Aporia

VOL. **XXI**

Aporia

Undergraduate Journal of the St Andrews Philosophy Society

VOLUME XXI

Aporia *is funded by the University of St Andrews Philosophy Society, which receives funds from the University of St Andrews Department of Philosophy, the University of St Andrews Students' Association, and independent benefactors.*

Letter from the Editor

Hello Folks,

I am Nigel Mika, the interim head of Aporia—I took over after the previous head editor(s) got a bit too busy to keep up with things—and I am a second-year philosophy and mathematics student from a small town in lower Michigan called Lowell. I came to St. Andrews due to the phenomenal philosophy, and more specifically, logic course they have, and I have never looked back. With the pandemic year this edition has been especially difficult to pull together due to the whole edition basically running via email chains with virtually no in-person contact at all. For this reason, I would like to give special thanks to everyone who has worked on this edition and persevered through this extremely difficult year. Especially my Co-interim editor Roberto who has really buckled down and done far more then I could have ever asked for. We have a really strong team. We have been looking forward to sharing this edition for a long while now and we are so glad to finally get it in your hands! A special thanks to the authors also for putting up with the many layers of bureaucratic nonsense this pandemic has brought on us which has only compounded by a lack of proper communication.

I would also like to thank the previous editors for helping with the transition into Roberto and my own hands, thank you Andrew and Isaac. I am also indebted to my parents for all the amazing opportunities, including this one, that they have afforded me; as well as God (in a Spinoza sense,) or the Universe, or the underlying One whatever you would like to call it, to grace me with the faculties and opportunities to go about and create this amazing product. Thank you to every teacher and professor I have ever had, and everyone who has ever encouraged me to follow my passions which have led me to endeavors such as this one. Thank you to everyone I have ever interacted with, considering time might not be real and we may be living these lives over and over again (b-theory to the extreme) it may be great to see you in the next version of this thing we call life and to do the same thing and enjoy it all the same. Thank you, Nicholas Melville,for being the best mentor I have ever had and always encouraging my passions and getting me into more and more philosophy at such a young age; as well as keeping me on a good life direction.

All I can say is I am grateful for the opportunity to even be thankful to all these amazing souls in my life in whatever capacity they are present. Now if you have read this whole introduction (it is quite long winded) thank you for the time you have taken to do that! Have a phenomenal read of this edition of Aporia and never stop philosophizing!

Cheers,

Nigel A. Mika

Acknowledgements

Nigel Mika
Editor-in-Chief

Luis Roberto Garcia Martinez
Deputy Editor

Editors

Jonathan Anderson

Isaac Bateman

Emily Berman

Franciszek Bryk

Taylor Cooperman

Cady Crowley

Andrew Fyffe

Dom King

Louisa McDonald

Hanna Palya

Michael Taylor

Sara Weissel

Reviewers

Anushrut Agrwaal Ramakrishnan

Adi Arora

Charlotte Bae

Spyrido Galatis

Ludovic Glasson

Alexander Graham

Wilson Jones

Victor Karl

Amber McIntosh

Louis Phelps

Rahul Prakash

Phillip Sink

Lara Thain

Gabrielle Torreborre

Maria Villanueva Josefina

Eirini Vryza

Clark Wang

Yifan Wang

Harriet White

Katie Pratt
Cover Art

Luis Roberto Garcia Martinez
Cover Design and Typesetting

Contents

Basic Income: The Left Libertarian Case for a Tax on Attention *Adam Binks* 1

'Testimonial Throttling' and Epistemic Injustice *Michael Calder* 8

Metalinguistic Negotiation and its Limits *William Chao* 23

Fictional Truth, Fictional Names: A Lewisian Approach *Sam Elliott* 37

"Deepfakes" and the End of the Photographic Age *Patrick Pan* 52

On the Difference Between Episodic and Autobiographical Memories *Gabriel Zaccaro* 65

Basic Income: The Left Libertarian Case for a Tax on Attention

Adam Binks

University of St Andrews

Taking a left libertarian account of distributive justice, Steiner argues for a basic income funded by a tax on land. In this essay, I argue along similar lines for a basic income funded by a tax on the involuntary drawing of attention. I first argue that the involuntary drawing of a person's attention denies them their liberty to direct their attention. I then show that attention is a production factor in some modern work, taking the paradigm case of advertising. With these premises, I conclude that when attention is a production factor, part of the product is owed to those whose attention was drawn — and extend this to argue for a universal basic income, funded by work which takes involuntary attention as a production factor, and is situated in public spaces.

1 Compensation for Liberty Lost

In this essay, I explore an interesting implication of left libertarianism. Left libertarianism is a theory of distributive justice: it makes claims about how benefits and burdens should be distributed across members of society.[1] Left libertarianism is a variety of libertarianism so, just like the more familiar right libertarianism, it grounds justice in moral (as opposed to legal) property rights. Left and right libertarians both take *self-ownership* as a fundamental moral right. Our bodies are our own property, and therefore, the fruits of our labour are also our property. But natural resources are the fruits of nobody's labour, so who's property are they? This is where left and right libertarianism come apart. Right libertarians say that the first individual to claim or mix their labour with a natural resource 'appropriates' it, making it their property. Left libertarians radically disagree, arguing it does not matter who gets there first, natural resources are owned equally by all — they are our collective natural inheritance.

For the purposes of this essay, let's assume a left libertarian account of distributive justice. As self-owners, we own the fruits of our labour. However, as Steiner argues, for our labour to generate value, some other things are required, which Steiner calls *production factors*.[2] While the nature of these production factors varies between kinds of labour, in many cases, the ultimate antecedents of them are natural resources and the land containing them. Therefore, though we own the portion of our product which is due to our

[1] Vallentyne, Peter, Hillel Steiner, and Michael Otsuka. "Why left-libertarianism is not incoherent, indeterminate, or irrelevant: A reply to Fried." *Philosophy & Public Affairs* 33, no. 2 (2005): 201-215.

[2] Steiner, Hillell. "Compensation for liberty lost: Left libertarianism and unconditional basic income." *Juncture* 22, no. 4 (2016): 293-295.

self-owned labour, there is often another portion which is due to natural resources, which are owned equally by all. Therefore, when any individual uses an area of land, they deny everyone else their equal liberty to use it. As a result, they owe everyone else compensation for this liberty denied — which Steiner argues could be levied in the form of a tax upon land and natural resources which is paid to all as a universal basic income.

In this essay, I make an argument among similar lines: on a left libertarian account of distributive justice, the involuntary drawing of attention is liberty denied. Thus, when human attention is used as a production factor, part of the product is owed to those whose attention was drawn.

2 Involuntary Attention as Liberty Lost

My argument relies on two properties of human attention. The first is that attention is scarce. Selectivity is a central property of attention: Mole describes attention as “the selective directedness of our mental lives”.[3] Some accounts of attention take its selectivity to be due to limitations in the brain’s capacity to process complex properties of multiple stimuli.[4] Meanwhile, other theories take selectivity to be the management of an excess in brain capacity.[5] For this essay, this discussion does not need to be resolved — both views are compatible with my argument. Both approaches take attention to be inherently selective — the amount of information that can be attended to is limited, and attention is a scarce resource.

The second property of attention that my argument rests on is that attention can be voluntarily directed by the subject, or involuntarily drawn by some external factor. This distinction is supported by the psychological literature — for example, Eimer *et al.* define involuntary attention as “processes of attending that are not elicited by intentions but by certain outside events”.[6] When a subject voluntarily directs their attention, this is an exercise of their self-ownership — they freely direct their own powers of perception and action. So, I argue for a liberty to direct one’s own attention. Additionally, voluntary direction of attention is a fundamental prerequisite for the exercise of well-established liberties, such as freedom of speech. To speak freely, the speaker must attend to the content of the utterance, and voluntarily perform the speech act.

Given that attention is scarce, when a portion of a subject’s attention is involuntarily drawn by something, this restricts the extent to which it can be voluntarily directed to other things.[7] Therefore, involuntary drawing of a subject’s attention restricts their liberty to direct their attention, and thereby denies other liberties that depend upon this.

[3] Mole, Christopher, "Attention". *The Stanford Encyclopedia of Philosophy*, Fall 2017 Edition (2017), 1.

[4] See Broadbent (1958), Deutsch (1963).

[5] See Friston *et al.* (2006), Neisser (2014).

[6] Eimer, Martin, Dieter Nattkemper, Erich Schröger, and Wolfgang Prinz. "Involuntary attention." In *Handbook of Perception and Action*, vol. 3. Elsevier Academic Press, 1996, 155.

[7] This claim is empirically supported, such as in Jonides (1981).

3 Attention as a Production Factor

Steiner's argument for a tax on land rests on land being an ultimate antecedent production factor in forms of work — in these forms of production, the producer owes compensation to everyone else for using land which is everyone's common inheritance. However, in some modern forms of work, very little land and natural resources are needed as production factors. Take the case of advertising — for concreteness, a billboard erected beside a motorway. The designer of the advertisement, as a self-owner, owns the fruits of their labour. This is not the total product, however, because a small amount of materials for the signage are a production factor, as is the small plot of land upon which it stands. Far more significant a production factor than these, though, is human attention. No matter how much labour the designer pours into the sign, or how much land and resource is used in its erection, it produces nothing if human attention is not drawn by it. Indeed, its product is roughly proportional to the amount of human attention drawn by it,[8] just as, for Steiner's argument, a coal extraction plant's product is roughly proportional to the amount of coal present in the seam it is built upon. Thus I argue that a significant production factor in advertising work is human attention.

In some cases, people pay attention to an advertisement voluntarily. In many other cases, however, attention is involuntarily drawn by an advertisement — indeed, this is often an explicit goal of marketing designers. Advertisements which employ bright colours, motion, large text, and evocative imagery exploit involuntary attentional processes which automatically direct attention to visual objects with these properties.

4 Basic Income Funded by an Attention Tax

Taking the claims justified thus far, we can now construct the central argument. When a person A uses the involuntary attention of person B as a production factor, A denies B's liberty to direct their attention freely. Therefore, A owes B compensation. After Steiner[9], we can apply the Kantian ends-means injunction – this denial of liberty is unjustified, as A uses B's attention as a means to the end of making a product, so does not treat B as an end in themselves. The material basis of the required compensation is conveniently apparent and calculable: B owns that proportion of the product which B's role as a production factor contributed. As a self-owner, products that are made with B as a production factor are owned by B to the extent that B contributed to the production. This is analogous to Steiner's land case — Steiner argues we should have a 100% tax on the portion of products that are due to natural resources and land, manifested as a 100% tax on the natural value of land, excluding any constructions built upon it.

I now expand this specific one-to-one argument for compensation for attentional

[8] The product is not directly proportional, because advertisements may be targeting specific audiences – an advertisement for wedding photographers may generate no additional product if more people who are not planning a wedding attend to it.

[9] Steiner, Hillel. "Silver spoons and golden genes: Talent differentials and distributive justice." In *The Moral and Political Status of Children*, 186. Oxford University Press, 2002

liberty lost, to a one-to-many basic income argument along the same lines. Take the case where A uses human involuntary attention as a production factor and situates their attention-drawing object in a public place. This could take the form of a flashing billboard displaying an advertisement visible from a public place, such as a town square. In this case, A denies the liberty *of all people* to enter that public place without their attention being involuntarily drawn, and used as a production factor — so used as a means to A's productive ends, not as an end in themselves. In cases like this, A owes compensation to *everyone* for liberty lost, because *all people* are denied the liberty to enter this public place without their attention being involuntarily drawn. Compensation owed to everyone can be paid in the form of a basic income. A should pay into the basic income fund an amount proportional to the degree to which the attention-drawing object involuntarily draws attention, multiplied by the duration over which the attention-drawing object is situated in public. This sets the level of compensation paid depending upon the total amount of liberty denied by A.

5 Objections and Replies

A first concern with the above argument might be that we have denied A, as a self-owner, ownership of their own product — after all, they contributed labour, so they own the fruits of their labour. This worry is taken care of by the method of calculating the compensation owed by A. In the one-to-one case, A owes B the proportion of the product which is due to B's involuntary attention as a production factor. This is clearly not the entire product, because without A's labour, perhaps in designing and erecting the advertisement, the contribution of B's involuntary attention would have not produced anything. Therefore, A does not owe B the whole product. Instead, the amount owed is greater than *none* of the product (because B's attention is a production factor), and less than *all of* the product (because A's labour is essential). Where the level is set within this range depends upon the relative productivity of the two contributing factors (and others, such as natural resources) — it must be settled case-by-case where this balance lies. This reasoning should be used in carefully setting the level of compensation owed in the one-to-many basic income case, such that A retains the fruit of their labour, and the liberty denied of all people is fairly recompensed.

