to study philosophy while I was young and I hated it. However, in 2019, I saw how studying philosophy was a good thing and I now want to pursue for its own sake. I form the desire to have had studied philosophy when I was young. If we allow the 'it is true that' talk about our future desires, we can say that my desire was satisfied when I was young. After all, when I was young I studied philosophy, and when I was young it is true that in 2019 I would desire that state of affairs. This seems implausible too. How can my desire be satisfied when I was young when I did not even have the desire then? In this scenario, intuitively, my desire is satisfied in 2019. Therefore, we should reject the 'it is true that' talk. When we ask whether a desire is satisfied at a certain time, it is irrelevant whether *it is true* that the event would obtain at a future time or a whether *it is true* that one would have the desire at a future time.

In contrast to *Time-of-Object* and *Unrestricted-Time-of-Desire*, all the other views seem to be able to accommodate the *All-Conditions-Met Principle*. Consider *No-Future-Time-of-Desire*. In the case of past-directed desires, a person benefits at the time of desire. By this time, the desired object has obtained and the person has the desire. So, both necessary conditions have been met. In the case of present-directed desires, a person benefits at the time that the desired object obtains and the person has the desire. So both necessary conditions have been met.

Concurrentism

Only two views can accommodate both the *Synchronic Resonance Constraint* and the *All-Conditions-Met Principle: No-Future-Time-of-Desire* and *Concurrentism*. In this section, I offer two reasons to prefer *No-Future-Time-of-Desire* over *Concurrentism*.

First, there are many cases involving past-directed desires which intuitively seem to affect one's well-being. Recall the case *Drunk* for example. If I desired not to have made a fool of myself the night before,

[28] See Bradley 2009: 24 for another argument on this point based on truthmakers.

it seems better for me if I did not make a fool of myself the night before. *Concurrentism* cannot easily account for how such past-directed desires intuitively seem to affect my well-being.

Second, various philosophers have argued that since desire-satisfactionists accept spatial distance in their theory, they should allow temporal distance as well.[29] Desire-satisfactionists often argue that things can be good or bad for us without making a difference in our experiences. For example, suppose you desire that your privacy is not constantly violated by people spying on you and making fun of you behind your back. Unknown to you however, your privacy is constantly violated by a group of people who spy on your life and laugh at you. This seems bad for you even though you do not know this. Such cases show that things can be good or bad for a person even though it does not affect their experiences. However, if we allow this spatial distance, why restrict temporal distance? As Duncan Purves says, "if [we] accept that spatial distance between the experience of the desirer and the object of her desire is irrelevant to whether her desire counts as satisfied, then why should temporal distance matter?"[30] *No-Future-Time-of-Desire* can easily accommodate this. *Concurrentism* however cannot does not allow temporal distance. For these two reasons, I think *No-Future-Time-of-Desire* should be preferred over *Concurrentism*.

From the Timing Question to the Benefit Question

One might object as follows. My *No-Future-Time-of-Desire* says something about the benefit question, namely that future-directed desires are irrelevant to well-being. However, both the *Synchronic Resonance Constraint* and the *All-Conditions-Met Principle* only address the timing question but not the benefit question. The *Synchronic Resonance Constraint* gives us a reason to restrict the time of benefit to the time one has the relevant desire. The *All-Conditions-Met Principle* gives us a reason to restrict the time of benefit to the earliest time that all the necessary conditions on receiving the benefit have been met. So it seems that the principles that I invoked cannot be used to justify my view on the benefit question.[31]

To illustrate, recall *Fireman*. The *All-Conditions-Met Principle* prevents me from being benefited when I was young because the desired object has not yet obtained then. The *Synchronic Resonance Constraint* prevents me from benefiting when I am 40 years old since the desired object does not resonate with me then. My two principles show that I could not have benefited when I was young or when I was 40 years old. My two principles however do not show that I do not benefit at all. After all, it is possible that I benefited (a) at a fusion time, (b) across my whole life, (c) at no time.

In reply, we can show that each of these options is problematic. Option (a) can be ruled out by the *Synchronic Resonance Constraint*. I do not have the desire across the fusion of time. Option (b) can be ruled out by the *Synchronic Resonance Constraint*. Since I do not have the desire across my whole life, I cannot be benefited across my whole life.

This leaves option (c), the idea that I benefited but at no time. This option however seems problematic. Suppose that if I did not become a fireman, my well-being has an average value of 50 across my life. If I became a fireman, would the average value be higher? Option (c) seems to entail that the average value would not be higher. This is because my well-being can have a higher average value across my life only if

[29] See Bradley 2009: 23, Dorsey 2013: 157-158 and Purves 2017: 803.
[30] Purves 2017: 803.
[31] I thank an anonymous reviewer for a discussion on this point.

I benefited at some point of time across my life. So option (c) seems to entail that I would have an average value of 50 across my life whether I became a fireman or not. But since my well-being would have an average value of 50 across my life whether I became a fireman or not, it is hard to see how I benefited from becoming a fireman. If I do not benefit at any time, then it seems that I do not benefit at all. Given this, we should conclude that in cases involving future-directed desires, since I do not benefit at any time, I do not benefit at all.

Conclusion

No-Future-Time-of-Desire is intuitive, and can accommodate both the *Synchronic Resonance Constraint* and the *All-Conditions-Met Principle* (which I have argued desire-satisfactionists should accept). Although *Concurrentism* can also accommodate both principles, *Concurrentism* has unintuitive results and faces the problem of allowing spatial distance while not allowing temporal distance. Therefore, *No-Future-Time-of-Desire* is superior to all existing views.

References

Bradley, Ben. *Well-being and Death*. Oxford: Clarendon Press, 2009.

Bruckner, Donald W. "Present Desire Satisfaction and Past Well-Being." *Australasian Journal of Philosophy* 91, no. 1 (2013): 15-29.

Bykvist, Krister. "The Moral Relevance of Past Preferences." In *Time and Ethics: Essays at the Intersection*, edited by Heather Dyke, 115-36. Boston: Kluwer Academic Publishers, 2003.

Dorsey, Dale. "Desire-satisfaction and Welfare as Temporal." *Ethical Theory and Moral Practice* 16, no. 1 (2013): 151-71.

Dorsey, Dale. "Why Should Welfare 'Fit'?" *The Philosophical Quarterly* 67, no. 269 (2017): 685-708.

Heathwood, Chris. "The Problem of Defective Desires." *Australasian Journal of Philosophy* 83, no. 4 (2005): 487-504.

Johansson, Jens. "The Time of Death's Badness." *Journal of Medicine and Philosophy* 37 (2012): 464-79.

Lin, Eden. "Asymmetrism about Desire Satisfactionism and Time." In *Oxford Studies in Normative Ethics*, edited by Mark Timmons, 161-83. Vol. 7. Oxford: UK: Oxford University Press, 2017.

Purves, Duncan. "Desire Satisfaction, Death, and Time." *Canadian Journal of Philosophy* 47, no. 6 (2017): 799-819.

Railton, Peter. "Facts and Values." In *Facts and Values and Norms: Essays Toward a Morality of Consequence*, 43-68. Cambridge: Cambridge University Press, 2003.

Rosati, Connie S. "Internalism and the Good for a Person." *Ethics* 106, no. 2 (1996): 297-326.

Sarch, Alexander. "Desire Satisfactionism and Time." *Utilitas* 25, no. 02 (2013): 221-45.

Silverstein, Harry S. "The Time of the Evil of Death." In *Time and Identity*, edited by Joseph Keim Campbell, Michael O'Rourke, and Harry S. Silverstein, 283-95. Cambridge: MA: MIT Press.

Sumner, L. W. *Welfare, Happiness, and Ethics*. Oxford: Clarendon Press, 1996.

The Problem of Disappearing Agency and Two Intuitions about Action: Defending the Causal Theory of Action

Jingbo Hu

Introduction: Two Competing Causal Frameworks for Action

Many philosophers of action agree that humans' actions are causal phenomena—when there is action, there is causation. However, they disagree on which causal framework should be used to analyze the notion of action. There are two competing frameworks—the event-causal framework and the agent-causal framework.

> Event-Causation: Action are bodily movements caused by certain motivating mental states (such as desires, beliefs, intentions, etc.).
>
> Agent-Causation: Action are bodily movements caused by the agent (which is a fundamental substance and cannot be *reduced* to the mental states).

To illustrate, if I open the bottle and drink the water, different stories will be told by the two accounts. The story from event-causation account is like this: I am feeling thirsty and have a desire to get water, and I believe that by opening the bottle I can get water, these mental states together generate an intention to open the bottle, and this intention subsequently cause my bodily movements. For the agent-causation account, the story is different: It is not the mental states such as beliefs, desires or intentions that cause my body to move; rather, it is me, or the self, a metaphysically fundamental substance, that causes the resultant bodily movement of opening the bottle.

Many philosophers endorse the event-causal framework for understanding action for it is more intelligible. Since they hold that causation in general is event-causation, this theory can place action into the orders of nature. Currently, the causal theory of action (CTA) which make use of the event-causal framework become the dominant action theory in the literature such that it is called 'the standard story'. However, opponents of CTA argue that the event-causal framework would make the agent disappear, which is now known as the problem of disappearing agency. Although different opponents have raised this problem in different ways, the most classic characterization is from David Velleman's paper 'what happens when someone acts'. Here is a quote:

> I think that the standard story is flawed in several respects. The flaw … is that the story fails to include an agent, or more precisely, fails to cast the agent in his proper role. In this story, reasons cause an intention, and an intention causes bodily movements, but … no person does anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part. (Velleman 1992)

However, defenders of CTA seem not to take this problem seriously. One possible explanation is that the presentations of problem seem to be vague and that it is not backed up by compelling arguments. For example, Markus Schlosser, a defender of CTA, writes that:

> [Proponents] (of the problem of disappearing agency) have not produced a single argument to support their case, and they have certainly not identified a philosophical problem. Their case is entirely based on intuition, and in some cases on mere metaphor and rhetoric. (Schlosser 2010)

Although I endorse CTA, I am sympathetic to the problem of disappearing agency. I believe that this problem reveals deep concerns about agency which requires attention from both sides. According to my own interpretation, the nature of the problem is that event-causal framework employed by CTA fall short of capturing certain key intuitions about agency:

> *Agent-Producing*: Action by nature is the agent bringing something about such that agent is the source and the initiator of her action.
>
> *Anti-Reduction*: There is no agency among mere happenings.

In this essay, I will discuss these two intuitions respectively. I will show that both intuitions as well as the arguments based on them are not enough to rebut the event-causal framework.

Agent-Producing

Agent-Producing is the intuition that an action must be produced or brought about by the agent herself. This intuition is *prima facie* compelling. Suppose Jack is flinching because of intense pain. Although his flinching is his bodily movement, we would not regard it as an action done by Jack. And probably the reason is that the flinching is not genuinely produced by Jack *himself*. Rather, the movement is brought about by the *pain.* Or consider the scenario of Mike who can act just like us. However, one day it is discovered that all of Mike's behaviour is produced and controlled by a remote evil neuro-scientist through a device implanted in his brain. Then, we no longer regard the bodily motions of Mike as genuine actions. Even worse, we no longer regard Mike as a genuine agent who can perform action by himself. This is because we now know that Mike is not the source of his behaviours. Below is a quote from the literature which represents the idea of Agent-Producing:

> In describing anything as an act there must be an essential reference to an agent as the performer or author of that act, not merely in order to know whose act it is, but in order even to know that it is an act at all. ... Another perfectly natural way of expressing this notion ... is to say that, in acting, I make something happen, I cause it, or bring it about. (Taylor 1966, pp. 109-111)

Admittedly, the intuition of Agent-Producing is generally shared—everyone who understands the word 'action' tends to agree with it. This widely shared intuition provides salient support to the agent-causation account according to which the agent figures as an unreducible element in the explanation of action. The event-causal framework, by contrast, seems to leave no place for the agent to bring about her own action. According to CTA, what take place during the action are just certain mental events causing the bodily movements. The agent in this story is more like an 'arena' for interactions of the mental and bodily events rather than a real initiator or producer of actions. Thus, CTA or the event-causal framework is criticized as failing to capture Agent-Producing.

Is the event-causal framework really incompatible with the intuition of agent-producing? Before addressing this question, we shall distinguish two level of considerations about action—the common-sense considerations and the ontological considerations. Roughly speaking, the common-sense considerations are what ordinary people think and mean when they are using the concept of 'action'. The relevant data and evidence for this kind of research comes from people's ordinary conception and semantic intuitions of ordinary people. In contrast, the ontological investigation of action concerns the metaphysical structure of action.[1] Thus, the debate on the causal framework regarding action then

[1] Though different philosophers of action may view the ontological project differently, a typical understanding of the ontological project of aim to providing ontological and sufficient conditions for

should be a debate about the ontological considerations. If this is the case, it is crucial to tell whether the intuition of agent-producing is relevant to the ontological investigation. And my answer is negative.

The feature of being widely shared suggests that the Agent-Producing is an intuition on common-sense level. In fact, the proponents of event-causation account of action can claim that their theories are also abiding by Agent-Producing as long as they argue that the event-causation account can accommodate the intuition of Agent-Producing. For example, Robert Kane, who is a proponent of event-causation account of action, writes that: "[d]oing without agent-causation in the non-occurrent sense does not mean denying agent-causation in the ordinary sense that agents act, bring things about, produce things, make choices, form their own characters and motives, and so on." Kane (1996, p. 123) What Kane proposes here is that one can be an agent-causalist on the common-sense level while being an event-causalist on the ontology level. To show that CTA can accommodate agent-producing, what proponents need to do is to argue that agent-causation can be realized by and reduced into event-causation. Clarke (2017) proposes a schema to reduce agent-causation into event-causation as follows:
Substance s caused event e2 just in case there was some event, e1, such that e1 involved s and e1 caused e2. (Clarke 2017, p.2)[2]

With this schema, a proponent of CTA can then analyse the common sense agent-causation in terms of evet-causation. For example, an agent (s) raises her hand up (e2) if and only if the relevant mental states of the agent (e1) cause her hand rising up (e2) in the appropriate way.

Opponents of CTA may probably not find such a reduction compelling. There is much room to debate about whether this kind of ontological reduction is tenable or not. The point is, the intuition of Agent-Producing seems to be an intuition remaining on the common-sense level and that it is not fine-grained enough to determine nuanced metaphysical issues. Merely from the intuition of Agent-Producing, it is difficult to reach the conclusion that an action is caused by the agent as an unreducible substance. Two more points can be used to substantiate my diagnosis.

Firstly, the claim of Agent-Producing seems not to provide many ontological indications for notions of 'agent' or 'producing'. Our ordinary conception of agent, or other related concepts, such as person, or the self, is ontologically vague. Specifically, only by referring to common sense, it is hard to answer the metaphysical questions such as whether the agent is a bunch of mental states or a non-reducible substance. This leaves much room for proponents of CTA to reconstruct the idea of agent-producing in terms of event-causation. Likewise, the ordinary conception of producing does not tell us much about metaphysics of causation either. From a naïve understanding, 'to produce' just means to changes through causal interactions. If the flash of thunder *caused* the forest to fire, naturally, we can say that the flash of the thunder *produced* the fire in the forest. This suggests that our ordinary conception of producing does not presuppose agent-causation. Of course, in the case of actions, the notion of 'producing' also involves intentionality or goals—if an agent raises her hand, then the agent raises her hand *intentionally* and that the agent raises her hand to achieve a certain goal. This again, is not a problem for CTA because the mental states cited as causes already involve intentionality.

---

action such that the conditions obtain not only in our actual word but in all other possible worlds that are similar to our actual world (e.g., Bishop 1990).

[2] Lowe (2008, Chapter 6) provides a similar schema for event-causalists to analyze substance-causation: Agent A caused event e if and only if there was some event, x, such that x involved A and x caused e.

Secondly, agent-causation is usually regarded as a special case of substance-causation (e.g., the flying ball cause the window to break). In our ordinary discourse of causation, we seem to use substance-causation talk and event-causation talk interchangeably to report the *same* causal phenomena. For example, we can say that a *ball* causes the break of the window, which is a typical claim of substance-causation. But we find it equally natural to say that the *movement* of the ball cause the break of the window, which is a claim of event-causation. In our ordinary discourse, we do not make a sharp distinction between substance-causation and event-causation. This may further suggest that our common-sense conception of causal phenomena is not fine grained enough to determine the metaphysical disagreement about causal frameworks. This lesson applies to the case of agency as well. Although we do employ agent-causation terms to describe agency and action, it does not follow that the use of this terms involves substantial ontological implications.

So far, I have argued that the intuition of Agent-Producing does not exclude the possibility of analyzing action in terms of event-causation. This is because, this intuition is not likely to be an intuition on ontological level and in effect it is inadequate to determine ontological issues about choosing the proper causal framework for action.

Anti-Reduction

To rebut the criticism from Agent-Producing, one key move is to argue that the intuition of agent-producing is not about ontological matters. Specifically, proponents of CTA can argue that Agent-Producing can be realized in an event-causal process. By doing this, they can claim to respect the Agent-Producing intuition without being an agent-causalist in the ontological sense. However, opponents may reply that, no matter how event-causations are constituted, there cannot be Agent-Producing. The assumption seems to be that events are something *merely happens* in the world so events cannot constitute actions. This is the main motivation behind the intuition of Anti-Reduction, according to which, there is no agency among mere happenings. The intuition of Anti-Reduction goes beyond our common-sense understanding of action and it operates on the level of ontological investigation. If Anti-Reduction is true, it will block any attempts to find agency within event-causations. In a word, Anti-Reduction is a straightforward denial of the reductive program inherent to CTA. Here is a quote from Melden which typically represents the concern of Anti-Reduction:

> It is futile to attempt to explain conduct through the causal efficacy of desire—all that can explain is further happenings, not actions performed by agents.... There is no place in this picture... even for the conduct that was to have been explained. (Melden 1961, pp. 128–29, quoted from Mele (2003))

Carlos Moya also provides a vivid scenario in order to pump our intuition of Anti-Reduction:

> … let us start with an episode that nobody would hesitate in classifying as an action, say, drinking a glass of water ...The water got into my mouth as an effect of gravity. The water getting into my mouth is a mere happening. This happening, in turn, was caused by the movement of the glass. Where is action in this? ... this movement can be said to be properly caused by my arm's and hand's movement, which in turn were caused by some muscles' contractions, which in turn were caused by some neurons' firings, and so on … Our everyday sharp distinction between actions and happenings begins to fade; it seems that we were calling 'action' what is in fact a series of causally related happenings. (Moya 1990)

The quoted passages express a similar worry that mere happenings cannot constitute agency: events by nature are something that just happens; from the story of CTA, however, process of action is only

constituted by events happening *passively*—we cannot find a single component which is genuinely *active*. We can reconstruct the argument from Anti-Reduction as follows:

P1: The event-causal framework employed by CTA reduces action into mere happenings.

P2: There is no agency among mere happenings. (Anti-Reduction)

C: Event-causal framework is inadequate to capture agency.

The force of this argument comes from the ambiguity of the term 'mere happenings'. After all, it sounds like an oxymoron to say that agency can come from mere happenings. The gloss of 'mere happenings' then becomes critical to the assessment of this argument. There are two possible readings for mere happenings—a narrow one and a broad one. From the narrow reading, the notion of 'mere happenings' means unintentional bodily movements such as eyes blinking or body trembling out of cold. This is the reading adopted by defenders of CTA. From this reading, the defenders of CTA would agree with P2, for it is both obvious and by definition true that unintentional bodily movements are not actions. However, they would disagree with P1 because they hold that CTA have the resources to distinguish action from mere happenings in the narrow reading.[3] Specifically, they will argue that what distinguishes action from mere happenings is that action must be caused by the appropriate motivating mental states in a non-deviant way.

Those who endorse this argument probably adopt a broad reading of mere happenings such that 'mere happenings' in the above argument is synonymous with 'events' or 'causal interactions among events'. However, this reading makes the intuition of Anti-Reduction (or P2 of the argument) significantly less compelling and even question-begging. Note that it is the aim of CTA to reduce the action into event-causation so that action can be placed in the orders of nature. Defenders of CTA may simply dismiss the Anti-Reduction as representing no truth but a nostalgia for the pre-scientific word view. Admittedly, the defenders of CTA cannot easily convince their opponents to give up Anti-Reduction.[4] The point is, however, that merely insisting in Anti-Reduction without providing further arguments would not make any achievements in the debate other than being trapped in a dialectic stalemate.