A potential objection to the basic income argument is that the liberty to direct one's attention has been shown, but it has not been shown that this liberty is one which extends to public spaces. A liberty to be naked might be argued for, yet this does not imply that such a liberty extends to public places. This objection usefully clarifies the importance of the liberty to direct one's own attention. As argued in 3, the ability to direct one's own attention is a prerequisite for other liberties such as freedom of speech. Freedom of speech certainly extends to public places, and so the denial of the liberty to voluntarily direct one's attention extends to public places too.

Throughout this essay, I have restricted claims about compensation for attentional liberty lost to cases where attention is drawn involuntarily as a production factor. I have thus far focused on attention used as a production factor as it is the most egregious case, because it breaks the Kantian ends-means injunction. However, the liberty to direct one's attention seems to be denied in cases where attention is drawn even *without* being a production factor. A worry arises that, if we expand the scope to all instances where attention is drawn involuntarily, there will be a massive proliferation of compensation owed. For example, if I plant some brightly coloured flowers in my garden, visible from the public street, these might involuntarily attract the attention of a passing pedestrian. That attention is not used as a production factor in anything, and I gain nothing. Yet I have still denied the pedestrian, momentarily, the ability to direct a portion of their attention. In response to this, it first should be noted that the vast majority of attention-drawing instances in the mass proliferation case will be extremely minor, such as this flowers example. Advertisements, on the other hand, are engineered to be maximally attention-grabbing. Thus the compensation owed will be a great deal lower in most attention-drawing instances where attention is not used as a production factor – so low, that we need not worry about fulfilling them. Additionally, due to the massive quantity of these generated, they will largely cancel out — I will owe you a miniscule amount, and you will owe me a miniscule amount, so in sum we owe each other virtually nothing.

I have responded to most of the unexpected obligations generated by mass proliferation, where small amounts of attention are drawn. But what of cases where large amounts of attention are drawn, yet we attract the person's attention for their own sake, satisfying the Kantian ends-means injunction? Most of these interactions take place in established relationships, because most people perform most of their interactions with people with whom they have an existing social connection. For example, my friend is visiting, and reading a book, and I bring them a delicious-smelling plate of curry. I have involuntarily drawn their attention, because the brain automatically directs attention towards newly perceived strong smells. Note that my friend may then voluntarily maintain their attention on the pleasant smell — yet, using the psychological definition of involuntary attention as "processes of attending that are not elicited by intentions but by certain outside events"[10], the initial attentional shift from the book to the curry smell is elicited by an outside event, the curry's arrival, so is classed as an involuntary drawing of attention. However, I argue that no compensation is owed, because by entering a friendship, an implicit contract is made to selectively forgo certain liberties with respect to one another. My friend has agreed to forgo the liberty of directing their own attention at all times when I am around.

What, then, of cases where the liberty to direct your own attention is denied, satisfying the ends-means injunction, but by someone with whom you have no pre-existing social agreement? For example, a stranger sees you stepping onto the road and shouts "watch out!" to warn you of an oncoming car. They have involuntarily drawn your attention, but for your own sake. In cases like this example, we might want to argue that you would have certainly consented to having your attention involuntarily drawn by the stranger, but there was no time for them to obtain your consent in advance.[11] However, allowing a person's liberties to be denied by others because we expect they would consent to such a denial generates potentially undesirable consequences. I suspect that the

[10] Eimer, Nattkemper, Schröger, Prinz, 155.

[11] Indeed, it might be difficult to obtain your consent without drawing your attention. This argument could arguably be extended by noting that the stranger is protecting your other liberties, such as the right to life, so denying the liberty to direct your own attention is worthwhile.

left libertarian would be suspicious of this line of reasoning, as it could be employed to justify paternalistic state action which breaches citizens' liberties. Further investigation is required here, going beyond the scope of this essay. Therefore, I restrict my argument for a tax on attentional liberty lost to cases where the attention is used as a production factor — the Kantian ends-means injunction is broken, and so there is strong justification for compensation being owed.

6 Conclusion

This essay has argued that, on a left libertarian account of distributive justice, we each have the liberty to direct our own attention. Some modern work, such as advertising, denies us this liberty by drawing our attention involuntarily, and uses this as a production factor. Therefore, those denied liberty in these cases are owed compensation – part of the product of the work should be paid to them. I then extended this argument to argue that such work, when situated in public, denies all people the liberty to enter that space without having their attention involuntarily drawn, and so compensation is owed to all – paid in the form of a universal basic income.

Bibliography

Broadbent, D. E. *Perception and Communication*. Pergamon, 1958.

Cohen, Gerald Allan. *Self-ownership, Freedom, and Equality*. Cambridge University Press, 1995.

Deutsch, J. Anthony, and Diana Deutsch. "Attention: Some theoretical considerations." *Psychological review* 70, no. 1 (1963): 80.

Eimer, Martin, Dieter Nattkemper, Erich Schröger, and Wolfgang Prinz. "Involuntary attention." In *Handbook of Perception and Action*, vol. 3, 155-184. Elsevier Academic Press, 1996.

Friston, Karl, James Kilner, and Lee Harrison. "A free energy principle for the brain." *Journal of Physiology-Paris* 100, no. 1-3 (2006): 70-87.

Jonides, John. "Voluntary versus automatic control over the mind's eye's movement." In *Attention and performance* 9, 187-203. Erlbaum, 1981.

Steiner, Hillel. "Silver spoons and golden genes: Talent differentials and distributive justice." In *The Moral and Political Status of Children*, 183-194. Oxford University Press, 2002.

———. "Compensation for liberty lost: Left libertarianism and unconditional basic income." *Juncture* 22, no. 4 (2016): 293-297.

Mole, Christopher, "Attention". *The Stanford Encyclopedia of Philosophy*, Fall 2017 Edition (2017).

Neisser, Ulric. *Cognitive Psychology: Classic Edition*. Psychology press, 2014.

Rosbergen, Edward, Rik Pieters, and Michel Wedel. "Visual attention to advertising: A segment-level analysis." *Journal of Consumer Research* 24, no. 3 (1997): 305-314.

Vallentyne, Peter, Hillel Steiner, and Michael Otsuka. "Why left-libertarianism is not incoherent, indeterminate, or irrelevant: A reply to Fried." *Philosophy & Public Affairs* 33, no. 2 (2005): 201-215.

‘Testimonial Throttling’ and Epistemic Injustice

Michael Calder

University of St Andrews

This essay provides a novel account of epistemic injustice by changing the standpoint of analysis from the marginalised to the oppressor. Previous investigations into epistemic injustice have shown how members of a marginalised group are harmed as knowers through their own speech. The framework that I will build — incorporating core elements of Fricker and Dotson’s work — focuses on speakers who truncate their own speech in conversation with a member of a marginalised social group, due to a bias against said audience. Testimonial throttling, at its core, is a restriction of access to the pool of knowledge due to bias. While a complete exposition of Fricker and Dotson’s work falls outwith the bounds of this essay, their accounts of ‘testimonial injustice’ and ‘testimonial quieting’, respectively, are instrumental in my account of ‘testimonial throttling’. After describing the foundations of this new account of epistemic injustice, I will propose a set of conditions along with thought experiment that describes a specific instance of testimonial throttling. Having defined testimonial throttling, I show that it covers a gap in the literature and provides insight into a vast array of resultant epistemic and practical harms. Before concluding I discuss a possible recourse for both the speaker and audience to combat testimonial throttling itself.

1 The Speaker: Why, and How

1.1 Why Investigate Testimonial Throttling?

1.1.1 The Road so Far: Fricker and Dotson

Fricker explains epistemic injustice by providing an account of testimonial injustice. This builds upon the concept of speaker credibility, which is the level of belief we have in the credibility of the person speaking. When a speaker has their credibility deflated they can in turn be “wronged specifically in her capacity as a knower”[1] — this is testimonial injustice. Specifically, the credibility deficit must be the result of bias, rather than due to the content of the speech, for it to constitute testimonial injustice.[2] Fricker uses Harper Lee’s ‘To Kill a Mockingbird’ to illustrate her definition. Tom Robinson, is unable to testify (both epistemically and legally) to the actions that resulted in his indictment — if he is to state that the white girl tried to kiss him, he will not be believed as a black man in a racist society.

[1] Fricker, Miranda, ‘Epistemic Injustice: Power and the Ethics of Knowing’, (2007), OUP, 20.

[2] Fricker, 22.

Fricker's other account of epistemic injustice, the more complex hermeneutical injustice, will help us clarify the nature of epistemic injustice. Fricker uses the example of women not having a term for sexual harassment as a paradigmatic case of hermeneutical injustice. Due to exclusion from the fields that investigate and construct concepts, and thus language, "such as journalism, politics, academia, and law",[3] women faced an obstacle that was not yet defined and to which they could not properly articulate an objection.

Fricker ties these two types of epistemic injustice to exemplify the nature of epistemic injustice clearly: "The wrongs involved in the two sorts of epistemic injustice, then, have a common epistemic significance running through them — *prejudicial exclusion from participation in the spread of knowledge*".[4] Epistemic injustice happens when people are excluded from the 'pooling of knowledge' (i.e. the knowledge we all share, contribute to, and draw from), due to bias. Testimonial injustice happens when this exclusion is based on "identity prejudice on the part of the hearer".[5] Hermeneutical injustice is when this exclusion is based on "structural identity prejudice in the collective hermeneutical resource".[6]

Dotson expands the discourse by defining 'epistemic violence'. She clarifies the importance of the audience on the "success of a speaker's attempt to communicate";[7] when the audience denies a speaker their full capacity to 'hear' they commit epistemic violence. This hinges on Dotson's definitions of 'reliable ignorance' and 'pernicious ignorance'. Pernicious ignorance is an extension of reliable ignorance ("ignorance that is *consistent* or follows from a predictable epistemic gap in cognitive resources"[8]) wherein the social context means that the reliable ignorance "causes or contributes to a harmful practice".[9]

To better understand what makes pernicious ignorance an epistemic violence Dotson provides an account of 'testimonial quieting', which occurs when "an audience fails to identify a speaker as a knower."[10] The broad nature of this concept is illustrated in Dotson's example: the epistemic position of black women in America. "Black women as belonging to an objectified social group, which hinders them from being perceived as knowers",[11] such consistent lack of recognition for black women's capacity as knowers exemplifies reliable ignorance. Moreover, given the obvious repercussions for such a belief, it is clearly a case of pernicious ignorance as well.

The other form that epistemic violence takes, as Dotson defines it, is through 'testimonial smothering', which occurs when a "speaker perceives one's immediate audience as unwilling or unable to gain the appropriate uptake of proffered testimony."[12] Dotson lists three issues at play which result in a speaker 'smothering' their own testi-

[3] Fricker, 152.
[4] Fricker, 152.
[5] Fricker, 152.
[6] Fricker, 152.
[7] Dotson, Kristie, 'Tracking Epistemic Violence, Tracking Practices of Silencing', *Hypatia*, no. 26 (2011), 238.
[8] Dotson, 238.
[9] Dotson, 239.
[10] Dotson, 242.
[11] Dotson, 243.
[12] Dotson, 244.

mony: (a) “the content of the testimony must be unsafe and risky”; (b) “the audience must demonstrate testimonial incompetence with respect to the content of the testimony to the speaker”; and (c) ”testimonial incompetence must follow from, or appear to follow from, pernicious ignorance”.[13] Dotson gives the example of discussion of domestic violence against black women “understood to corroborate stereotypes concerning the imagined “violent” black male”[14]. As epistemic violence occurs in “a failure of an audience to communicatively reciprocate, either intentionally or unintentionally”, the ‘violence’ of testimonial smothering is the audience’s failure to demonstrate testimonial competence due to their pernicious ignorance.

Having seen the development and implementation of the terms ‘epistemic violence’ and ‘epistemic injustice’ — they are epistemic as they relate to knowledge and are linked by the harms they cause — we now have our framework for building an expansion of the field. ‘Testimonial throttling’, as I will define it, develops from both of these epistemological investigations.

1.1.2 Why Must We Investigate Bias-induced Speaker Truncation?

(i) The lack of investigation into the common instance of biased speakers truncating their speech; (ii) The necessary importance of investigating those who control access to information.

(i) Dotson’s work on ‘testimonial smothering’ relies on the analysis of speech truncation in instances wherein the speaker perceives bias against themselves. In Fricker’s testimonial injustice, the injustice occurs when a speaker has deflated credibility due to bias. A gap exists between these two: when the speaker is biased against their audience and thus truncates their testimony. Such an occurrence is ubiquitous in societies rife with identity prejudice, that is to say all societies.