If the opponents of CTA want to get out of the stalemate, what they need to do is to buttress the intuition of Anti-Reduction by providing independent justification or explanation. Otherwise, this argument may not have much appeal to those who deny the intuition of Anti-Reduction at the very beginning. For example, one can argue that event-causations cannot constitute agency because of the existence of causal deviance (e.g., Sehon 1997). Since this move would transform the problem of disappearing agency (under the current interpretation) into the problem of causal deviance and go beyond the scope of this essay, I am not going to discuss this option here.

Equally, if the defenders of CTA manage to explain away the intuition of Anti-Reduction other than simply rejecting it, they can gain dialectic advantages. This is what I am trying to do now. My argument heavily relies on the work from Neal Judisch (2010). In his paper, Judisch tries to explain why the concept of agency persistently resists reduction. He proposes that the notion of agency is 'conceptually irreducible' because our conception of agency essentially involves a special phenomenology of action.

[3] For example, see Schlosser (2010)

[4] For example, Hornsby (2004) writes:' it is only to those in the grip of the naturalistic conception of what happens when someone acts that it could seem that differences between actions had always to be recorded as differences between causally efficacious items that produced them.' The opponents of CTA may argue that only those biased by the obsession of naturalism would reject to accept Anti-Reduction.

Specifically, whenever we act, we possess a special feeling of doing something, which ensures us that the action is something done by us rather than something happens to us. It is this subjective agentive experience that makes it awkward to analyze action within the event-causal framework. Judisch correlates this problem with the 'hard problem' in philosophy of mind, according to which there is always an 'explanatory gap' between our subjective phenomenal consciousness and the physical matter. The phenomenal consciousness of the mental is notoriously difficult to be analyzed with physical terms because of the subjective characteristic. For Judisch, this difficulty also applies to the reductivism program of action because the agentive experience he invokes also has a subjective aspect. He claims that any reductive programs have to put the target under an objective perspective, which makes the subjective aspect disappear. Judisch quotes Nagel to strengthen his position:

> Something peculiar happens when we view action from an objective or external standpoint. Some of its most important features seem to vanish under the objective gaze. Actions seem no longer assignable to individual agents as sources, but become instead components of the flux of events in the world of which the agent is a part. (Nagel 1986, p. 110)

Judisch further suggests that 'it is because the phenomenology of agency is absent from third-person conceptualizations that action theories developed from this perspective will inevitably appear to ignore something significant' and that 'what is apparently missing from the naturalistic account is a particular *phenomenal* conception of ourselves as sources of our conduct' (p. 103). For Judisch, this ignored aspect of agency can only be accommodated from 'the internal point of view' (p. 104).

Judisch's explanation, if correct, will probably make Anti-Reduction irrelevant to rebut CTA. Firstly, according to this explanation, the source of Anti-Reduction is the experience of action. It is the subjective experience of action that resists the attempt of reduction. However, CTA does not aim to reduce or explain the subjective experience of action. Although CTA usually falls under the name of reductionism, there are actually two different reductive programs going on: the first is to reduce the mental (especially the phenomenal consciousness) properties into the physical properties; and the second to reduce action into event-causation. Defenders of CTA, especially those with a naturalist orientation, would admit that a more thorough understanding of human agency should include the reduction of the metal properties (the experience of agency in particular) into the physical properties. However, this task should be better left to philosophers of mind. Admittedly, for a certain version of CTA, even it is successful, it would just be one step of the entire project of naturalizing agency. Apart from cracking the hard problem of phenomenal conscious, the naturalization of agency requires several different philosophical achievements, say, the naturalization of non-derived intentionality, and the naturalization of rationality. To fully understand agency, we definitely need input from other philosophical areas and that one should not demand and expect too much from CTA.

Secondly and relatedly, according to Judisch's explanation, the problem is rooted in the distinction of external perspective/internal perspective—the agentive experience can only be done justice to from an internal, subjective, first personal perspective. However, even though distinction of internal/external perspectives cuts across the distinction of event-causation/agent-causation, these two distinctions can be dissociated. True, within the event-causal framework, we are more inclined to view actions from an external perspective. But this is not necessary. Even within the agent-causal framework, the action can be viewed either under an internal perspective or an external perspective. An action can be cashed out in terms of agent-causation from external/third-personal perspective: the agent causes her hand to rise by exercising her agential power. Or it can be cashed out in terms of event-causation from the

internal/first-personal perspective: my intentions caused my hand to raise.[5] Thus, according to this explanation, even if the subjectivity of agentive experience resists any reductive analysis, we have no reason to blame the event-causal framework endorsed by CTA.

Concluding Remarks

In this essay, I have discussed a particular interpretation for the problem of disappearing agency, according to which the event-causal framework is intuitively incompatible with human agency. I pinpoint two intuitions about agency related to the problem—the Agent-Producing and the Anti-Reduction. I have argued that the intuition of Agent-Producing is operating on common-sense level and that it is not adequate to decide appropriate causal frameworks for action. I have also argued that the Anti-Reduction will probably not shared by defenders of CTA and that this intuition should be supported by further arguments or explanations. Merely invoking the intuition of Anti-Reduction can hardly serve the dialectic aim that it is supposed to serve. In addition, I provided an explanation which, if correct, would explain away the intuition of Anti-Reduction. Therefore, I conclude that these two intuitions about actions cannot rebut CTA and the event-causal framework.

References

Bishop, J. (1990). *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge University Press.
Clarke, R. (2017). Free Will, Agent Causation, and "Disappearing Agents." *Noûs*, 76–96.
Hornsby, J. (2004). Agency and Actions. *Royal Institute of Philosophy Supplement*, *55*, 1–23.
Kane, R. H. (1996). *The Significance of Free Will* (Vol. 110). Oxford University Press.
Lowe, E. J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford University Press.
Melden, A. I. (1961). *Free Action*. Routledge.
Mele, A. R. (2003). *Motivation and Agency*. Oxford University Press.
Moya, C. (1990). *The Philosophy of Action: An Introduction*. Polity Press.
Nagel, T. (1986). *The View From Nowhere*. Oxford University Press.
Schlosser, M. E. (2010). Agency, Ownership, and the Standard Theory. In A. Buckareff, J. Aguilar, & K. Frankish (Eds.), *New Waves in Philosophy of Action* (pp. 13–31). Palgrave-Macmillan.
Sehon, S. R. (1997). Deviant Causal Chains and the Irreducibility of Teleological Explanation. *Pacific Philosophical Quarterly*, *78*(2), 195–213.
Taylor, R. (1966). *Action and Purpose*. New York: Humanities Press.
Velleman, J. D. (1992). What Happens When Someone Acts? *Mind*, *101*(403), 461–481.

[5] Perhaps some would think that the problem is not that we cannot provide an event-causation description from the first personal perspective. Rather, the problem is that whenever we try to provide such a description, it would fail to accurately capture our first personal experience of action. I think this is a fair objection and I am going to deal with it elsewhere for the limitation of space here.

A Weak Knowledge Condition of Responsibility in Tracing

Kerong Gao

Introduction

It is widely accepted that in order to be morally responsible, an agent must meet a certain kind of knowledge condition—namely, that for an agent to be responsible for an outcome of her action, she must be able to foresee the outcome *when* she acts. However, the knowledge condition is often not met by an agent at the time when she acts. Vargas gives a case where Luis runs over some pedestrians on his way home from a bar in which he got drunk (Vargas, 2005, 269). When he is driving, he cannot foresee the outcome due to his drunkenness, which implies that he is not responsible. But we intuitively think that he is responsible. In order to solve the problem here, the notion of tracing is introduced into the knowledge condition. Vargas uses the following condition as an exemplar throughout his argument:

> (KC) For an agent to be responsible for some outcome (whether an action or consequence) the outcome muse be reasonably foreseeable for that agent at some suitable *prior* time. (Vargas, 2005, 274)

According to KC, moral responsibility depends not on an agent's epistemic state at the time when she acts, but on the epistemic state at a time prior to her action. The notion of tracing is shown in the process of tracing an agent's epistemic state that influences her moral responsibility from a time when she acts to a time prior to her action. The tracing will attribute moral responsibility to her even though she cannot foresee the outcome of her action when she acts. For instance, in the drunken case the danger of driving home in the state of drunkenness is foreseeable for him when he decides at a prior time whether to have a drink, Luis is responsible for running over the pedestrians (Vargas, 2005, 270).

Although the notion of tracing works well in the drunken case, Vargas argues that there are still some cases where it does not work. He envisages four problematic cases to show this, one of which is "Jeff the Jerk": Jeff is a middle-aged manager in a company and is a jerk; one day he laid off his employees in an altogether rude and insensitive fashion (Vargas, 2005, 271). Vargas thinks that the best candidate for tracing is the moment when Jeff decided whether to become a jerk at the age of 15 for the sake of attracting his female classmates (Vargas, 2005, 275-276). Vargas holds that it is not reasonable for a 15-year-old boy to foresee that, as a middle-aged manager decades later, he would lay off his employees in a rude way. If so, Vargas argues, Jeff would not be responsible for his way of laying off employees (Vargas, 2005, 277). However, our intuition is that he is responsible. Thus, Vargas concludes that either we accept that people are less responsible than we often think, or we need to refine the current knowledge condition to avoid such scenarios as Jeff the Jerk (Vargas, 2005, 287-288).

Fischer and Tognazzini (F & T) challenge Vargas' conclusion. They argue that there *is* suitable prior time when Jeff can be reasonably expected to foresee the relevant outcome, the one for which they think Jeff can be held responsible. The key, they indicate, is how to describe the outcome (F & T, 2009, 537). They list three possible ways of describing it.

> (Outcome 1) Jeff fires *those* employees who work for *that* company on *that* precise day in *that* precise manner.

(Outcome 2) Jeff fires some of his employees at some company or other at some point in the future in a despicable manner as a result of his jerky character.

(Outcome 3) Jeff treats some people poorly at some point in the future as a result of his jerky character. (F & T, 2009, 537)

From Outcome 1 to Outcome 3 the description becomes more and more general. They claim that Outcomes 1 and 2 "set the epistemic bar too high" (F & T, 2009, 537), but Outcome 3 is one which he can be reasonably expected to foresee when he decided to become a jerk. They explain that this is because it is reasonable for *a 15-year-old boy* to foresee that he would treat some people poorly in the future as a result of his jerky character (F & T, 2009, 538).

Having explained Vargas' argument and F & T's response, I will suggest a weak knowledge condition (WKC) to deal with cases like Jeff the Jerk where tracing seemingly cannot explain moral responsibility which we want to attribute to him.

(WKC) For an agent to be responsible for a bad outcome O, (a) the agent must be able to foresee that her decisions at actions at time 1 will bring about *a certain state X*, and (b) the agent should have known that X risks O at time 2 in the future.

I will argue, for instance, that if Jeff could foresee his decision's outcome—*a jerky character*, then he is responsible for his later way of firing his employees, given that it is his jerky character that causes his later insensitivity when he fires his employees. There is no need to ask whether Jeff could foresee the *action* of firing the employees in an offensive way in Vargas' particular description, or the *action* of treating some people poorly in F & T's general description 20 years later when he decided to become a jerk. What matters is whether he could foresee that his choice would lead to a bad character. If he could foresee it but still made a decision to become a jerk, it is he who got himself into a situation where in his later life he is insensitive to the relevant moral concerns. On the other hand, I will argue that if he could not foresee that his choice would lead to a bad character, then he is not responsible for his way of firing his employees.

A Weak Knowledge Condition (WKC)

To begin with, I will present two reasons why there is no need to ask whether Jeff could foresee any of Outcome 1, 2, or 3 when he decided to become a jerk. First, I think that Vargas can easily respond to F & T's rejection. In F & T's view, when an appropriate description of the relevant outcome is available, like Outcome 3, we can reasonably trace Jeff's responsibility back to the moment when he decided to become a jerk. It is noteworthy that the success of the tracing then hinges on the plausibility that a 15-year-old boy could have foreseen Outcome 3. If there is no decisive consensus on whether it is plausible for a 15-year-old boy to foresee Outcome 3, Vargas could revise Jeff's case a little bit so that the decision to become a jerk was made at the age of 13, 10 or even younger. The younger the decision was made, the less plausible he could have foreseen Outcome 3.

Second, asking whether an agent could foresee any description of a remote outcome seems to miss a crucial point that tracing involves. Perhaps the case about Luis could help figure out the point. When it is shown that Luis is insensitive to the relevant moral concerns when he runs over some pedestrians, we look for the cause of his insensitivity. Once we know the cause is drunkenness, we ask whether Luis could foresee the outcome of drinking, i.e. drunkenness, when he decided to have a drink. By analogy, when it is shown that Jeff is insensitive to the relevant moral concerns when he fires his employees, we look for the cause of his insensitivity. Once we know the cause is his jerky character, we should ask

whether Jeff could foresee the outcome of behaving offensively, i.e. a jerky character, when he decided to behave offensively. It is enough to ask whether Luis could foresee drunkenness when he decided to have a drink, or whether Jeff could foresee a jerky character when he decided to behave offensively. Hence, I propose a weak knowledge condition (WKC) for attributing moral responsibility.

> (WKC) For an agent to be responsible for a bad outcome O, (a) the agent must be able to foresee that her decisions at actions at time 1 will bring about *a certain state X*, and (b) the agent should have known that X risks O at time 2 in the future.

Here, whether the agent could directly foresee or even take into consideration O when she made a decision at a prior time is not the point. The point is that it is X, the result of *her own* decision that risks O. What we really care about is whether she could foresee the resultant state X when she decided to act in a certain kind of way. A certain state gives her a tendency to act in a certain way after she forms the state. In this sense, when we examine her moral responsibility, we value the question of whether she knowingly forms the state, which makes tracing her responsibility to her knowingly forming the state sensible.

A concern about WKC might arise here that the knowledge that X is bad might also be influenced by the age at which the agent made the decision which brought about X. I admit the possibility that the agent could not know, because of her rather young age, that X is bad at the time when she first made the decision. However, in normal cases it is quite plausible that she got relevant feedback later through her life so that she had the relevant resources to realize that X is bad. I would say that if she got the cognitive resources but decided to overlook them, then the ignorance about the badness of X would not excuse her responsibility. Therefore, unlike Vargas, I think the fact that the 15-year-old Jeff could not foresee that his decision would risk his firing his employees in a rude fashion will not absolve him of responsibility. Unlike F & T, I think neither will the fact that the Jeff who was younger could not foresee that his decision would risk his treating people poorly, because once he later got the relevant resources to realize that X risks O but decided not to make a change, he became responsible. I propose that if the young Jeff could foresee that his decision would brought about X, then he is responsible for later outcome O, granted that X risks O. Thus, my proposal can avoid both Vargas' and F & T's stalemates. I think it demanding for an agent to take into consideration a remote outcome O which is described either in a particular way or in a general way when she made a choice many years ago. In particular, the remote outcome is not the direct result of her choice. Her choice directly results X and then X risks O.

In effect Vargas has already touched the tension between whether the agent could foresee a remote action caused by a certain state which results from her prior choice and whether the agent could foresee the state which results from her prior choice. Vargas seems to talk about KC in both of the two senses. On one hand, as noted above in the case about Jeff, he asks whether the 15-year-old Jeff could be reasonably expected to foresee that he would fire his employees in an offensive way when he decided to become a jerk. That is to say, Vargas thinks that the responsibility for an action should be traced back to a prior time when *the action* was reasonably foreseeable for the agent. On the other hand, when he elaborates on KC, he seems to use it in a weaker sense. He says that "if I cannot currently foresee the outcome of one of my actions, but the reason why is because I knowingly made sure that I would not have that information, we can trace responsibility back to that prior moment when my later ignorance was reasonably foreseeable" (Vargas, 2005, 274). Here he does not trace the responsibility back to a prior moment when *an action* (or an outcome of an action) was reasonably foreseeable, but to a prior moment when *a certain state* (i.e. ignorance) that causes the action was reasonably foreseeable. In an example he brings up, a woman cannot currently foresee the outcome of not immunizing her children, and the reason is that she deliberately decided not to access any information about

immunization (Vargas, 2005, 274). Vargas remarks that her failure to foresee the outcome is not reasonable because she culpably worked herself into a situation where she could not foresee the outcome (Vargas, 2005, 274). Put simply, the ignorance about the outcome of not immunizing is reasonably foreseeable for her when she decided not to learn about the relevant information. This sense of reasonable foreseeability is different from the sense in which putting her children into a physically risky situation as a result of the ignorance is reasonably foreseeable for her when she decided not to learn about the relevant information.

Then I will show why we should use WKC if we want to relieve the tension between KC and our intuition on whether an agent is responsible. In order to do so, I will show the reason why Luis's case is compatible both with WKC and with KC, while Jeff's case is only compatible with WKC. In the sense of KC, we ask whether Luis could be reasonably expected to foresee running over pedestrians when he decided to drink. In the sense of WKC, we ask whether Luis could be reasonably expected to foresee getting drunk when he decided to drink. The answers to both questions are affirmative, because it is much easier or common to connect drinking with car accidents. However, when it comes to Jeff's case, it is problematic to expect 15-year-old Jeff could foresee he would later fire his employees in an offensive way when he decided to become a jerk. The problem derives from a fact that it is rather less common to connect becoming a jerk with firing some people. I think that the reason for the less common connection does not lie in the particularity of firings, because car accidents can be as particular as firings. The reason is that the frequency of becoming a jerk followed by firings is lower than the frequency of drinking followed by car accidents. Thus, the success of tracing in the sense of KC depends on the degree to which the content of a prior decision (drinking or becoming a jerk) is connected with a later action (car accidents or firings) in ordinary life. The lower the degree of connection, the less successful the tracing. Nonetheless, if we still have an intuition that the agent is responsible when the connection is less frequent, there will be a tension between our intuition and the tracing. In order to relieve the tension, I contend that WKC instead of KC is used in tracing.

Moreover, WKC can accommodate cases like the vaccine scenario while F & T's suggestion cannot. Suppose that we take their suggestion to judge the woman's responsibility—we examine whether she could foresee a coarse-grained outcome of her not immunizing her children when she decided not to learn about the relevant vaccine information. Suppose that the coarse-grained outcome is that her children were in a bad physical situation. Now the connection between not accessing vaccine information and her children's illness is not as frequent as the connection between becoming a jerky and treating people poorly. This is because we cannot describe her children's illness as *some people* getting sick, while we can describe firing his employees offensively as treating some people poorly. In the vaccine case, the objects (i.e. her children) toward which she is responsible are an important element that cannot be generalized as *some people* when we judge whether she is responsible. By contrast, in the case about Jeff, the objects (i.e. his employees) toward which he is responsible can be generalized as *some people* without affecting our judgement about whether he is responsible. If the former connection is less frequent, then it is more reasonable that she could not foresee her children's bad health when she decided not to learn about the relevant vaccine information. But we would not be willing to conclude that she is not responsible for her children's illness. Nonetheless, as noted in the last paragraph, WKC can accommodate the vaccine case much more easily. WKC uses X to explain responsibilities for a range of Os without involving how close the connection between decisions at time 1 and (particular or general) O at time 2 is.

To conclude, Vargas illustrates KC in a weak sense in which the responsibility can be traced back to a prior moment when *a certain state* that causes a later action was reasonably foreseeable. However,

when discussing Jeff's case, he utilizes it in a strong sense in which the responsibility can be traced back to a prior moment when *a certain action* (or a certain outcome of an action) was reasonably foreseeable. It is this divergence that makes Jeff's case problematic.

The Judgement of Responsibility According to WKC

In this section I will argue that Jeff's case is no longer problematic when KC is replaced by WKC, and that WKC is more compatible with our ordinary intuition on moral responsibility compared with KC.

Suppose that it is true, as Vargas puts, that Jeff is insensitive to the relevant moral concerns when he fires his employees due to his jerky character. Given that his character makes him insensitive to the moral concerns, according to WKC it is necessary to inquire whether he could foresee that his choice would lead to the bad character when he decided to behave jerkily.

Now I will consider the hypothesis that Jeff could foresee his choice would lead to the bad character. Vargas adds some details about 15-year-old Jeff's situation (Vargas, 2005, 275-276). According to Vargas' description, it is Jeff's own decision to become a jerk so as to attract his female classmates. At the beginning he had worries about some possible consequences of this decision, such as his friends would laugh at his new behaviors or would find out the reason why he changed his behaviors. However, he finally decided to give it a shot. Even to his surprise, the transition was easy, quick and successful. He did not receive any negative feedback from his parents or other people during the transition. And he successfully achieved his goal, i.e. attracting some of his female classmates. According to these details, Jeff had control over what kind of person he wanted to be, because through deliberation he thought that the best way to realize the goal of attracting his female peers was to behave jerkily. Also, because he was worried that other people would perhaps complain his new behaviors, he could foresee his decision would make him a bad person. Thus, he was responsible for becoming a jerk, and even Vargas himself accepts this conclusion. Vargas says that "Jeff was responsible for becoming a jerk" because "he undertook the decision freely, he could reasonably foresee that his resultant behavior might have benefits with his female classmates, and that some of his friends might complain about his new behavior and so on". (Vargas, 2005, 277). In a word, Jeff could foresee his choice would lead to the bad character. Because his bad character risks his way of firing his employees in his later life, he is responsible for firing his employees in the offensive way.