(ii) The oppressor has the greatest propensity to commit epistemic injustice; they are able to restrict access to knowledge as they, by definition, hold power. Moreover, there is a motive for the oppressor to suppress access to knowledge in order to continue their literal and epistemic subjugation of the marginalised groups they oppress. This comes to fruition as a ‘self-fulfilling prophecy’ which will be discussed at length in Section 2.2.

1.1.3 Objections to the Investigation

Feminist epistemology takes the standpoint of the oppressed. While it may seem that analysing the speech of the powerful (as testimonial throttling sets out to) goes against this shared methodological aim, this is not the case. While it is true that “feminist standpoint theorists have been explaining the importance of starting our thinking or our research from the lives of marginalized people”[15], investigating the powerful can shed much light

[13] Dotson, 244.

[14] Dotson, 245.

[15] Garry, Ann, ‘Intersectionality, Metaphors, and the Multiplicity of Gender’, *Hypatia*, 26, no. 4 (2011), 828.

on the epistemic injustices that marginalised people face.

Throughout this paper I will refer to the 'marginalised' or 'oppressed'; and the 'oppressor'. I understand that these terms have their definitions debated to great extent in the literature; later work on testimonial throttling could include reference to specific theories of oppression but I feel it falls outside the bounds of this work. Thus, I use them and intend them to be understood in simple, relational terms.

1.1.4 Throttling's Advantage and the Term Itself

While it is primarily a complementary theory: there are inherent advantages to testimonial throttling as a mode of analysis over Fricker and Dotson's work. While the harms resulting from testimonial throttling serve to best elucidate its importance, I posit that methodologically, throttling is easier to diagnose.

Throttling is easier to spot in social contexts, both historically and in new instances. As an outside observer, in order to diagnose an instance of testimonial throttling, we merely have to see a difference in speech due to bias on the part of the speaker. In the case of testimonial smothering, a further demand is that we must conclude that the speaker is changing their speech due to (a) *her perception* of (b) *the audience's potential for bias* against her group. This calculation on the part of the speaker necessitates a subjective analysis that is not present in the cut and dry case of testimonial throttling. Given our goal should be to end epistemic injustices, the simplicity of testimonial throttling makes it easier to propose solutions to it.

I would like to make a note on the term 'throttling'. While it evokes violence, I find this necessary, much as Dotson uses smothering, given the injustice at play. I intend its use not only as an homage to Dotson but as a double meaning - with 'throttle' understood to denote:

control, the throttle of a car, controlling power output just as testimonial throttling controls knowledge output; and ***abuse*** with throttling as a physical act of violence that controls the subject, just as limiting access to the pool of knowledge abuses the marginalised.

1.2 Defining Throttling

1.2.1 The Conditions

Testimonial throttling occurs when:

(1) Speaker 'P' truncates speech to an audience 'Q'.

  (1.a) Q must be the marginalised member of the conversation in which the truncation takes place.

    (1.a.i) Q's status as 'marginalised' is relational to P, it can be but need not be that they are structurally oppressed.

    (1.a.ii) P has knowledge that Q lacks, while they may not necessarily oppress the marginalised elsewhere, situationally they hold power over the knowledge.

(2) The truncation reduces Q's access to the pool of knowledge.

  (2.a) This knowledge would benefit Q and is relevant to the conversation.

(3) P's reason for the truncation is bias against Q.

  (3.a) Specifically, bias results from *identity prejudice.*

    (3.a.i) Thus, Q's status as a knower is devalued.

(4) Q is harmed as a knower due to the conjunction of (1), (2), & (3).

Specifically:

- While it is likely that the marginalised audience is structurally oppressed given the nature of such knowledge exchanges, it should not be a necessary condition for testimonial throttling. This allows for cases wherein 'outliers' of oppressed groups find themselves in positions of power over knowledge but continue to act as an oppressive force to others, even to their own communities.[16] However, the central case of testimonial throttling should be understood as that which represents the structural disadvantage of the access to knowledge of the marginalised.
- In 3.a.i. Q's status as a knower being devalued mirrors Fricker's *speaker credibility* in cases of testimonial injustice. Just like Tom Robinson's credibility was deflated and so he wasn't believed a competent witness, Q is considered an incompetent audience.

1.3 Testimonial Throttling in Action

1.3.1 A Thought Experiment

This case not only shows a common and clearly diagnosable example of testimonial throttling but will also serve as a solid basis for analysing the epistemic and practical harms that arise from instances of testimonial throttling.

[16] A black republican senator, for example, is a member of a marginalised group still capable of testimonial throttling against minority communities in his capacity as a senator.

We are to imagine that a white man goes to a job centre, where he is given ample resources on emerging industries he may be suited to and instruction on the best way to prepare for a job interview. He follows the advice and secures a good job in a relevant industry.

A black woman with the same qualifications and background goes to the same job centre and meets with the same person. However, due to their preconceived bias that black women are 'lazy' and 'stupid', the job centre employee believes that the black woman would not benefit from the information he is able to provide her. Thus, he doesn't bother to explain the current industrial dynamics, nor does he give any advice on securing such a job, merely handing the black woman some readily-available pamphlets and giving non-specific advice. The black woman leaves the job centre with little new or relevant information.

1.3.2 Applying our Conditions

We set out conditions for testimonial throttling in section 1.2.1. I will now use the thought experiment to elucidate these conditions.

In this case, the employee takes the place of 'P', and the black woman 'Q'.

(1) P has clearly truncated their speech directed at an audience Q, the speech granted to the white man before Q shows the extent of this truncation.

> **(1.a)** Q is marginalised and relies on P for access to the pool of knowledge, due to her position as job seeker.
>
> > **(1.a.i)** In this case Q is also a member of a structurally oppressed group, black women, who have been systematically discriminated against in the job market.
> >
> > **(1.a.ii)** P has knowledge due to their position as a job centre employee that Q lacks.

(2) The truncation reduces Q's access to the pool of knowledge that would allow her to advance her job seeking ambitions.

> **(2.a)** The truncated part of the speech contains knowledge highly important and relevant to the job seeking, i.e. the reason for the conversation.

(3) P's reason for the truncation is bias against Q.

> **(3.a)** Specifically, bias results from P's racial bias: *identity prejudice.*
>
> > **(3.a.i)** Thus, Q has her status as a knower devalued by P not due to any substantive reason

(4) Q is harmed as a knower due to the conjunction of (1), (2), & (3).

1.3.3 Truncations of Speech Without Testimonial Throttling

It is important for a precise account of testimonial throttling to consider cases that appear to be truncations of speech, perhaps even due to bias, that do not meet the entire criteria. Tight criteria and acknowledgement of possible problem cases should bulwark testimonial throttling against possible objections.

Case 1 - When truncating testimony is not an epistemic harm & not due to bias.

In this case someone is truncating speech for epistemically beneficial reasons. A public lecture on a scientific discovery that is attempting to inform non-specialists will use different language to a professor informing his students of the same subject. The speech truncation in this case is not done either to cause harm, nor is it epistemically harmful - it is done to aid understanding of a topic that can be easily understood to be outside the grasp of the general public. In this case, simplification of technical terminology aids access to the pool of knowledge.

Case 2 - When truncating testimony is hurtful & due to bias but not an epistemic harm.

In this case we see a truncation of speech, due to bias, that may cause emotional harm to the audience but cannot qualify as an epistemic harm. For example, an employer in the service industry may truncate their speech when speaking to an employee they presume unable to understand 'proper english'. In saying something like "Clean... toilet... now" rather than "once you're done with your current task please can you go on to clean the toilets", the employer might hurt their audience's feelings and are acting on their bias but clearly there is no epistemic injustice at play. While speech has been truncated, the part that has been removed is not contributory to the pool of knowledge and while the worker may feel demeaned they have not been harmed in their capacity as a knower.

2 What's Happening to the Audience: Epistemic and Practical Harms

2.1 Epistemic Ramifications of Throttling

2.1.1 How Epistemic Injustices are 'Epistemic:'

Pohlhaus Jr. provides a useful set of criteria that describes the epistemological nature of the injustices present in the literature. I will first describe the criteria as they relate to Fricker and Dotson's work and then how each similarly applies to 'testimonial throttling'.

Pohlhaus Jr. helpfully uses both Fricker and Dotson to relay the main epistemological significance of epistemic injustice: that the injustice is being done to "knowers *as* knowers."[17] She describes how in Dotson's case it is a knower's testimony that is suppressed (testimony is uncontroversially understood to be of great epistemic significance) and Fricker's that it is made "difficult for particular knowers to know what it is in their interest to know".[18] Similarly, 'testimonial throttling' is the action of suppressing one's access to the pool of knowledge, thus it wrongs the audience specifically in their capacity as knowers.

Furthermore, Pohlhaus Jr. describes these injustices as epistemic due to their inducing epistemic dysfunction,[19] that which (negatively) affects one's status or ability as an epistemic agent. The resultant harms of epistemic dysfunction, "distorting understanding and stymieing inquiry"[20] , are clearly present in cases of testimonial throttling. The audience's understanding is distorted as they are unaware of the knowledge they lack. Similarly, their ability to inquire is reduced by the lack of access to the full pool of knowledge. Thus, 'testimonial throttling' meets Pohlhaus Jr.'s second criterion as an action that causes 'epistemic dysfunction'.

The third criterion is an extension of the first two in that it states such harms arise either "within [... or ...] through the use of, our epistemic practices and institutions".[21] Epistemic practices are all those which engage with the social pool of knowledge; Pohlhaus Jr. notes that 'school curricula' is an epistemic institution. If curricula are a locus of epistemic injustice then the application of a curriculum can surely produce instances of testimonial throttling. Given testimonial throttling's focus on speech truncation that affects access to the pool of knowledge it aids our evaluation of the harms "an epistemic institution causes in its capacity as an epistemic institution."[22]

2.1.2 Epistemic Objectification

There is a debate in the literature as to whether the primary harm that occurs in cases of testimonial injustice is an instance of 'epistemic objectification' or 'epistemic othering'. An understanding of this debate will help frame 'testimonial throttling' in the literature whilst serving to better understand the exact nature of the resultant epistemic harms.

Fricker posits that the harm *objectifies*, as the epistemic agent is treated as an object or "a mere source of information rather than as an informant due to one's prejudices".[23] This is based upon a criteria of objectification, proposed by Nussbaum, that identifies seven ways in which one may be objectified.[24] Fricker specifically focuses on Nussbaum's third criterion 'inertness', that states "the objectifier treats the object as lacking in agency, and perhaps also in activity".Nussbaum, 257

[17] Pohlhaus Jr, Gaile, 'Varieties of epistemic injustice', in: *The Routledge Handbook of Epistemic Injustice'* (2017), 213.

[18] Pohlhaus Jr, 213.

[19] Pohlhaus Jr, 213.

[20] Pohlhaus Jr, 213.

[21] Pohlhaus Jr, 213.

[22] Pohlhaus Jr, 214.

[23] McGlynn, Ann, 'Objects or Others? Epistemic Agency and the Primary Harm of Testimonial Injustice', *Ethical Theory and Moral Practice* 23 (2020), 832.

[24] "Instrumentality, denial of autonomy, inertness, fungibility, violability, ownership, and the denial of subjectivity". from: Nussbaum, Martha, 'Objectification', *Philosophy and Public Affairs* 24, no. 4 (1995), 257.

McGlynn notes that three problem cases are typically used to motivate the claim that the harm should instead be considered an instance of epistemic othering:

(i) When there is a failure to consider the speaker as an inquirer despite treating them as an informant rather than a mere source of information.[25]

(ii) When credibility excess leads to a testimonial injustice.[26]

(iii) When one lies in spite of being considered capable of telling the truth.[27]

I adopt McGlynn's approach to these problem cases. In order to solve these problem cases and account for testimonial throttling within the framework of epistemic objectification we need only turn to the rest of Nussbaum's criteria. McGlynn refers to the specific implementation of each of Nussbaum's criteria in the epistemic objectification account of harm as 'epistemic analogues'.[28]

McGlynn argues that (ii) can be solved by considering Nussbaum's "Fungibility" criterion, wherein "the objectifier treats the object as interchangeable with other objects of the same type, or with objects of other types"[29]. Similarly, (i) can be overcome when one considers it an example of Nussbaum's second criterion: a 'denial of autonomy' in which "the objectifier treats the object as lacking in autonomy and self determination"[30]. Finally, (iii) is actually not a problem case. Fricker deals with this in her example of Tom Robinson's testimony as he is believed to have been insincere: she considers "both competence and sincerity as epistemic [...] she takes the capacity to convey one's knowledge to others as essential to the very possession of knowledge".[31]

Having asserted the applicability of the objectifying account of epistemic harm in Fricker's 'testimonial injustice', its use for describing the harms inherent to testimonial throttling is clear. When a speaker truncates their speech due to bias they are, in fact, epistemically objectifying their audience in accordance with four of Nussbaum's criteria for objectification, specifically through the analogues of: (a) "inertness: the objectifier treats the object as lacking in agency, and perhaps also in activity."[32]; (b) "denial of subjectivity: the objectifier treats the object as something whose experiences and feelings (if any) need not be taken into account."[33]; as well as the aforementioned (c) 'denial of autonomy' and (d) 'fungibility'.