Then I will consider the opposite hypothesis that Jeff could not foresee his choice would lead to the bad character. Suppose he lived in a small town where people treated each other in an offensive manner once doing so would bring them good or would at least not bring them harm. People taught this principle to their descendents. The principle was one custom in the town. People regarded being jerky as a necessary condition to be successful. Jeff observed this kind of phenomenon from his birth. Since he has never seen other ways of acting, he naturally became a jerky person without making a choice to do so. Being a jerky person was the only option he could think of. The custom was so overwhelming that he even did not question it at any one moment. He had no access to information about other ways of living. In this situation it is hard to think that Jeff could foresee his choice would lead to the bad character. The place in which he grew up actually did not provide him with *background knowledge* that being a jerk would be bad.[1] Thus, I would say that Jeff is not responsible if things turned out this way. Since this case

[1] This scenario is inspired by Wolf's case about JoJo:
JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to

is relatively rare, we most often do not consider it when we intuitively judge that he is responsible. Nonetheless, it is noteworthy that this kind of case did happen in the past, such as our misogynistic grandpa or slaveowners in certain cultures where they were brought up to be unquestioning about the status quo.

Suppose that this Jeff once left his town and got negative feedback from strangers enough for him to realize that his jerky behaviors were bad. But he still decided to sustain his old behaviors. Then I would say that this Jeff is responsible since there is a now prior time, namely when he gets this feedback he comes to know that his current character or pattern of behaviors was bad but decided not to make any change.

To conclude, according to WKC, we can judge whether Jeff is responsible according to whether he could foresee the resultant state of his choice to behave offensively, given that the state accounts for his later insensitivity when he acts.

Conclusion

In this paper, I have proposed a weak knowledge condition to break the stalemate Vargas brings up. And the WKC can accommodate more cases than F & T's suggestion that the agent should be able to foresee general actions. These are cases where a certain state does not entail even a general action but still risk the general action, like in the vaccine case. According to WKC, if an agent could foresee that her decisions at actions at time 1 would bring about a certain state X, given that X risks a bad outcome O later at time 2, then she is responsible for O. On the other hand, there exist some rare cases where the agent could not foresee that her decisions at actions at time 1 would bring about a certain state X because she is in a situation where she is educated not to question what she has been taught. Then even though X risks a bad outcome O later at time 2, she is not responsible for O.

References

Fischer, John and Tognazzini, Neal. 2009. The Truth about Tracing. *Nous* 43(3): 531-556.
Vargas, Manuel. 2005. The trouble with tracing. *Midwest Studies in Philosophy* 29 (1): 269-290.
Wolf, S. 1987. Sanity and the Metaphysics of Responsibility. In Ferdinand Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge University Press: 46-62.

accompany his father and observe his daily routine. In light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things, he acts according to his own desires. Moreover, these are desires he wholly wants to have. When he steps back and asks, "Do I really want to be this sort of person?" his answer is resoundingly "Yes," for this way of life expresses a crazy sort of power that forms part of his deepest ideal. (Wolf, 1987, 53-54)

On the Validity of the Consequence Argument

Liang Jin

Introduction

The consequence argument is widely accepted as a successful defense to incompatibilism. Here is one of its expression:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not *up to us* what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not *up to us*. (van Inwagen, 1983: p.16, p.56)

In a recent paper, Marco Hausmann (2018) defines the "unavoidability" and "realizability" and accordingly argues that the consequence argument is not valid. This paper will argue that Hausmann's definition involves a significant mistake and develop a better definition. By doing so, this paper will provide a defense of the validity of the consequence argument.

The Consequence Argument and Some definitions

The world is governed by determinism, if and only if, "For every instant of time, there is a proposition that expresses the state of the world at that instant. If p and q are any propositions that express the state of the world at some instants, then the conjunction of p with the laws of nature entails q." (van Inwagen, 1983: p.65) Let P_0 represent for a proposition that becomes true at a remote past instant t_0 (assuming it was before humans were born), L for all laws of nature, and P for any true statement of state, the following formulation can be thought of a corollary of determinism:

[D] $\Box(P_0 \wedge L) \rightarrow P$)

There are some true propositions that no one can render them false. van Inwagen introduced an "unavoidability operator", N, to formulat them:

[N0] $Np =_{df} p$ and no one has, or ever had, any choice about whether p.

There are two inferential rules and two highly plausible promises about N:

[α] $\Box p \vdash Np$
[β] $Np, N(p \rightarrow q) \vdash Nq$
[1] NP_0
[2] NL (van Inwagen, 1983: p.94)

According to these propositions, van Inwagen formulizes the consequence argument as follows:

(1) $\Box(P_0 \wedge L) \rightarrow P$) D

(2) $\Box(P_0 \rightarrow (L \rightarrow P))$ (1), *Exportation*

(3) $N(P_0 \rightarrow (L \rightarrow P))$ (2), α

(4) NP_0 1

(5) $N(L \rightarrow P)$ (3), (4), β

(6) N*L* 2

(7) N*P* (5), (6), β (van Inwagen, 1983: pp. 94-95)

According to a common understanding, having free will means have control on something. But the conclusion of the argument (N*P*) indicates that nothing is up to us. Therefore, if determinism is true, no one has free will.

Afterwards, van Inwagen redefined unavoidability as:

> [N1] N*p* = *p* and every region to which anyone has, or ever had, *exact* access *is a sub-region* of *p*. (van Inwagen, 2000: p. 8)

This definition is developed in the light of *logical diagram* (which will be explained below). In recent papers, he rewrites it more comprehensibly:

> [N1*] An untouchable proposition is a true proposition that is such that nothing that anyone is or ever has been able to do *might* have had the consequence that it was false. (van Inwagen, 2008: p. 452)[1]

By explaining "have choice about" in N0, some misunderstandings and rebuttals are avoided.[2]

No matter how unavoidability is defined, N*p* is combined of "*p*" and another proposition. The latter generally be written as "it is not realizable that ¬*p*", and be formulized by the means of the *realizability operator*, R, which is interchangeable with N via the following rule:

> [NR] N*p* = $p \wedge \neg R \neg p$

In the light of NR, the realizability that corresponding to N1 and N1* can be defined respectively as follows:

> [R1] R*p* = There is someone who has, or ever had, exact access to a sub-region of *p* or a region that overlaps with *p*.
>
> [R1*] R*p* = There is at least one agent who is (or has been) able to *φ* such that, if he *φ*-ed, it *might* have had the consequence that *p*.

The definition of R reflects understandings towards the non-causal relationship between an agent's ability of act and the truth value of a proposition: by performing the action, the

[1] The term "untouchable" can be treated as equal to "unavoidable". It is a little confusing why the definition N1 can be rewritten as N1*. I will explain that later.

[2] The most well-known refutation among these can be seen in McKay & Johnson, 1996.

agent renders a proposition true or false. In a recent paper, Marco Hausmann assumes that the relationship is *metaphysical grounding*, a reason-explanatory relationship: the truth of the proposition is metaphysically grounded in the of action, and the action can be counted as a reason why the former obtains. Accordingly, he rearranges van Inwagen's definition (R1 and N1) as:

> [R2] R*p* = There is someone who is (or has been) able to do something such that, if he did it, it *might* be a reason why *p*.
>
> [N2] N*p* = *p* and it is not the case that there is someone who is (or has been) able to do something such that, if he did it, it *might* be a reason why ¬*p*. (Hausmann, 2018: p. 4937)

Hausmann claims that it is vague to say an action "*might* be a reason why …", so it should be discussed under two situations: (A) the action *only might* be a full/sufficient reason why *p*; or (B) the action *also might* be a partial/insufficient reason why *p*. Accordingly, there are two possibly correct definitions of R:

> [R2A] R*p* = There is someone who is (or has been) able to *φ* such that, if he *φ-ed*, *φ-ing might* be a *full* reason why *p*.
>
> [R2B] R*p* = There is someone who is (or has been) able to *φ* such that, if he *φ-ed*, *φ-ing might* be a *full-or-partial* reason why *p*. (Hausmann, 2018: p. 4938)[3]

Then he argues that based on any of those two definitions, the consequence argument can be demonstrated as invalid.[4]

Analysis: Why Hausmann Is Wrong

From Logical Diagram to "Might"

The first step of analyzing Hausmann's definitions is to see whether and how R2 correctly grasps the idea that N1 tries to grasp. According to van Inwagen (2000), N1 is understood by means of the *logical diagram*. The last "*p*" in N1 stands for a region in the *logical space*, which means "the ensemble of logical possibilities, a universe composed of all possible-and-existing states of affairs and all possible-and-non-existing states of affairs" (Bunnin & Yu, 2009: p.398). The region *p* is a union region of all regions that contains the state of affairs *p*. If an agent is able to *φ* and, as a consequence of *φ*-ing, the proposition *p* becomes true, he has access to region *p*. By "has *exact* access to *p*", van Inwagen refers to such a situation that one has access to region *p* but does not have access to any of its sub-regions.
According to R1, there are two situations for an agent to be able to render *p* true. They can be diagramed in figure 1:

[3] The corresponding definitions of N can be easily deducted from definitions R2A and R2B, so I will not present them here.

[4] I will not introduce Hausmann's argument further, because that is not closely connected with my thesis.