2.1.3 Applicability of Objectification to Throttling

To illustrate my case I will show how each of these are observable in cases of testimonial throttling. Using the thought experiment set out in Section 1.3.1, we can see the applica-

[25] McGlynn, 834.
[26] McGlynn (2020), 834
[27] McGlynn, 835.
[28] McGlynn, 842.
[29] McGlynn, 833.
[30] McGlynn, 833.
[31] Hawley, Katherine, 'Trust, Distrust, and Epistemic Injustice' in: *The Routledge Handbook of Epistemic Injustice* (2017), 72.
[32] Nussbaum, 257.
[33] Nussbaum, 257.

bility of each mode of objectification:

(a) The classic means of objectification in Fricker's cases is treating the agent as inert. In our case we see the audience being treated as inert through the assumptions made on the part of the speaker. By allowing their bias to influence their speech truncation they take no measures to engage with the audience as an epistemic agent and thus are treating them as though they lack agency. They see a 'black woman' and their minds are made up (due to their bias) as to their audience's capacity.

(b) The audience are having their subjectivity denied as they are judged and treated not due to their individual experiences or feelings but instead solely as a representative of their group; conversely, the white man is treated as a unique individual whose feelings and experiences are taken into account.

(c) The speaker is clearly denying the audience their autonomy, by restricting their access to knowledge they are denying their right to self-determination. The restricted knowledge, on account of their membership of a marginalised group, restricts their ability to flourish. The audience, having been defined as 'a black woman' and having been tarred through bias as being 'undeserving' of access to the full pool of knowledge, is seen as not befitting the full range of opportunity to develop themselves that would be granted to 'a white man'.

(d) The audience being treated as 'fungible' can be seen as an extension of (b). While the denial of subjectivity means that the audience is being seen solely as 'a black woman' rather than a complex individual, they are being demeaned by the fact that the speaker is considering them to be interchangeable with any other member of the group 'black women'. The white man is considered an individual and thus given specific, relevant information; the black woman is given the small amount of information that the employee would give any black woman.

Thus, we have framed the primary epistemic harms of 'testimonial throttling', with reference to several leading authors in the field of feminist epistemology. Moreover, we have used the case given in Section 1 to exemplify these harms.

2.2 Practical Harms Arising from Epistemic Harms

Having described the epistemic harms that arise from instances of testimonial throttling I will move on to the practical implications of these harms.

2.2.1 Individual and Group Harms

Anderson claims that Fricker's 'testimonial injustice' fails to remedy the structural nature of epistemic injustice as "her remedies in both cases [individual and structural injustices] stress individual virtue".[34] Anderson claims Fricker's "depict[ion] as a transactional injustice",[35] cannot account for structural epistemic injustice. However, I contest that Anderson is merely not considering enough instances of transactional injustices. Adding more accounts of transactional injustices (testimonial throttling being merely one), will show that transactional injustices can account for structural injustices.

Anderson proposes a case:

> there is no transactional injustice in refusing to offer a job to an unqualified applicant, the fact that members of a disadvantaged group cannot get good jobs because they have been unjustly denied opportunities to qualify themselves for these jobs justifies the judgment that their lack of access to good jobs is a structural injustice.[36]

Anderson is correct that testimonial injustice alone is inadequate to account for the asymmetry of access to the pool of knowledge; it is for this reason I consider throttling a significant expansion of the literature. The case she gives is a paradigmatic example of the structural repercussions of an instance of testimonial throttling — as shown in my thought experiment. Testimonial throttling is one of many 'transactional' instances that can lead to structural injustices. As the literature expands, we will have accounts of many more varieties of transactional instances of epistemic injustice that will further our understanding of the resultant structural injustices. Deconstructing these structural injustices from their point of conception (instances of transactional injustice) will provide us with the best tools for combatting injustice.

However, Anderson is correct in noting that a full treatise on the remedies for structural epistemic injustices "would require many books".[37] Importantly, this analysis has shown the intrinsic link between individual and group harms: each instance of transactional injustice against individuals contributes to the structural injustices the group (of which the individuals are members) must face.

2.2.2 Personal Harms

It takes little imagination to see how the repercussions of testimonial throttling can be severe and widespread. I see these practical harms as ways in which the audience are harmed as persons — in contrast to them being harmed as knowers. The first harm I will discuss is both an epistemic and practical harm, it is perhaps the most significant of all the harms that an account of testimonial throttling divulges: the self-fulfilling prophecy.

The self-fulfilling prophecy can be best understood practically, through our original thought experiment. The aftermath of our thought experiment can conceivably go

[34] Anderson, Elizabeth, 'Epistemic justice as a Virtue of Social Institutions', *A Journal of Knowledge, Culture and Policy* 26, no. 2 (2012), 165.

[35] Anderson, E., 165.

[36] Anderson, E., 169.

[37] Anderson, E., 171,

as follows. The white man's ability to get a good job will be greatly improved via his easy access to the pool of knowledge. The black woman will struggle to get a similar job due to her lack of access to the same pool of knowledge. Her lack of access to the pool of knowledge will result in her genuinely having less knowledge, thus 'confirming' one of the original stereotypes that resulted in speech truncation: that black women are stupid. Furthermore, her lack of knowledge in this area will result in her finding it more difficult to get a job. Her resultant unemployment will 'confirm' to her prejudiced observers that she is lazy, the other stereotype for which she was excluded from access to the pool of knowledge. In summary: the result that the white man gets the job and the black woman is forced to go onto unemployment benefits, reaffirms both the employee's racist beliefs and the white man's belief of superiority. I see this as a reflection of the consequences Fricker discussed, in which those who commit testimonial injustices end up limiting their own access to knowledge:

> So the preservation of ignorance that p, where p is the propositional content of what was said, may often entail further missed epistemic opportunity [...] We might express this by saying that testimonial injustice tends to preserve not only immediate ignorance but also inferentially ramified ignorance.[38]

However, in our case those that commit the injustice limit their *victims'* access to knowledge.

The self-fulfilling prophecy is thus, the confirmation of the original biases - that led to testimonial throttling — by the effect testimonial throttling has on its audience. This in turn can lead to further testimonial throttling. For instance, in our case, the black woman going onto unemployment benefits will increase the percentage of black unemployment, leading the job centre employee to reaffirm his belief that black women are lazy and make him even less likely to provide black women with adequate resources in the future. The self-fulfilling prophecy causes harm to the audience over and over again, as knowers and as persons.

The self-fulfilling prophecy is not only a symptom of oppressive society but one of the causes, an enabler of what Mills calls the 'domination contract', that in contrast to social contract theory, "society is basically coercive, with injustices and social oppression being the norm".[40] Without properly assessing instances of testimonial throttling we permit the coercive nature of society that states "United States has historically been a racially flawed liberal democracy"[41] rather than the United States has historically been a white supremacist polity."[42] Testimonial throttling addresses systemic practices that create and reinforce oppressive dichotomies, particularly racist theories of supremacy, which harm the audience as persons.

Beyond the self-fulfilling prophecy there are numerous practical harms that result from instances of testimonial throttling. The black woman is clearly harmed as a

[38] Fricker M., & Jenkins, K., 'Epistemic Injustice, Ignorance, and Trans Experience', in: [39] (2017), Routledge, 270.

[40] Mills, Charles W., 'Philosophy and the Racial Contract', in: *The Oxford Handbook of Philosophy and Race* (2017), Oxford: Oxford University Press, 6.

[41] Mills, 7.

[42] Mills, 7.

person by her lack of ability to get a job: financially as well as her ability to achieve self-determination. The possibilities for similar harms are endless. The black student who was not adequately prepared for his university applications due to his teachers' bias misses out not only on the financial opportunity that a degree permits but also to become better educated. The minority defendant who through the bias of lawyers and the court isn't fully explained the consequences of a plea bargain and relinquishes her freedom.[43] Testimonial throttling clearly accounts for harms that negatively affect the daily lives of those on the receiving end. For this reason, testimonial throttling is a social justice issue as much as it is an epistemic concern.

2.3 Alleviating a Meta-Harm & Ethics of this Investigation

This paper has addressed a meta concern. We have provided those who have been subject to testimonial throttling a concept which makes their experience intelligible. Before this investigation, those on the receiving end of testimonial throttling were unable to make communicatively intelligible something which was particularly in their interests to be able to render intelligible. This has the effect of ending one hermeneutical injustice. To not discuss these concerns would be to engage in "the active production and preservation of ignorance by those in privileged positions".[44]

Pohlhaus Jr. warns, when writing on epistemic injustices: "we would do well to consider [...] the ways in which this essay might itself participate in and perpetuate epistemic injustice".[45] Thus, I have tried to ensure that this essay does not engage in throttling nor any other act of epistemic injustice. From my perspective, a privileged one where I am not a member of a marginalised community, I have attempted to spell out my case clearly so as to not create an impasse for any potential readers.

2.4 What Can be Done About This, by the Speaker and the Audience.

Remedies for epistemic injustices are difficult due to the wide-reaching, structural and individual biases that form their cause. This puts philosophers in the position that their proposed solutions can be deemed 'grandiose' or 'naïve'. However, instances of testimonial throttling can be reduced on the part of the speaker, and the audience, being aware of throttling, can prepare to fight against it.

On the speaker side, Sullivan points to a number of ways in which we can begin to stamp out epistemic injustices in the criminal justice system. His most relevant point is, to "increase efforts to make judges and juries more aware of the assumptions they bring to

[43] Cox, Jane, & Sacks-Jones, Katharine, 'Double Disadvantage: The Experiences of Black, Asian and Minority Ethnic Women in the Criminal Justice System', *Agenda, the alliance for women and girls at risk'*: Report (April 2017), https://weareagenda.org/wp-content/uploads/2017/03/Double-disadvantage-FINAL.pdf, 8.

[44] Tuana, Nancy, 'Feminist epistemology: the subject of knowledge', in: *The Routledge Handbook of Epistemic Injustice* (2017), 132.

[45] Pohlhaus Jr, 14.

their interpretation of the meaning of our social practices."[46] The most important barrier to stopping testimonial throttling is ignorance of the act. If, as we have shown, implicit bias can cause testimonial throttling, then society, and individuals, must work hard to acknowledge their own biases. Sullivan notes that: "The retraining and wide-ranging dialogue needed for an assessment of the objectivity of our beliefs would require a significant investment of time and energy".[47] While I do not doubt the difficulty of this undertaking, in the past year alone we have seen the meteoric rise of Black Lives Matter, a movement whose purpose is to raise awareness of structural and individual racism in society. As racial justice movements grow, those who control access to knowledge will inevitably be forced to consider their own role in a system that perpetuates injustice, epistemic or otherwise. Thus, the clearest way of decreasing the instances of testimonial throttling is through increasing awareness for those who perpetrate it. In Dotson's terminology this would be education to weed out pernicious ignorance, in Fricker's this would be ensuring that credibility is not unjustly deflated. As Sullivan puts it "active ignorance of our own ignorance" is no excuse.[48]

As for the audience, hoping that awareness of testimonial throttling will enable those on the receiving end to prepare for it is only one small step - becoming aware that sources of knowledge can be biased, motivates searching for knowledge elsewhere. To contest testimonial throttling, to call it out when it is apparent, is the clearest (if tremendously difficult) means of raising awareness. To ask that of the marginalised, could be asking too much, but as we have seen with the rise of racial justice movements this century - it can be done.

3 Conclusion

I have shown that testimonial throttling merits a place in the literature given its wide reaching implications and ubiquity. Ideally, the discernment of cases of testimonial throttling should be made a priority for those wishing to fight for not only epistemic justice but social justice. Eliminating instances of testimonial throttling will increase access to the pool of knowledge, in turn preventing the 'self-fulfilling prophecy' and other resultant harms. While these aims are no mean feat, such idealism is not without merit. Epistemology by its very nature can be a force for change: as knowledge builds so does power.

Bibliography

Anderson, Elizabeth, 'Epistemic justice as a Virtue of Social Institutions.' *A Journal of Knowledge, Culture and Policy* 26, no. 2 (2012): 165.

Anderson, Luvell, 'Epistemic Injustice and the Philosophy of Race' in: *The Rout-*

[46] Sullivan, Michael, 'Epistemic Justice and the law', in: *The Routledge Handbook of Epistemic Injustice* (2017), 300.

[47] Sullivan, 301.

[48] Sullivan, 301.

ledge Handbook of Epistemic Injustice (2017): 144.

Cox, Jane, & Sacks-Jones, Katharine, 'Double Disadvantage: The experiences of Black, Asian and Minority Ethnic women in the Criminal Justice System' *Agenda, the alliance for women and girls at risk:* Report (April 2017)

Dotson, Kristie, 'Tracking Epistemic Violence, Tracking Practices of Silencing', (2011). *Hypatia* 26 (2011): 236-257.

Fricker, Miranda, *Epistemic Injustice: Power and the Ethics of Knowing* Oxford University Press, 2007.

———, & Jenkins, Katharine, 'Epistemic Injustice, Ignorance, and Trans Experience', in: *Routledge Companion to Feminist Philosophy*. Routledge, 2017: 268-278.

Garry, Ann, 'Intersectionality, Metaphors, and the Multiplicity of Gender', *Hypatia* 26, no. 4 (2011): 826-850.

Hawley, Katherine, 'Trust, distrust, and epistemic injustice', in: textitThe Routledge Handbook of Epistemic Injustice (2017): 69-78.

McGlynn, Aidan, 'Objects or Others? Epistemic Agency and the Primary Harm of Testimonial Injustice', Ethical Theory and Moral Practice 23 (2020): 831-845.

Mills, Charles W., 'Philosophy and the Racial Contract', in: *The Oxford Handbook of Philosophy and Race,* Oxford University Press (2017).

Nussbaum, Martha, 'Objectification', *Philosophy and Public Affairs* 24, no. 4 (1995). 249-291.

Origgi Gloria, & Ciranna, Serena, 'Epistemic Injustice: the case of digital environments', in: *The Routledge Handbook of Epistemic Injustice* (2017): 303.

Pohlhaus, Gaile, 'Varieties of epistemic injustice', in: *The Routledge Handbook of Epistemic Injustice* (2017): 13-26.

Sullivan, Michael, ''Epistemic Justice and the law', in: *The Routledge Handbook of Epistemic Injustice* (2017): 293-302.

Tuana, Nancy, 'Feminist epistemology: the subject of knowledge' in *The Routledge Handbook of Epistemic Injustice* (2017): 125-138.

Metalinguistic Negotiation and its Limits

William Chao

Yale University

This paper defends the idea that disputes which do not feature conflicts in literally-expressed contents could express genuine disagreement. Using the model of metalinguistic negotiation and Stalnakerian common ground, the paper argues that many such disputes are driven by the conversational parties' disagreements in the meaning of expressions. The disputants convey and settle their disagreement pragmatically, negotiating the meanings of terms under controversy by using instead of mentioning the terms. The paper further examines how the disputants collect cues from the conversation to become aware of the metalinguistic nature of their dispute and explains why such an account is compatible with semantic externalism, by clarifying the scope and limits of metalinguistic negotiation.

1 Introduction

The paper seeks to defend the possibility that dispute between speakers could express genuine disagreement, even if the speakers mean different things by the same term they both use and literally express rationally compatible content. This is against an accepted way of reasoning, which insists that the speakers must mean the same thing by their common term in order to genuinely disagree with each other. To explain such a possibility, the paper starts by introducing the metalinguistic negotiation model, according to which conversational participants negotiate the proper meaning to use for certain common terms by using instead of mentioning the very term. The paper further clarifies this model by reviewing it in light of the Stalnakerian common ground. It sees the efforts of conversational participants to decide which meaning to use for the common term as a negotiation regarding what to include within their interpretative common ground, the common information body about language presupposed by conversational participants for effective communication to happen.

The paper then proceeds to respond to two challenges against metalinguistic negotiation model raised by Herman Cappelen, a speaker-error objection that speakers engaging with these disputes do not see themselves as debating the meaning of words, and the externalist criticism that the model assumes a control of speakers over meaning that does not exist, which makes it incompatible with mainstream semantic externalism. Regarding the speaker-error objection, the paper points out that speakers approach these disputes with the default assumption that there is no meaning difference of the terms they use, and it oftentimes only becomes clear for the speakers when they collect cues from their conversational proceedings that demonstrate their divergence in language use. For the externalist argument, the paper clarifies that metalinguistic negotiation can only directly act on what people think, believe, presuppose, and their other attitudes toward words' meaning and reference, instead of the actual meaning of terms. The paper then takes a step further, by exploring the possibility of metalinguistic negotiation to indirectly contribute to the facts grounding the meaning of terms, by influencing speakers' attitudes toward the meaning of terms and changing the use patterns.

2 Metalinguistic Negotiation Model, Explained

2.1 Mark Sainsbury's "fishy" case

In his short paper *Fishy Business*, Mark Sainsbury tried to make sense of a 19th century courtroom debate between an inspector and an oil merchant on whether whale oil should be taxed as fish oil. Both parties recognized that whales are lung-using, air-breathing mammalian sea creatures but disagreed on the relevance of these facts. The inspector used "fish" to designate all sea creatures while the merchant emphasized that mammals cannot be fish.[1] Do the speakers' different ways of classification correspond to different meanings of the expression "fish"? The oil merchant and the inspector applied radically different methodologies of classification and, as a result, used the term "fish" in systematically different ways. The oil merchant applied the term in such a way as to never include mammals, while the inspector applied the term in such a way as to include marine mammals. The difference holds true even when all the relevant factual information about whales is on hand, making it difficult to pinpoint the factual basis of their different classifications, unless they mean different things by the term "fish." This, at the very least, provides prima facie reasons for thinking that the merchant and the inspector meant different things for the term "fish."

Call the above argument *different-meaning interpretation* of the "fishy" case. It quickly faces a challenge that it fails to capture the observation, which we would like to preserve, that the dispute between the merchant and the inspector expressed genuine disagreement.[2] The accepted view of disagreement tells us that it requires conflict in content literally expressed: if a dispute between speaker A and B expresses a genuine disagreement, in which speaker A asserts that p and speaker B asserts that q, it must be rationally incompatible for anyone to accept both p and q. It suggests that for any dispute to express genuine disagreements, the speakers cannot talk past each other. According to this view, if any dispute involves assertions of the form that the *F*s are *G*s and that the *F*s are not *G*s, the speakers must mean the same thing by the terms "F" and "G" to assert something rationally incompatible and avoid talking past each. This view of disagreement, which requires conflict in content literally expressed, helps draw a semantic conclusion about "F" and "G" from our intuition of disagreement, and challenges the different-meaning

[1] Sainsbury, M., "Fishy Business," *Analysis* 74, no. 1 (December 2013): 3.

[2] The observation that there is a genuine disagreement is worth defending for many reasons. The dispute would not have stopped had they agreed that whales are lung-using mammalian sea-creatures. There is also good evidence suggesting that the dispute heavily centered on whether whales are fishes instead of other possible disagreements, since the trial largely proceeded with the jury hearing evidence from eminent anatomists (saying that whales are not fish), merchants and seafaring men (mostly, but not in every case, saying that whales are fish).

interpretation of the “fishy” case, because there is no conflict in the content literally expressed.[3] If the merchant and the inspector meant different things by the term “fish”, the merchant asserted, in paraphrases, that whales are cold-blooded vertebrate animal living wholly in water, while the inspector asserted that whales are sea creatures. The speakers asserted compatible propositions, and did so by virtue of the fact that they intended to mean different things by the same term “fish”. It is not rationally incompatible for either of them to accept both of the two propositions. The different-meaning interpretation, therefore, seems unable to validate our observation that the “fishy” case involves genuine disagreement.

2.2 Non-canonical Disputes and Metalinguistic Negotiation

The above challenge to the different meaning interpretation is misplaced because it mistakenly presumes that only *canonical disputes* — disputes that involve literal expression of incompatible content — could express genuine disagreement. It is wrong to assume, merely based on the recognition of genuine disagreement, that the relevant dispute should involve the literal (semantic) expression of incompatible contents. *Non-canonical disputes* — disputes centered on information that is not conveyed semantically — could express genuine disagreements as well.[4] For this paper, I am using the term *dispute* to refer to any linguistic practice that appears to evince or express a genuine disagreement. And I am using the term *disagreement* to involve conflicting attitudes of speakers toward certain content, such as acceptance, beliefs and desires, rather than what they utter. For two speakers to disagree with each other there should be some objects p and q (e.g. propositions, plans, etc.) such that A accepts p and B accepts q, when it is rationally incompatible for anyone to accept both p and q.[5] However, the content in conflict may well be conveyed pragmatically without being literally expressed.

While the distinction between semantics and pragmatics is hard to draw, the paper will make the distinction between the sentence meaning, the linguistic meaning of a sentence-type, and speakers’ meaning, which consists of what is said and what is implicated. This is the maximalist interpretation of this distinction, which emphasizes the difference between “primary” contextual processes and “secondary” contextual processes.[6] Primary pragmatic processes are contextual processes that help determine what is said, the proposition expressed by the sentence, while secondary pragmatic processes are inferential processes that presuppose what is said, take it as input, and yield further propositions as output.[7] For this paper, if a speaker conveys that p *semantically*/literally, the proposition that p belongs to what is said, while what is conveyed *pragmatically* belongs to what is implicated.

With the above concepts clarified, consider the following case of disagreement centered on information conveyed via implicature of relevance.

[3] Plunkett, D. and Timothy Sundell., “Disagreement and the Semantics of Normative and Evaluative Terms”, *Philosophers’ Imprint*, 13, (2013): 3.

[4] Plunket, Sundell, 7.

[5] Plunkey, Sundell, 11.

[6] Recanati, F., “What Is Said”, *Synthese*, 128 (2001): 79.

[7] Recanati, 79.

(1a) Sally was able to solve the last problem on the test.

(1b) But Sally chose to quit before she got there.

The literal contents of (1a) and (1b) are rationally compatible, but there is an incompatibility between (1b) and the implicature of (1a), namely that Sally in fact solved the last problem on the test. The genuine disagreement expressed by the dispute centered on this incompatibility.[8] The example demonstrates that the possibility of non-canonical disputes to express genuine disagreement. Therefore, one could argue that in the "fishy" case the genuine disagreement between the merchant and the inspector did not center on the propositions they asserted but on what the speakers pragmatically conveyed in addition. The merchant and the inspector, when they made the assertions, each pragmatically advocated for the meaning of "fish" he deemed appropriate. The two speakers pragmatically proposed two meanings of "fish" incompatible for one to accept at the same time. The conflict in content, in this way, lies within what is pragmatically conveyed within the conversation instead of its literally-expressed semantic content.

The different-meaning interpretation of the "fishy" case, which locates the conflict in what the speakers convey pragmatically, helps shed light on a series of similar questions across various fields of philosophy. When progressives and conservatives debate whether gay marriage should be legal, do they mean the same thing by "marriage"? When a sports radio host and his audience argue whether a racehorse could be an athlete, do they mean the same thing by "athlete"?[9] When missionaries and cannibals debate whether collecting human scalps is good, do they mean the same thing by "good"?[10] If one believes that in these disputes the speakers mean different things by their common terms, we could analyze such non-canonical disputes as *metalinguistic negotiations*, disputes where participants use (as opposed to mention) expressions whose meanings they disagree about in order to resolve this very disagreement.[11]

2.3 Why Metalinguistic Negotiations are Worth Having

Metalinguistic negotiations are worth having, because how we use our words matters. Conversational participants of metalinguistic negotiations feel the need to continue their disputes, even when it becomes common ground that they intend to mean different things by the same term, because they have independent reasons for the incompatible meanings of the expression they each pragmatically advocate for. In the above "fishy" case, although the merchant and the inspector both recognized that they intended to mean different things by the term "fish," they continued their dispute because they accepted different classifications and had economic consequences tied to the trial. In debate between gay right activists and conservatives over gay marriage, even if the conservatives

[8] Plunkett, D. and Timothy Sundell., "Disagreement and the Semantics of Normative and Evaluative Terms", *Philosophers' Imprint,* 13, (2013): 12.

[9] Ludlow, P., *Living Words Meaning Underdetermination and the Dynamic Lexicon* (Oxford: Oxford Univ. Press, 2018): 78.