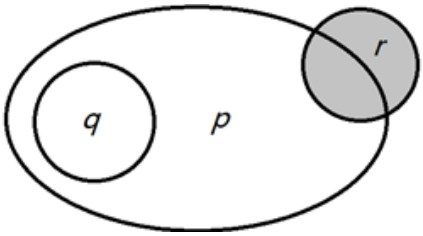

Figure 1

If an agent has exact access to a sub-region of *p*, such as *q*, he is able to *φ* and by *φ*-ing he *ensures* that *p*, so he is able to render *p* true. On the other hand, if he has exact access to a region that overlaps with *p*, such as *r*, he is able to *φ*, and by *φ*-ing what he can *ensure* is that it *might* be the case that *p*[5] In this case, the agent is also considered as able to render *p* true. To sum up, R1 means that it is in one's power to render a proposition *p* true if he is able to *φ* in such a way that his *φ*-ing would have a consequence that the *p might* be true.
This explanation is similar to R2 and, I believe, equivalent to R1/R1*. Therefore, I think this is one reason why Hausmann believes that he can rewrite R1 as R2. It appears that another reason why so is that Hausmann believes that the expression "it *might* have had the consequence that *p*" in R1* is equal to "it *might* be a reason why *p*" in R2. I will now check whether it is right.

Full Reason, Partial Reason, and "Might"

Hausmann's definitions are developed in terms of the *metaphysical grounding*, which is defined as such:

> *f* is fully grounded in *g* = *g* is a reason why *f*, and, as a matter of metaphysical necessity, *f* obtains if *g* does.
>
> *f* is partly grounded in *g* = *g* is a reason why *f*, and, it is metaphysically possible that *f* fails to obtain even if *g* does. (Correia and Schnieder, 2012: p.21)

Agents render a proposition true by performing an action that provides reasons to it. "*φ*-ing *might* be a reason why *p*" in R2 means that *p* is partly grounded in *φ*-ing, viz., there might be some possible worlds in which *φ*-ing is not a reason why *p*. I can come up with two such situations. The first is that the "grounded in" is unfit for the relationship between *φ*-ing and *p* because metaphysical grounding is *factive* while *p* fails to obtain.[6] The second possible situation is that there is no metaphysical grounding relationship between *φ*-ing and *p* — even if they both obtain, *φ*-ing is not a reason why *p*.

Now a contradiction emerges: the second situation does not consistent with van Inwagen's understanding of realizability. As I analyzed before, according to R1, it is in one's power to render a proposition *p* true if he can *ensure* that his *φ*-ing would have a consequence that it

[5] According to van Inwagen, the diagram can also express the probability of *p*'s being true: it equals to the ratio of the area of the overlap between *r* and *p* to the area of *r*.
[6] "Factive" is such a property of metaphysical grounding that if *f* grounded in *g*, both *f* and *g* must obtain.

might be the case that *p*. If *φ*-ing and *p* both obtain, I think the definition R1 must imply that *φ*-ing is a reason why *p*. Because, if not, I cannot see in what sense the agent can *ensure* that his *φ*-ing makes it might be the case that *p*.

In other words, there is a crucial difference between van Inwagen's understanding of realizability and Hausmann's. When van Inwagen mentions that "*φ*-ing might have a consequence that *p*", he is talking about two situations: (1) it must be the case that *p* when the agent *φ*-ed; and (2) it may not be the case that *p* when *φ*-ed. Both situations presuppose that *φ*-ing is a reason why *p*. However, suppose that both *φ*-ing and *p* obtain, Hausmann's wording "*φ*-ing might be a reason why *p*" involves three possibilities: (a) *φ*-ing is a full reason why *p*; (b) *φ*-ing is a partial reason why *p*; (c) *φ*-ing is not a reason why *p*. The situation (1) is corresponding with the possibility (a) while (2) is corresponding with (b). The possibility (c) has no counterpart. It is an extra that Hausmann brings when he rewrote van Inwagen's definition.

We can also consider this in terms of logic diagram. It is not difficult to see that possibilities (a) and (b) correspond to regions *q* and *r* in figure 1. Possibility (c), in the other hand, adds another possible region in figure 1, and hence forms figure 2:

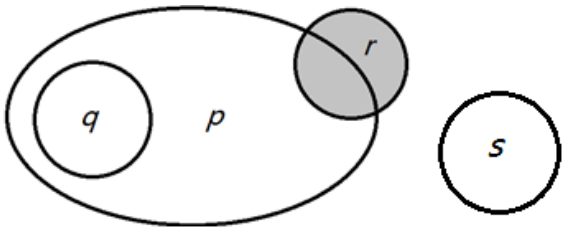

Figure 2

In situation (c), *φ*-ing is not a reason why *p*, so the corresponding region *s* to which the agent has exact access is neither a sub-region of, nor a region that overlaps with, *p*. Therefore, the reason why Hausmann is wrong is that his definition of realizability implies a consequence that van Inwagen's definition cannot accept.

Understanding Realizability

A New Definition of Realizability

Ruling out the possibility (c), viz., ruling out that "*φ*-ing is not a reason why *p* even if they both obtain", I find that the difference between R2A and R2B does not matter anymore: if *φ*-ing *might* be a *full* reason why *p*, it also *might* be a *full-or-partial* reason why *p*. Because if *φ*-ing *might* be a *full* reason, while cannot be a *full-or-partial* reason, why *p*, the only possibility that left for *φ*-ing would be "being a *full* reason why *p*", which means that it is necessary for *φ*-ing to be a *full* reason why *p*. That is not a correct understanding of R1/R1*. Therefore, R2A is not a worth-considering form of R2. My argumentation is going to focus on R2B.

As I have said, the "full reason" in R2B corresponds to the "a sub-region of *p*" in R1 while the "partial reason" corresponds to "a region that overlaps with *p*". However, the "might" in R2B corresponds to nothing in R1 but brings an unacceptable consequence. Therefore, the "might" in R2B is unnecessary and misleading. Deleting the "might", R2B can be revised as follows:

> [R2*] R*p* = There is at least one agent who is (or has been) able to *φ* such that, if the history and laws of nature have been fixed, if *p* is true and the agent *φ*-ed, *φ*-ing *would* be a *full-or-partial* reason why *p*.

I believe that R2* grasps what van Inwagen intends to express via R1 more successfully than R2.

However, R2* cannot successfully respond to Hausmann's refutation against the consequence argument. Let's see how Hausmann's refutation (2018) works. He finds out a logical consequence of rule α and β, the *principle of disjunction*: N*p* ⊢ N(*p*∨*q*). Then he argues that the *addition of realizability*, R*p* ⊢ R(*p*∧*q*), is a valid inferential rule under definition R2B. At last, by the aid of the addition of realizability, Hausmann argues that there are counterexamples against the principle of disjunction: if there is a proposition that its negative is realizable, the principle of disjunction would be incorrect. So, the principle of disjunction is invalid. The failure of the principle of disjunction indicates that either α or β is wrong.

I think his argument is right. However, if we use R2* to replace R2B, the addition of realizability remains highly plausible: if *φ*-ing is a full-or-partial reason why *p*, it is a full-or-partial reason why *p*∧*q*.[7] Therefore, If R2* is the right understanding of realizability, the consequence argument would be invalid.

But I do not think the consequence argument fails, because I believe that the definition R2*, though more plausible than R2, is wrong. There are two reasons for that. The first is that R2* fails to meet some intuitive understandings of realizability. Consider the following case:

> [RESCUE] A child falls into the water beside Tom. Tom cannot swim at all, but he jumped into the water recklessly to save the child. Because of luck (say, the water wave that Tom aroused pushes the child ashore), the child is rescued. Assume that Tom is the only agent in this case.

Let *p* = "Tom tries to save the child", *q* = "Tom has skills of swimming". R*p* is true while R*q* is false.[8] Suppose that other conditions are fixed in such a way that if someone who has skills of swimming tries to save the child would rescue him successfully, *p*∧*q* can be treat as "the child is rescued by Tom". R(*p*∧*q*) is intuitively false. But according to R2* and the addition of realizability, R (*p*∧*q*) is true because R*p* is true.

---

[7] Of course, *φ*-ing may not be a full reason why *p*^*q*, but it must be a partial reason why *p*^*q*. A partial reason is a full-or-partial reason. Therefore, *φ*-ing is a full-or-partial reason why *p*^*q*.

[8] Of course, we can assume that Tom can render *q* true by learning to swim, but at that very instant in RESCUE case, Tom cannot learn to swim immediately. Hence it can be said that Tom is not able to render *q* true at that instant.

Secondly, R2* is not equivalent to R1. The REACUE case can be illustrated by logical diagram in the following figure 3:

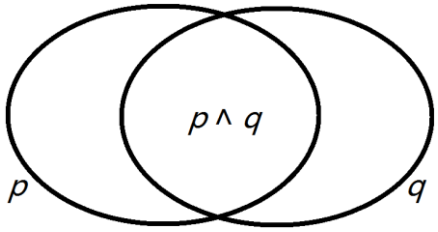

Figure 3

When R*p* is true and R*q* is false, Tom has exact access to *a* sub-region of *p* or a region that overlaps with *p*, but does not have exact access to any sub-region of, or any region that overlaps with, *q*. Since Tom actually cannot swim, I cannot see any chance for Tom to have exact access to a sub-region of, or a region that overlaps with, $p \wedge q$.[9] Therefore, according to R1, as long as both *q* and R*q* are false, R($p \wedge q$) is false even though R*p* is true. The addition of realizability is false according to R1 while is true according to R2*. Hence R2* is not an equivalent rewrite of R1.

To sum up, R2* is neither a good understanding of realizability nor an accurate rewrite of R1. Therefore, even if Hausmann's argument is correct and the consequence argument that based on R2* is invalid, its failure does not imply that the consequence argument itself, or the version that developed based on R1, is invalid.

3.2. A Newer Definition of Realizability

No matter using "might" or "partial reason", the ideas that the definitions R2 and R2* intend to capture are same. That is, an agent can render *every* proposition that expresses a *probable* consequence of his actions true. I think the reason why R2 and R2* fail is that this idea itself is not correct. Now I will examine this idea.

It is more or less vague to say that I am *able* to do something. There are, of course, something I am definitely able to do and something I am definitely not, but there is also something that I am neither definitely able to do nor definitely not. Consider the following three questions in three cases (assume that the protagonist in each case is the *only* agent in the case):

(1) Amy is a surgeon. She has an 80% chance to successfully operate a liver transplant. Is she able to successfully operate today's liver transplant?

[9] It is important to emphasize that Tom actually cannot swim, or he is *not inside* any sub-region of *q* or any region that overlaps with *q*. Because if not, viz., if he can swim, even it is not in his power to render "Tom has skills of swimming" true, it would be possible for Tom to having exact access to a sub-region of *p*^*q* or a region that overlaps with *p*^*q*.