[10] Hare, R.M., *The Language of Morals* (Oxford: Oxford Univ. Press, 1991): 148.

[11] Plunkett, D. and Timothy Sundell., "Disagreement and the Semantics of Normative and Evaluative Terms", *Philosophers' Imprint,* 13, (2013): 10.

make clear that they take “marriage” to mean a union between a male and a female, the activists will want to insist that this meaning is unacceptable in order to advance same-sex marriage rights.[12] Similarly, a feminist may advocate that “woman” should refer to people that are subordinated in virtue of their observed or imagined female bodily features, because such an ameliorated meaning of “woman” will advance the feminist cause.[13] All these examples demonstrate that there are all types of reasons for speakers to battle over the meanings of particular lexical items.

Speakers mostly accommodate others’ uses of words that differ from theirs to facilitate communication, unless they have fair enough reasons not to. These reasons concern the expressive aspects of lexical items that one cannot stipulatively do away with (for example, one cannot claim to use derogatory terms in a benign manner), the practical benefits of adopting certain meanings of terms that are fixed in legal documents (“marriage”, “rape”, “people”, etc.), the possibility of advancing social justice by advocating for meaning revision of various terms (“queer”, “woman”, etc.), and many more. The reasons to engage with metalinguistic negotiations may vary, but the case is strong enough to conclude that speakers can have good reasons to negotiate the use of expressions that they do not want to accept.

What makes metalinguistic negotiations worth having distinguishes them from *merely verbal disputes*, which arise solely in virtue of miscommunication and can be resolved after well-received clarification. Merely verbal disputes are non-canonical disputes that involve different meanings for common expressions but convey no genuine disagreements. These disputes arise solely in virtue of miscommunication: one speaker may take the other speaker to mean something different from what the other speaker takes herself to mean by a particular expression, and both speakers actually accept the other’s intended way of the using the expression after clarification.[14] Once it is common ground what each speaker takes the expression to mean, there is no disagreement between the speakers for them to continue the dispute.

2.4 In Light of the Conversational Common Ground

We could better see how metalinguistic negotiations unfold in practice in light of Robert Stalnaker’s common ground model. *Common ground*, identified as the common beliefs about what is mutually accepted, is both the end of the communicative action which seeks to add contents to the common ground and the means available to the speaker that is necessary for communication to take place. The speaker *accepts* that p as long as the speaker treats it as true for some reason, and it is *common belief* that p among a group of believers if and only if all believe that p, all believe that all believe that p, all believe that all believe that all believe that p, etc. The common ground includes the information that must be available in order for the utterance to be reasonably taken as an act of communi-

[12] Both examples here involve a legal setting in which words and expressions written in law could not be easily changed, a fact that motivates speakers to advocate for different meanings of the same term to use when the court and legal system interpret and apply the law. These legal disputes involving different meanings are typical metalinguistic negotiations that are worth having.

[13] Haslanger, S., “Gender and Race: (What) Are They? (What) Do We Want Them To Be?”, *Noûs*, 34, no. 1 (2000): 52.

[14] Chalmers, D.J., “Verbal Disputes”, *Philosophical Review*, 120, no. 4 (January 2011): 526.

cation.[15] This part of common ground includes the presuppositions the speakers expect the audience to have ready for use in making sense of what they say, many of which are tied to particular words and accompany their use. This information body within the common ground is what Mark Richard refers to as the interpretive common ground (ICG).[16][17] For example, when a competent English speaker speaks of bachelor using "bachelor", the speaker expects to be recognized as talking about an unmarried male and expects the audience to access this idea via proper interpretation. The speaker normally presupposes that "bachelor" refers to unmarried male, and hence, that bachelors are unmarried males. The latter follows by disquotation, which any competent speaker will hold to be analytically true. Given the speaker's presupposition that "bachelor" refers to unmarried males, the speaker will have pro tanto reasons to think that anyone who does not share this presupposition uses the term to mean something else. The kind of analyticity I am introducing here could be seen as follows: that bachelors are unmarried males is true for the speaker solely in virtue of the presupposition that "bachelor" refers to unmarried male. This argument is inspired by Kevin Scharp's analysis of constitutive principles for concepts as principles to guide interpretation. If a speaker rejects a principle that one takes to be constitutive for a concept in conversation, one then has a pro tanto reason to think that they do not mean the same thing by the word in question.[18]

The genuine disagreements in metalinguistic negotiations, which are conveyed pragmatically, are then expressed by manifesting presuppositions about the ICG. If a participant pragmatically conveys certain ways of using a term that violate part of what the other participant presupposes to be part of the ICG, the latter may refuse to accommodate by blocking the proposed way of using the term from becoming part of the ICG. The speakers then proceed to dispute whether the proposed way of using certain expression ought to be added into the ICG, negotiating the meaning of the expressions. In the fishy case, the merchant presupposed that "fish" is used to talk about cold-blooded vertebrate animal living wholly in water, and that fishes are cold-blooded vertebrate animals living wholly in water. On the other hand, the inspector presupposes that "fish" is used to talk about sea-creatures, and that fishes are just sea-creatures. Therefore, when the inspector asserted that whale oil is fish oil, the manifest occurrence of this assertion became common ground, which not only proposed to add the semantic content of the assertion to the common ground, but also pragmatically conveyed the metalinguistic message that it is appropriate to use "fish" to talk about whales. This, at the very least, violates the merchant's presupposition that fishes are vertebrate animals, which follows analytically from the presupposition that "fish" does not refer to mammalian sea creatures. Maybe it was still not clear to him that the inspector presupposed something completely different about how to use the term "fish", but it gave the merchant enough reason to start the metalinguistic negotiation.

[15] Stalnaker, R., "Common Ground", *Linguistics and Philosophy*, vol. 25, no. 5/6 (2002): 704.

[16] Richard, M., "The A-Project and the B-Project", *Conceptual Engineering and Conceptual Ethics*(2020): 363.

[17] While I am using Richard's term ICG to describe the presuppositions that a speaker expects her audience to recognize in order to make sense of what she seeks to communicate, I am distancing myself from his view that meaning of lexical items is, to a first approximation, interpretive common ground. In fact, my take on meaning is quite different from this view, which I will detail in my response to the externalist objection to the metalinguistic negotiation model in part three of the paper.

[18] 18 Scharp, K., "Philosophy as the Study of Defective Concepts", *Conceptual Engineering and Conceptual Ethics* (2020): 397.

2.5 The Issue of Dependence Between Presuppositions

One final note about metalinguistic negotiation model is that we must resist the misleading way of explaining canonical disputes in virtue of what is pragmatically conveyed about meaning. For example, when a college student meets a flat-Earther, the flat-Earther asserts that the Earth is flat. The college student responds by asserting that the Earth is not flat. Following the spirit of metalinguistic negotiation, one could argue that the two speakers seek to adjust the meaning of the term “Earth” in their upcoming dispute. When the flat-Earther asserts that the Earth is flat, the college student must have found it to be a violation of what she presupposes to be part of the ICG, that it is inappropriate to use “Earth” to designate something that is flat and that the flat-Earther probably means something different by the expression “Earth.”

Such analysis is blatantly wrong. The dispute between the flat-Earther and the college student is a canonical one in which the asserted propositions are incompatible. The Earth is not flat, and the flat-Earther asserts something false. There are also relevant factual disagreements between the speakers about the Earth: whether the Earth is a flat disc with the Arctic Circle in the center and whether there is a 150-foot-tall wall of ice around its rim. It is not a case of non-canonical dispute, let alone metalinguistic negotiation. Moreover, that the Earth is flat, or that the Earth is not flat, is not something that speakers presumably hold to be analytically true. That the Earth is not flat, for the college student, does not hold true solely in virtue of the meaning of the term “Earth”. Conversely, the college student only presupposes that it is inappropriate to use “Earth” to designate something that is flat in virtue of her belief that it is commonly accepted that the Earth is not flat. This is something that she possibly learns by listening to her parents as a child or reading a textbook, not just by mastering how to use the term “Earth.”

This illustrates the issue of *dependence* between presuppositions. The speakers presuppose that the *F*s are *G*s if and only if they presuppose that “the *F*s” refer to Gs. In some cases (type A), the speaker presupposes that the *F*s are *G*s because one presupposes that “the *F*s” refer to *G*s. In other cases (type B), the speaker presupposes the latter because one presupposes the former. The “fishy” dispute, as well as the other cases of metalinguistic negotiations, belongs to type A, while the flat-Earth dispute, as well as other cases of canonical disputes, belongs to type B. The presupposition that fish is not mammalian depends on the presupposition that “fish” refers to vertebrate animals, while the presupposition that “Earth” does not designate something flat depends on the presupposition that the Earth is not flat. It is even possible that the college student does not believe that this rule of appropriate use of the term “Earth” is common ground, if she does not expect everyone to accept that the Earth is not flat.[19] The distinction between type-A and type-B cases of dependence among presuppositions should help explain why the metalinguistic negotiation model should not apply to canonical disputes.

[19]19 While the idea that one should not expect someone to know that the Earth is not flat may seem wild (a poll in 2018 suggested that only two thirds of American millennial believe that the Earth is round), think of how experts adjust their presuppositions when addressing technical problems to their friends who have no expertise in the field. The experts may not presuppose that an average user of what they hold to be jargons should only use the words in certain ways, because the experts do not expect an average speaker to have the needed background expertise that drives one to use the words in these ways.

3 Speakers' Misconceptions of Metalinguistic Negotiations, Explained

One objection to the above picture of metalinguistic negotiation is that the model oftentimes goes against the speakers' own reflections and self-reports of what they are doing in what the model takes to be metalinguistic negotiations.[20] Many times the speakers are under the impression that they are not debating how to use particular words. This is the problem raised by Herman Cappelen against the metalinguistic model, that the speakers regard themselves as having a debate about torture, not "torture", when they debate whether waterboarding is torture. The discrepancy between the model's prediction and what the speakers take themselves to be doing in the waterboarding debate requires further explanation.[21] The paper has two responses to this speaker-error objection. First, it is difficult for speakers to recognize the non-canonical and metalinguistic nature of their disputes, but once they do it becomes difficult for Cappelen to explain what is going on. The speakers usually spend much time looking for potential factual disagreements, reflecting on what they presuppose to be analytically true of the expressions they use, and collecting cues from their ongoing dispute, before they realize that the literal contents of their assertions are compatible and confirm that their dispute is non-canonical and metalinguistic. The well-known dispute about whether Pluto is a planet well demonstrates such difficulty. When the astronomers at first tried to resolve their dispute over whether Pluto is a planet, they each brought about facts they thought to be relevant to the evaluation concerning Pluto's planet-hood. It soon became clear to many astronomers that they did not have a consensus over the meaning of the term "planet" because they failed to see why the facts raised by their colleagues were even relevant. The fact that astronomers could not reach a consensus about what is relevant in evaluating Pluto's planethood finally convinced them that they lacked a common way of using the term "planet" in the first place, revealing the metalinguistic nature of their dispute. Just like what happened in the Pluto case, conversational participants approach their non-canonical, metalinguistic disputes with the default assumption that it does not involve meaning differences. In other words, they do not expect to run into contents about words' meanings conveyed pragmatically against what they take to be within the ICG. Many times, the face that conversational parties mean different things by the same word only becomes salient as speakers continue their non-canonical metalinguistic dispute. Peter Ludlow calls the cues in conversation that allow conversational participants to recognize the meaning differences "triggers."[22] The trigger could be the realization that one conversational party has a broader or narrower modulation of a term ("fish" includes mammals or "athlete" excludes non-human animals). It could also be the realization of the need for both parties to sharpen a meaning ("planet") to resolve certain problems ("Is Pluto a planet?"). Nonetheless, given enough triggers from the conversation, the speakers will be aware that there is a meaning difference between them. They will realize that the genuine disagreement behind their dispute is the different meanings they pragmatically advocate for

[20] Cappelen, H., *Fixing Language: an Essay on Conceptual Engineering* (Oxford: Oxford University Press, 2018): 174.

[21] Cappelen, 175.

[22] Ludlow, P., *Living Words Meaning Underdetermination and the Dynamic Lexicon* (Oxford: Oxford Univ. Press, 2018): 40.

in conversation.[23]

Cappelen then has difficulty explaining these cases when speakers later admit that their dispute is about meaning differences. While the metalinguistic negotiation model allows for the conversational parties to gradually uncover the fact that their dispute is non-canonical and metalinguistic, Cappelen must insist that the conversational parties initially have the correct attitude that their disagreement is over one subject matter, only to get lost upon further evidence and reflections. If the speakers take in all the triggers and conclude that they are engaging with a dispute concerning the meaning of the contested term, Cappelen has to argue that these speakers now are mistaken about the nature of the dispute they have. Moreover, he has to say, somewhat implausibly, that their initial attitude is correct, but the one they have after they receive more information from the ongoing dispute is mistaken.

Besides the above difficulty of Cappelen's challenge, a simpler response is that most of the time speakers are disposed to seek factual ground for their normative disagreement. They are terrible at overtly normative disputes, so they tend to find some factual disagreement to distract themselves from the normative aspect of their disputes. This may happen because speakers have trouble specifying what they mean by certain expressions or what particular normative standards they have in mind when using the expressions under controversial circumstances. Feeling reluctant to appear as if they do not know what they mean when using these normative terms, the speakers may avoid directly engaging in the normative disagreement, driven by the recognized social expectation that they should know what they mean or what normative standards they uphold. Getting caught not knowing what you are talking about looks bad. Influenced by these social conditions, people tend to have long grinding debates about whether a dish tastes delicious, whether a picture looks beautiful, or whether it is morally good to do something or not in a unique fashion. They keep raising all types of details and facts that the other party may find irrelevant and talking past each other, without pausing to clarify that they share different aesthetical or ethical views or specify their normative standards. Direct engagement in normative disagreement is simply not the norm and there is no good reason to suspect that metalinguistic negotiation, with its normative aspect concerning the meanings of expressions, does not fall on the same line.

4 Externalist Arguments and the Limits of Metalinguistic Negotiations

Another important objection to the metalinguistic negotiation theory is that it is incompatible with semantic externalism.[24] The theory seems to assume that the speakers have

[23] An interesting question is whether it is still rational for the speakers to continue using instead of mentioning the term under dispute after it becomes common ground that the speakers mean different things by the same expression and aim to resolve the meaning disagreement. The speakers certainly could start to mention the term instead of using it to more directly litigate the meaning of the term. However, if the arguments for certain meaning of the term can be conveyed pragmatically, it can still be reasonable to continue the dispute by using the term and litigating the meaning of the term pragmatically.

[24] Cappelen, H., *Fixing Language: an Essay on Conceptual Engineering* (Oxford: Oxford University Press, 2018): 173.

control over the meanings of the terms they use. It takes for granted that the speakers could cooperatively resolve their metalinguistic dispute by battling over their ICG, raising reasons for or against certain meanings, and settling on a meaning that the speakers eventually come to accept. The "fishy" case could be resolved with the inspector and the merchant agreeing that "fish" means sea-creatures. The astronomers could resolve their dispute by having IAU sharpen the meaning of the term "planet." However, according to semantic externalism, this assumed control is largely illusory.[25] Many externalist arguments have shown that the grounding facts for meaning consist in part of various external, non-psychological facts: meanings ain't in the head. External factors shape, influence and determine the intension and extension of linguistic expressions and speakers' mental contents. These factors include features of the past such as the introductions of expressions (initial dubbing and "baptism" events), the causal and historic factors of how people use certain expressions, experts and social institutions, and the list goes on. In other words, presupposing words to mean something doesn't make it so.

Here are two examples inspired by Burge (1979) and Kripke (1980). If two naïve patients talk to each other and agree that they both have arthritis in the thigh, their joint agreement about how "arthritis" ought to be used will not change the fact that both of them have been using the term in a wrongful manner. If two speakers agree that from now on when they speak of "Jane" it no longer designates Jane but Henry, their consensus will not change the initial event of Jane's parents' naming her Jane in the past or disrupt all the causal chains connecting those who use the name and the initial naming event. The list of external *metasemantic facts* — facts that ground the meaning of linguistic expressions — is long and open. If all these metasemantic facts that fix the meaning of the terms are outside the speakers' control, it makes acts of metalinguistic negotiation hard to justify, because there is no way for interlocutors to fruitfully control the relevant metasemantic facts of meanings that they aspire to negotiate.[26]

My take of this externalist objection is that it demonstrates the limits of metalinguistic negotiations. At first glance, the scope of metalinguistic negotiations could be narrow and temporary. It works only on a conversation-by-conversation basis. Speakers may not use the meanings of expressions that they agree on in one conversation to navigate their future linguistic practices after the conversation is over. Their ICG updates proposed by their negotiation process can be fleeting, only effective within limited contexts and timeframes much like Ludlow's conception of *microlanguages*, one-off fleeting languages he thinks humans build and discard on a conversation-by-conversation basis.[27] Even if the speakers could agree on the meaning of certain expression in their conversation, the agreement does not necessarily have any influence on the conversation next door or anyone else that uses the expression. It is then important to recognize the difference between *metasemantic bas*, the grounding facts for meaning and reference, and *metasemantic superstructure*, which consists of our beliefs, presuppositions, preference, intentions, theories, and other attitudes about meanings and reference, what they are and what they ought to be.[28] Given that metalinguistic negotiations take place over the ICG, the changes pro-

[25] Cappelen, 74.

[26] Sterken, R.K., "Linguistic Intervention and Transformative Communicative Disruptions", *Conceptual Engineering and Conceptual Ethics* (2020): 422.

[27] Ludlow, P., *Living Words Meaning Underdetermination and the Dynamic Lexicon* (Oxford: Oxford Univ. Press, 2018): 3.

[28] Cappelen, H., *Fixing Language: an Essay on Conceptual Engineering* (Oxford: Oxford University Press, 2018): 59.

posed and made to the ICG throughout the process should be recognized to occur within the metalinguistic superstructure of the disputed expression instead of the metasemantic base. Metalinguistic negotiations aim at changing what the speakers accept certain expressions to mean, instead of what certain expressions actually mean. In other words, the negotiating efforts do not directly change the meanings of certain expressions but change the meanings the speakers accept for the sake of communication.

The difference between metasemantic superstructure and metasemantic base can be seen in how people communicate through encrypted messages. The speakers settle on a set of rules to take what certain expressions mean in their conditioned conversational context, without making a commitment about what these expressions mean for them in other contexts or the actual meanings of the expressions. In response to the "Lack of Control" principle of Cappelen that speakers have little or no control over what words mean, I would like to propose an "In Control" principle that speakers mostly have control over what they accept to be the meaning of certain linguistic expressions in their ICG through joint efforts of metalinguistic negotiations.[29] It is worth noticing that the "In Control" principle does not require more than the basic premises of Stalnakerian common ground model, which assume that speakers do at least have control over their common ground and what they accept. Therefore, semantic externalism is not directly at odds with the metalinguistic negotiation theory, since it only requires the speakers to have control over what they accept to be the meanings of linguistic expressions, not control over what the expressions actually mean. A more cynical view is that the speakers could make and even force certain people to misinterpret sentences containing particular expressions by exploiting the speakers' charitable interpretative nature, cooperativity and their common goal of successful communication. A more charitable reading could instead argue that the end of successful communication justifies the means of "forcing" misinterpretation. If all that the speakers want is effective communication, they could use the expressions as tools in whatever ways as long as they serve the purpose of successful exchange of information. Such an overly simplistic view apparently neglects all the other things speakers want to achieve through linguistic practices, which might as well constitute good or bad reasons for speakers to force misinterpretations that do good to them. However, it does explain how metalinguistic negotiation should be compatible with semantic externalism.

While it has been conceded that metalinguistic negotiation does not *directly* legislate the actual meaning of the term the disputants are using, it could still contribute to diachronic meaning changes, and influence these meaning shifts under the view of externalism. After all, in the "fishy" case, the inspector and the merchant cared about how "fish" is used in their court dispute because they believed the dispute would affect the court-sanctioned meaning of "fish" and, in turn, affect taxation. But even the idea that the court decision could effectively change the public meaning of terms requires further scrutiny. The dispute could still affect the taxation result, even if the court, without changing the public meaning of "fish", forced a misinterpretation of the term in the legal context. The court's ruling may only work on the metasemantic superstructure of these expressions, by forcing the participants of trials to accept certain meanings of expressions in court with the aid of the state and its law enforcement, without actually impact-

[29] Cappelen, 72.

ing the metasemantic bases and actual meanings of the expressions.[30] It is even unclear whether the court has the supposed control over the public meaning of these expressions since the court may not have much control over the metasemantic base of words after all. The argument is that the metasemantic mechanism that generates current meanings and drives meaning changes remains largely *inscrutable.*[31]

It then becomes difficult to come up with a clear and complete picture of how individual metalinguistic negotiation each contributes to possible public meaning shifts. Still, I want to end the paper by making a few proposals. Consider the example of the diachronic meaning shift of "salad". Not too long ago, "salad" meant a cold dish served with high preponderance of green leaves. A concoction of cold cut fruit would not qualify as salads. Nowadays, however, "salad" could designate various warm leaf-free concoctions and concoctions of cold cut fruits. The meaning of "salad", both its intension and extension, is observed to have shifted through time.[32] It is not hard to postulate how metalinguistic negotiations on a conversation-by-conversation basis could contribute to such diachronic meaning shift. Maybe at certain point some chefs without much deliberate coordination started calling their new warm leaf-free concoctions "salad" in their menus. Many customers rejected this practice, since they found it inappropriate. Many conversations took place, involving "this is not a salad," "the salad has no green leaves," or "this is not the salad that I ordered." Through metalinguistic negotiations, some customers would accept the proposed changes to the ICG between them and the chefs, and carry the changes with them in future conversations, presupposing that "salad" could be used to talk about warm leaf-free concoctions. The proposed update to the metasemantic superstructure then occurs on a large scale as words spread and changes are continuously being made to the metasemantic superstructure of "salad", which, in an inscrutable and yet-to-be-explored manner, continuously influence the metasemantic base of the word "salad". When the change becomes salient enough, someone makes the observation that the meaning of "salad" has changed. Therefore, as long as an externalist semantic theory allows for and seeks to explain the diachronic meaning shifts, metalinguistic negotiation, with its limits and scope of influence over the ICG and metasemantic superstructures of interlocutors, could still contribute to and explain certain types of meaning changes in various ways that require further investigation. For example, if externalists are to believe that complex use patterns of lexical items overtime constitute part of the metasemantic base of language, metalinguistic negotiations, which revise speakers' accepted way of using terms and influence how they proceed to use those terms, apparently influence speakers' use patterns and contribute to the changes of the metasemantic base. It is not hard to imagine that countless non-canonical metalinguistic disputes over whether a gay couple could *marry* each other, whether a husband could *rape* his wife, or whether whale oil is fish oil contribute to the supposed meaning changes of "marry," "rape" and "fish oil," as speakers, via metalinguistic negotiations, discard the old ways and accept new ways of using these terms and bring about use patterns of these terms that constitute part of the metasemantic base.[33]

[30] Cappelen, 76.

[31] Cappelen, 73.

[32] Dorr, C. and Hawthorne, J., "Semantic Plasticity and Speech Reports", *Philosophical Review*, 123, no. 3 (January 2014): 284.

[33] The same analysis might not apply to natural kind terms, and could work at its best in explaining meaning changes of terms of groupings that people agree are a matter of choice.

5 Conclusion

It is best to explain non-canonical disputes that seem to express genuine disagreements literal expression of incompatible content through the model of metalinguistic negotiation. While the model validates the observation that speakers mean different things in these disputes, it also proves to be compatible with semantic externalism. Metalinguistic negotiations only directly work on what the speakers take the words to mean instead of the words' actual meanings that are partially fixed by external non-psychological factors. However, this does not mean that metalinguistic negotiations cannot contribute to meaning changes of words at all. They still influence and explain meaning changes by exploiting an inscrutable relationship between their metasemantic superstructure and metasemantic base that requires further investigation.

Bibliography

Burge, Tyler. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4 (1979): 73–121.

Cappelen, Herman. *Fixing Language: An Essay on Conceptual Engineering.* Oxford University Press, 2018.

Chalmers, David J. “Verbal Disputes.” *Philosophical Review* 120, no. 4 (2011): 515–66.

Dorr, Cian, and John Hawthorne. “Semantic Plasticity and Speech Reports.” *Philosophical Review* 123, no. 3 (2014): 281–338.

Hare, Robert M. *The Language of Morals.* Oxford University Press, 1991.

Haslanger, Sally. “Gender and Race: (What) Are They? (What) Do We Want Them To Be?” *Noûs* 34, no. 1 (2000): 31–55.

Kripke, Saul A. *Naming and Necessity*. Harvard University Press, 1980.

Ludlow, Peter. *Living Words Meaning Underdetermination and the Dynamic Lexicon.* Oxford University Press, 2018.

Plunkett, David & Sundell, Timothy (2013). "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13 (23):1-37.