(2) Bob is a gambler. He has a 50% chance to make the coin fall heads. Now he is going to toss a coin. Is he able to make the coin fall heads?
(3) Cathy is a terrible swimmer. She has a 10% chance to survive a flood. There is a flood in her hometown these days. Is she able to survive the flood?

Let *p* = "Amy has successfully operated the liver transplant", *q* = "Bob has made the coin fall heads", *r* = "Cathy has survived the flood". These three events are consequences of Amy, Bob and Cathy's actions, but none of them is a necessary consequence. According to R2*, all answers to these three questions should be "Yes". Intuitively, I think the answer to question (1) is "Yes", while the answers to (2) and (3) are "No". Maybe others have different intuitions. No matter how, these three answers are unlikely to be the same. That means not *every* proposition that is a *probable* consequence of an action is realizable. Which of them are realizable and which are not? This question is related to the vagueness of realizability.

It is difficult to draw a clear line in the "vague zone" and thus separate those propositions that are realizable from those that are not. However, I don't think it is impossible. Among propositions that express possible consequences of my actions, the realizable ones can be thought of as those whose probability of being true would rise to a relatively high level because of my action; the unrealizable ones can be thought of as those whose probability of being true is relatively low and would not rise to a relatively high level because of the action. Correspondingly, the definition of realizability and unavoidability shall be presented as such:

> [R3] R*p* = There is someone who is (or has been) able to *φ* such that, if he *φ*-ed, *φ*-ing would be a reason why the probability of *p*'s obtaining has risen to a relatively high level.
>
> [N3] N*p* = *p* and it is not the case that there is someone who is (or has been) able to *φ* such that, if he *φ*-ed, *φ*-ing would be a reason why the probability of ¬*p*'s obtaining has risen to a relatively high level.

Let me make some comments on this definition. Firstly, the realizability of *p* has nothing to do with the probability of *p*'s obtaining, but with whether the agent's action can make the probability rise. For example, in surgeon Amy's case, Amy is able to operate a liver transplant, not because the success rate of liver transplant is 80%, but because Amy can make the probability of the success of the surgery rise from 0 to 80%.

Secondly, definition R3/N3 still uses the terminology "reason", which means I do not give up the metaphysical grounding method. The factivity of the metaphysical grounding is still valid. Amy's ability of operating a liver transplant means that if she operates it, its chance of success *does* rise to a relatively high level.

Finally, the "relatively" in the above definition means that the realizability is *context sensitive*. Consider Amy over again. If the average success rate of a liver transplant operation is 97%, I would say that Amy is *not* able to operate the liver transplant. Because even though 80% seems high, it is still a lot lower than 97%.

Given the above definition, I can answer Hausmann's refutation against the validity of the consequent argument. According to R3, the inferential rule that Hausmann uses to argue against the consequence argument based on R2B, the addition of realizability (Rp ⊢ R(p∧q)), is no longer valid. Because when the problem of realizability turns into a probability problem, realizability can be calculated according to Bayes Rule:

[Bayes Rule] P(A∧B) = P(A) × P(B|A) = P(B) × P(A|B)

According to the Bayes Rule, the probability of p∧q depends on not only the probability of p's obtaining, but also the probability of q's. Therefore, even if Rp is true (that is, there is an agent who is able to φ such that if he φ-ed, φ-ing would be a reason why the probability of p's obtaining has risen to a relatively high level), as long as Rq is false (viz., the probability of q's obtaining will not increase because of φ-ing), the probability p∧q may not rise to a relatively high level. That is to say, R (p∧q) may still be false. Therefore, if R3 and N3 are correct definitions of realizability and unavoidability, Hausmann's refutation would no longer be a challenge to the validity of the consequence argument.

Conclusion

In this paper, I argue that Hausmann's argument against the validity of the consequence argument fails because his definition of realizability commits two significant mistakes: (1) the "might" he introduces in R2 is misleading; (2) when the "might" is deleted, the new definition R2* does not meet an important intuition about realizability. Then I develop a definition (R3/N3) that meets our intuition better, and this new definition can be used to answer Hausmann's argument. By doing so, I think this paper provides a successful defense to the validity of the consequence argument.

I have to admit that R3/N3 is a tentative definition. More work needs to be done to formalize or defend it. But I will not do that in this paper. What I am doing is only to point out that there is an alternative way (a better way, maybe) to understand realizability and unavoidability, and that the consequence argument remains valid according to this understanding.

References

Blum, A. (2003). The Core of the Consequence Argument. *Dialectica*, *57*, 423-429.

Bunnin, N. & Yu, Jiyuan, (2009). *The Blackwell Dictionary of Western Philosophy*, Oxford: Blackwell.

Correia, F., & Schnieder, B. (2012). Grounding: an opinionated introduction. In F. Correia & B. Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality,* 1–36. Cambridge: Cambridge University Press.

Hausmann, M. (2018). The Consequence Argument unground. *Syneses*, 195 (11), 4931-4950.

Lewis, D. (1981). Are we free to break the laws? *Theoria*, *47*, 113-121.

McKay, T., & Johnson, D. (1996). A Reconsideration of an argument against compatibilism. *Philosophical Topics*, *24*, 113-122.

O'Connor, T. (2016). Free will, *Stanford Encyclopedia of Philosophy*. URL = <http://plato.stanford.edu/archives/sum2016/entries/freewill/>
Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
—— (2000). Free will remains a mystery. In J. E. Tomberlin (ed.), *Philosophical Perspectives, 14*, 1-19. Oxford: Blackwell.
—— (2008). The Consequence Argument. In P. Van Inwagen & D. Zimmerman (eds.), *Metaphysics: The Big Questions,* (2nd ed.), 450-456. Oxford: Blackwell.
Widerker, D. (1987). On an argument for incompatibilism. *Analysis*, 47, 37-41.

On Moral Responsibility for Implicitly biased behaviors

Xiaolong Wang

Introduction

This paper will focus on the the question about whether people is morally responsible (i.e., should be blamed with moral reactions like resentment, indignation, anger, etc.) for their immoral behaviors influenced by their implicit bias which, as a cognitive state in human mind discovered by psychologists, refers to a kind of unconscious and prejudiced attitude toward people in various social groups classified by gender, sexual orientation, age, race, religion, region, illness and so on. (Cf. Brownstein, 2017) In considering this question, I will evaluate the possible answers according to *the merit-based view*, i.e., that one is morally responsible is justified if he/she satisfies some conditions and therefore deserves blame.
[1] (Cf. Eshleman, 2016) Thus, the work includes two steps: first, to find out the best theory of the conditions of holding moral responsibility for behaviors; second, to test whether implicitly biased behaviors satisfy these conditions and therefore should be blamed.

Here is how I proceed to answer the main question. In section 2, I will introduce The Standard Theory that there are two conditions needed, i.e., epistemic and control, and then introduce the externalism's interpretation on the epistemic condition and the ecological control view's interpretation on the control condition. To make them stronger, I combine the two interpretations into a specific version of The Standard Theory named The Combined Theory, given which people should take moral responsibility for their implicitly biased behaviors if the knowledge of implicit bias is available in the environment. In section 3, I will argue against The Combined Theory on both the epistemic and control condition. In section 4, by modifying The Combined Theory in order to avoid the problems of it, I will formulate and defend my suggested theory named The New Combined Theory, whose answer to the main question is that people should take moral responsibility for their implicitly biased behaviors if the knowledge of implicit bias is available in the environment and they have acquired the knowledge.

The Combined Theory

The Standard Theory

For the first step to consider the question, what is the potentially best theory of holding moral responsibility for behaviors? An appealing and influential answer is The Standard Theory, initiated by Aristotle and developed by many philosophers, holding that one should take moral responsibility for his/her behavior if *both* the conditions below are satisfied:

[1] I will not evaluate according to the consequentialist view, that is, that one is morally responsible is justified if the blame would bring desired consequences, e.g., changing the subject's behavior. (Cf. Brownstein, 2017)

(1) *the epistemic condition*: if one is aware of what he/she is doing and the potential consequences and moral valence (moral or immoral) of the behavior, or one does not have the awareness but should have had it. (Cf. Rudy-Hiller, 2018)
(2) *the control condition*: if one has control over his/her behavior, or one does not have the control but should have had it. (Cf. Eshleman, 2016)

According to this theory, it seems that people should not take moral responsibility for their implicitly biased behaviors, because they are not aware of implicit bias and therefore can not control, and there is no good reason to say they should have had the awareness and control. However, since the standard view is too general or somewhat vague–crucially, (Q_1) what kind of awareness and control is needed and (Q_2) what is the criterion of judging that one should have had the awareness and control–it still leaves specific judgments on implicitly biased behaviors open to question. Thus, in answering the two questions above, philosophers developed different interpretations.

Interpreting The Epistemic Condition: The Externalism View

Washington and Kelly (2016) propose an externalism interpretation of the epistemic condition of The Standard Theory. They maintain the epistemic condition is that one is aware of what is done and the consequences and moral valence, or one does not have direct or indirect awareness but should have had it, *if the knowledge of implicit bias is available in the epistemic environment*, *and he/she occupies the relevant social role*.

This is an externalism answer to (Q_2), i.e., whether people should have been aware depends not only on their own or internal ability but also on their external epistemic environment and social roles. The externalists support this by raising two different cases. (Cf. Washington and Kelly, 2016) Suppose a hiring committee, composed of white men, is browsing the resumes from candidates (varying in races) to decide who to employ. Despite they are explicit anti-racialists and deliberate carefully, it turns out that they implicitly prefer the white-sounding names and accordingly more white candidates are employed. In case A, the hiring committee made the decision in 1930, when people knew nothing about implicit bias. While in case B, the committee did the same thing in 2019, when there have been more and more researches, publicity and presentations, and they ever heard of the term 'implicit bias' but never went to learn more. Externalists assert that we tend to hold that only the committee in case B should take moral responsibility, because when the knowledge is available, the committee as a social role whose function is to judge justly, they have the obligation to or should acquire the knowledge.

Interpreting The Control Condition: The Ecological Control View

Some philosophers (Holroyd, Scaife and Stafford, 2016; Holroyd and Kelly, 2016) provide an ecological control view in interpreting the control condition of The Standard Theory. They hold that the control condition means one has *direct or indirec*t control over his/her behavior, or one does not have any control but should have had it. Having direct control means people can currently intervene in the behavior on their own while having indirect control means they can intervene in the behavior with the help of some props often in advance.

This view also asserts people do have a special kind of indirect control over implicitly biased behaviors, i.e., ecological control, which means intervening the behaviors (including the implicit biased ones) by appealing to some ecological props that exploit or change the subliminal mechanisms and factors in the environment. (Cf. Holroyd and Kelly, 2016) Implicit bias, as a member of subliminal mechanism, is built in the human mind in the 'if-then' form, e.g., the implicit racial bias that "if the name on the resume is black-sounding, then this is a lazy guy". Thus theoretically, the ecological props could exploit the mechanism of implicit bias to intervene in two ways: (1) changing the environment or adopting some cognitive procedures in order to prohibit the occurrence of the antecedent, e.g., using special techniques to hide the names on the resume; (2) changing the relation of implication between the antecedent and consequent through cognitive training and practices. And there are such ecological props in reality, such as counter-stereotypical pictures, cognitive procedures, implementation intentions, training program and so on. (Cf. ibid)

The Combined Theory

I think the externalism view and the ecological control view form a natural combination and so I will combine them to formulate The Combined Theory and illustrate its pros. This theory, as a specific version of The Standard Theory, holds that one should take moral responsibility for his/her behavior if *both* the conditions below are satisfied:

(1) the epistemic condition: one is *directly or indirectly* aware of what is done and potential consequences and the moral valence of the behavior, or one does not have the awareness but should have had it, if the relevant knowledge is available and he/she occupies the relevant social role.
(2) the control condition: one has direct or ecological control over his/her behavior *if the relevant knowledge, skills, and props are available and he/she occupies the relevant social role*., or one does not have the control but should have had it.

The Combined Theory makes the externalism view tidier because it borrows the distinction between *direct or indirect* from the ecological control view. Having direct awareness means people are aware of something through consciousness and having indirect awareness means they are aware *only* by knowledge. For example, now I am directly aware of my feeling of hunger, and I am indirectly aware that there are many neural activities in my brain (by neuroscience knowledge) though I can not be directly aware of them. This view is reasonable because we would intuitively hold the hiring committee morally responsible if they have had the knowledge of what implicit bias is and its features.

And The Combined Theory helps the ecological control view solve an apparent problem. The ecological control view says people do have ecological control, but what about those who have not acquired the props like knowledge and skills yet? Now The Combined Theory could say that they do have because the props have been available in the environment.

So according to The Combined Theory, people should take moral responsibility for their implicit behaviors if the knowledge and props are available in the environment and he/she occupies the

relevant social role., which gets them into a situation where they have ecological control because they have and should have been indirectly aware of implicit bias.

Objection to The Combined Theory

Objection to The Combined Theory on The Epistemic Condition

One of the claims from The Combined Theory about The Epistemic Condition is that (S1) one should have been indirectly aware of the implicitly biased behaviors, if the knowledge of implicit bias is available in the epistemic environment, and he/she occupies the relevant social role. Basically, I will argue that (S1) is too demanding and very likely to lead to the injustice of blame. Here is my argument:

> (P_1) If one can *not* do something, i.e., does not have the ability (physical capacity, knowledge, and skill) and opportunity (from now on I will use 'can' in this sense), then it is *wrong* to say he/she should do that. (Vranas, 2007) (The should-implies-can principle)
>
> (P_2) People can not be indirectly aware of the implicitly biased behaviors, even if the knowledge of implicit bias is available in the epistemic environment, and he/she occupies the relevant social role.
>
> (C) Therefore, (S1) is wrong.

Vranas (ibid) defends that version of should-implies-can principle by drawing solid arguments. If denying (P_1), to be consistent, one has to accept some counterintuitive claims, e.g., John should jump into the lake to save the drowning kid, although John can not swim.

(P_2) is also true. First, even though the knowledge of implicit bias is available in some epistemic environments, the social roles there still can not fully determine themselves to acquire the knowledge. It is contingent that someone comes to have the knowledge from the newspaper while someone does not often browse any media and not come to know. If The Standard Theory says one can be indirectly aware of something or have the knowledge based on contingency, then how could the obligation be justified on the basis of contingency or (epistemic) luck? If so, it would lead to further problems with a special kind of social injustice. In the information era, once discovered, the knowledge would spread quickly and pervasively and thus be available in a large epistemic environment. However, it seems unjust to say both the hiring committees in a small developing country and New York (the existing locations are contingent features), who have not yet acquired the knowledge of implicit bias, have the obligation or should acquire the knowledge. Although The Combined Theory could say the knowledge in New York is more available and therefore the committee has a stronger obligation to acquire it, it remains difficult to measure the degree of availability in everyday blaming practice, and it is still too demanding for the committee in the small developing country (maybe even also for the one in New York). This objection also works well in other cases. Once some new medical discoveries in some fields have been published on the top journal, do we think all the doctors (as special social roles) in that field in the world, who are accessible to the internet and media, have the obligation or should come to grasp them? The intuitive answer would be "No" because there are 'less perfect' (in the professional sense) doctors everywhere without the necessary obligation to be 'perfect' doctors.

Objection to The Combined Theory on The Control Condition

The first claim here I will reject is that (S2) having indirect control is sufficient for one to be morally responsible. This claim would get many blameworthy behaviors off the hook since people always have some indirect control over their behaviors. Suppose that John invites his friend Amy to his home to enjoy a movie. During the movie, a robber destroys the wooden door of the house, enters their room, puts the pistol into John's hand and manipulates John's hand to shoot at Amy, and finally takes all the properties away. To our intuition, John should not be morally responsible for Amy's death, but if (S2) is true then John should, because John has some indirect control over this emergency, e.g., he could have changed a stronger secure door for his house as a prop to intervene in the occurrence of the tragedy.

The second claim needed to be rejected is that (S3) one has ecological control if the relevant knowledge, skills, and props are available. To say one has ecological control or can ecologically control means one has the ability (physical capacity, knowledge, and skill) and opportunity to ecologically control. I reject this in a similar way in which I did in the last part, that is, even though the knowledge and skills (cognitive props) and opportunities (environmental props) related to implicit bias are available in the environment, people still can not determine themselves to acquire them, and they do not deserve blame for lacking them because there are too many contingent factors out of control.

If The Combined Theory maintains that people can acquire them and then reason that people should be morally responsible, then it is even more demanding and unjust than the one in the last part, because having ecological control involves social and economic capital. Consider this case. Suppose that for a whole process of hiring interview, it turns out that the hiring committee might have many kinds of potential implicit biases which could be triggered by features of the candidates together with some factors in the environment, and in this world all of the cognitive and environmental props have been discovered and developed well in the academia and therefore are available. But most these props, e.g., "eliminating implicit bias" training program, lectures on skills about blocking implicit bias and environmental devices, are not free of charge. Now a big company A can afford and equips its hiring committee with these props, and as a result, the committee does not have implicitly biased behaviors. While a small company B can not afford this (or they would go bankrupt) and its hiring committee does have some implicitly biased behaviors. Given this view, company B would be held morally responsible. But the only difference between company A and B is on their capital. How could morality be a matter of money?

4. The New Combined Theory

I have argued that The Combined Theory is problematic to some extent and therefore should not be adopted to determine whether people should take moral responsibility for their implicitly biased behaviors. Now I will try to modify The Combined Theory into The New Combined Theory, trying to avoid the problems found before.

The New Combined Theory, also as a new version of the standard view, holds that one should take moral responsibility for his/her behavior if *both* the conditions below are satisfied:

(1) the epistemic condition: one is directly or indirectly aware of what is done and potential consequences and the moral valence of the behavior, or one does not have the awareness but should have had it, if the relevant knowledge is available and *he/she has already had it*.
(2) the control condition: one has *direct control or ecological control (only for the subliminal mechanism causes like implicit bias)* over his/her behavior, if the relevant knowledge, skills, and props are available and *he/she has already had them*, or one does not have the control but should have had it.

Let me explain the idea. As an interpretation of The Standard Theory, for (Q2), The New Combined Theory holds that direct awareness or indirect awareness (knowledge) is needed for the epistemic condition while direct control or ecological control only on subliminal activities (not general indirect or ecological control) needed for the control condition. (Notice, this says nothing about whether people do have the awareness and control.) The former part of this answer fits with our intuition that one should take moral responsibility for the behavior if he/she has knowledge of it (even if he/she has no direct introspection); and the latter is not susceptible to the fact that people always have some general indirect control and thus avoids problems like the one in the shooting case above.

And The New Combined Theory modifies the enlightening but somewhat problematic externalism view from The Combined Theory into a new version which I name "soft externalism view". Here is the idea. It is an externalism view of the condition of moral responsibility because whether one is morally responsible depends not merely on his/her own situations but also on the external environment. Human beings are experiencing the "progress" of knowledge and morality, because empirical research has discovered and will keep exploring what people can do and can not do, and according to the should-implies-can principle, people come to know what they do not have the obligation to do, probably what they should do and what they are or are not morally responsible for. However, even if academic research produces the knowledge, skills, and techniques, it is too demanding to say people should acquire all of the discovery because sometimes they can not do that on their own and there are many contingent factors out of control. Thus, this view is "soft" because though accepting the externalism view, it adds an extra condition, and hence the view is that one should know the discoveries or one has the ecological control if the discoveries are available in the environments *and he/she has already known them*. That is to say, even after the discoveries are made and thus available, we still should not blame one for his/her implicitly biased behaviors until he/she come to know the discoveries about what it is and grasp the props about how to intervene.

Conclusion

In this article, for the question about moral responsibility for implicit bias, I firstly introduce The Combined Theory, according to which people should take moral responsibility for their implicit behaviors if the knowledge is available in the environment. Then I raise objections to this view and finally defend The New Combined Theory, given which people should so if the knowledge and props are available in the environment and they have had the knowledge and ecological props.

References

Vranas, Peter B.M. "I ought, therefore I can." *Philosophical Studies* 136, no. 2 (2007): 167-216.
Holroyd, J. D., Robin Scaife, and Tom Stafford. "Responsibility for implicit bias." *Philosophy Compass* 12, no. 3 (2017): e12410.
Holroyd, J. D., and Daniel Kelly. "Implicit bias, character and control." (2016).
Washington, Natalia, and Daniel Kelly. "Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias." (2016).
Brownstein, Michael, "Implicit Bias", The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/ implicit-bias/>.
Rudy-Hiller, Fernando, "The Epistemic Condition for Moral Responsibility", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/>.
Eshleman, Andrew, "Moral Responsibility", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/ moral-responsibility/>.