Recanati, François. “What Is Said.” *Synthese* 128, no. 1/2 (2001): 75–91.

Richard, Mark. “The A-Project and the B-Project.” *Conceptual Engineering and Conceptual Ethics* (2020): 358–78.

Sainsbury, Mark. “Fishy Business.” *Analysis* 74, no. 1 (2013): 3–5.

Scharp, Kevin. “Philosophy as the Study of Defective Concepts.” *Conceptual Engi-*

neering and Conceptual Ethics (2020): 396–416.

Stalnaker, Robert. "Common Ground." *Linguistics and Philosophy* 25 (2002): 701–721.

Sterken, Rachel Katharine. “Linguistic Intervention and Transformative Communicative Disruptions.” *Conceptual Engineering and Conceptual Ethics* (2020): 417–34.

Fictional Truth, Fictional Names: A Lewisian Approach

Sam Elliott

University of Edinburgh

The account of fictional truth proposed by David Lewis in his seminal 1978 paper "Truth in Fiction" remains of central importance to much contemporary discussion of this issue — namely, how we should analyse what is, so to speak, 'true in a fiction'. Despite this, Lewis says relatively little about fictional names as such, nor have Lewis's views on fictional names received much scholarly attention — surprising, given the extent to which the issues of fictional truth and fictional names overlap. In this paper I argue that Lewis's account of fictional truth forces us to adopt an account of fictional names as non-rigid designators, whose reference is fixed satisfactionally at a given world. However, as such, I argue that Lewis's account is vulnerable to challenges analogous to Kripke's criticisms of classical descriptivism: namely, that this account is seemingly incompatible with intuitively coherent patterns of 'counter-fictional' reasoning.

1 Introduction

For as long as names have been considered a worthy subject of philosophical inquiry — that is, at least, since Frege — fictional names have posed a problem: we can say seemingly true things with fictional names, which seemingly do not refer to any individual (at least, not to any ordinary individual). How can this be?[1]

The solutions to this problem bifurcate along the lines of this tension. In order to resolve the problem, we could allow that fictional names do (or, at least, can) refer — but then to what do they refer? Alternatively, we could reject the possibility of saying true things with fictional names — but then how are we to explain our seeming ability to do just that?

In this paper I will trace the route through this garden of forking paths followed by David Lewis in his seminal paper "Truth in Fiction,"[2] pausing briefly to justify his choices at each junction. I will then consider how the commitments Lewis makes lay the groundwork for a plausible account of fictional names. However, I argue that such an account faces some robust difficulties. I leave it open to the reader to decide whether, in light of these difficulties, we ought to prefer to continue down this path.

[1] Gottlob Frege, "Über Sinn und Bedeutung," Zeitschrift für Philosophie und philosophische Kritik 100 (1892): 25–50. Translated as "On Sense and Reference" in Translations from the Philosophical Writings of Gottlob Frege, ed. and trans. Max Black and Peter Geach (Oxford: Blackwell, 1980).

[2] David Lewis, "Truth in Fiction," *American Philosophical Quarterly* 15, no. 1 (1978): 37-46, https://www.jstor.org/stable/20009693.

2 A Problem of Fictional Names

We can say seemingly true things with fictional names. Consider this simple comprehension question on Shakespeare's *Hamlet*:

(1) Hamlet is Danish.

This sentence seems true. If a student, say, were to utter (1), in the context of a discussion about *Hamlet*, we should (plausibly) take them to have said something true.[3]

If we follow the surface grammar of (1), we may take it as having the grammatical form of a typical subject-predicate sentence: ascribing a property — being Danish — to an individual, referred to by means of a proper name, 'Hamlet'. A standard truth-conditional analysis of such a sentence would hold that it is true if and only if the referent of the name satisfies the property expressed by the predicate.

However, it is tempting to suggest that, almost axiomatically, *fictional names* like 'Hamlet' do not refer to anyone — if they did, they would not be genuine *fictional names.*[4] A standard treatment of proper names would hold that a proper name (on an occasion of use) refers to an individual by virtue of being part of a name-using practice, originating in that individual being dubbed with that name.[5] However, almost by definition, our use of the name 'Hamlet' did not originate in some person being dubbed with the name. We might be tempted to say, then, that Shakespeare "just made up the name:"[6] it does not refer to any individual; it is an empty name.

We have an obvious tension here: if 'Hamlet' does not refer to anyone, it cannot be that the referent of 'Hamlet' satisfies the property expressed by the predicate 'is Danish'; so (1) cannot be true.

This tension can be spelled out in three claims:

(A) The sentence 'Hamlet is Danish' is true.

(B) The name 'Hamlet' does not refer to any individual.

(C) The sentence 'Hamlet is Danish' is true iff: (i) 'Hamlet' refers to some individual; and, (ii) the referent of 'Hamlet' satisfies the property expressed by 'is Danish'.

All three claims are intuitively plausible; however, as I hope is clear, they cannot all be held in conjunction.

[3] Graham Priest, "Sylvan's Box: a Short Story and Ten Morals," *Notre Dame journal of Formal Logic* 38, no. 4 (1997): 573-582. This use of a 'comprehension test' is inspired by Priest.

[4] 4 We assume here and throughout that 'Hamlet' is a paradigmatic example of a genuine *fictional name.*

[5] Saul Kripke, *Naming and Necessity* (Hoboken: Wiley-Blackwell, 1981). This standard treatment follows a rough *causal picture of reference,* as popularised by Kripke.

[6] David Kaplan, "Bob and Carol and Ted and Alice," in *Approaches to Natural Language,* ed. Jaakko Hintikka (Dordrecht: Reidel, 1973), 505.

This tension is not unique to the name 'Hamlet'. Fictional names pose a general problem for semantic analysis, as we can use them to ascribe properties to fictional characters; making (seemingly) true claims, despite their (seemingly) failing to refer.

3 Do Fictional Names Refer?

Many suggested resolutions to this tension start by rejecting B, holding that fictional names like 'Hamlet' do refer — though not to ordinary flesh-and-blood persons. Such solutions may (broadly) be described as 'realist', insofar as they hold that — *ontologically* speaking — there are such things as fictional characters, and that fictional names refer to these characters.[7] However, we should then ask: if fictional characters are 'real', and fictional names refer to them, what sort of thing are they — metaphysically speaking? Within this broad ontological 'realist' position, there is significant divergence in answering this metaphysical question.[8]

One common 'realist' approach is to follow the Meinongian line that fictional characters are *non-existent* individuals.[9] Parsons summarises such a position:

> "Sherlock Holmes, for example, is an object that is a detective, solves crimes, ..., and doesn't exist. His nonexistence doesn't prevent him from having (in the actual world) quite ordinary properties, such as being a detective."[10]

This, however, is chiefly a negative thesis — fictional characters are not existent. We should be inclined to ask the Meinongian for some positive metaphysical thesis to supplement this claim. Again, suggestions on such positive theses diverge. Parsons, for instance, holds that fictional characters (qua non-existents) are "concrete correlates of sets of properties;"[11] in contrast, Zalta and Stokke maintain that fictional characters are "roles" or "individual concepts," specified by sets of properties.[12]

If these suggestions are already beginning to look too metaphysically obscure for our tastes, we might prefer to backtrack a little, and consider an alternative realist line. The most common such alternative would, likely, follow the thesis that fictional characters are *abstract* individuals. Again, such a thesis is primarily negative — fictional characters are *not* concrete — and ought to be supplemented with some positive metaphysical

[7] However, in order to knit with our intuitions, such fictional characters must be different in kind to ordinary flesh-and-blood people.

[8] Amie Thomasson, *Fiction and Metaphysics* (Cambridge: Cambridge University Press, 1999). Thomasson draws this distinction between the *ontological* and *metaphysical* issues of fictional characters: *'are there such things as fictional characters?'*; and (if so), *'what kind of thing is a fictional character?'*.

[9] 9 Fred Kroon and Alberto Voltolini, "Fictional Entities," *The Stanford Encyclopedia of Philosophy* (2018), https://plato.stanford.edu/archives/win2018/entries/fictional-entities; Andreas Stokke, "Fictional Names and Individual Concepts," *Synthese* (2020): 1-31. Per Kroon and Voltolini, and Stokke, we should be careful about attributing any of these views too stringently to Meinong himself. Rather, we say that these views are broadly "Meinongian."

[10] Terence Parsons, "Fregean theories of fictional objects," *Topoi* 1 (1982): 81.

[11] Parsons; Kroon and Voltolini.

[12] Edward N. Zalta, *Abstract Objects: An Introduction to Axiomatic Metaphysics* (Dordrecht: Reidel, 1983); Andreas Stokke, "Fictional Names and Individual Concepts."

claim. Thomasson is a forceful defender of one such thesis;[13] arguing that fictional characters are *artifacts*: "created, dependent abstracta present in the actual world";[14] "created objects dependent on such entities as authors and stories."[15]

I will not consider the relative merits of these metaphysical options here. There are, however, as noted by Martin and Schotch, general reasons to be wary of adopting any such ontologically realist line.[16] For one, all such lines are ontologically costly, requiring that we accept an ontology containing not only ordinary, concrete, existent individuals, but some other sort(s) of individuals too (e.g. *non-existents*, or *abstracta*). If available, a solution to the problem which did not require such ontological commitments might be considered theoretically preferable. Second, even granting such ontological commitments, we may (reasonably) be sceptical of the legitimacy of attributing ordinary properties — such as being a detective, or being prince of Denmark — to these *'metaphysically-other'* individuals. Such attributions, however, are necessary to preserve the intuitive truth-value of sentences like (1).[17]

Considering the thorny issues which lie in store, should we venture down the realist path, we may be inclined to choose an alternative 'anti-realist' track. The starting point for any such track would be to uphold B — our original intuition, that fictional names do not refer (to any kind of individual). Since this 'anti-realism' amounts to a dismissal of the ontological question, the difficult metaphysical questions which troubled the realist simply do not arise. However, by upholding B, the anti-realist is forced to conclude that (1) is not true (at least, not strictly speaking).

What Lewis attempts to show,[18] first of all, is that there is a path available to the anti-realist which allows them to uphold ¬ A and B, and yet give a satisfying explanation of the intuitive truth of (1). He argues that it is available to us to say that, although (1) is not strictly true, it can be used to say something true. More precisely, he claims, when someone uses or utters the phrase "Hamlet is Danish," we may (in certain contexts) take them to have implicitly asserted, not (1), but the more complex sentence:

(2) In *Hamlet*, Hamlet is Danish.

We draw an implicit distinction here between what a speaker *utters* — *the precise words they use* — and what they *assert* — very roughly, *the point they express*.

In effect, the addition of the prefix 'In Hamlet...' makes explicit that the claim being asserted is *metafictional*: it does not concern what is true *simplicite*, but rather what

[13] Amie Thomasson, "Fiction, Modality and Dependent Abstracta," *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 84, no. 2 (1996): 295-320; "Speaking of Fictional Characters," *Dialectica* 57, no. 2 (2003): 205-223.

[14] Thomasson, "Fiction, Modality and Dependent Abstracta", 296.

[15] Thomasson, 301. For others, see: John Searle, "The Logical Status of Fictional Discourse," *New Literary History* 6, no. 2 (1975): 319-332; Peter van Inwagen, "Creatures of Fiction," *American Philosophical Quarterly* 14, no. 4 (1977): 299–308, https://www.jstor.org/stable/20009682; Nathan Salmon, "Nonexistence," *Noûs* 32, no. 3 (1998): 277–319; Saul Kripke, *Reference and Existence: The John Locke Lectures* (Oxford: Oxford University Press, 2013).

[16] Robert M. Martin and Peter K. Schotch, "The Meaning of Fictional Names," *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 26, no. 5 (1974): 378.

[17] Kripke, *Reference and Existence.* Kripke maintains such a view: broadly, that fictional characters are abstract objects, but that — as such — they are not eligible to satisfy ordinary properties such as being Danish, or being a detective; instead they satisfy fictional-analogues of ordinary properties, such as being *fictionally-Danish*, or being a *fictional detective.*

[18] Lewis, "Truth in Fiction."

is ‘true-in-*Hamlet*’.[19] Though (1) is not true *simpliciter*, it is true-in-*Hamlet*; or true, when situated within the scope of the prefix ‘In *Hamlet*…’.

This may be used as a general strategy for understanding metafictional discourse: we may take seemingly straightforward utterances as (abbreviated) *metafictional assertions* — implicitly prefixed with an operator of the form ‘In such-and-such a fiction…’ (or, a *fiction-operator*).[20]

This path leaves us with the following commitments:

(¬**A**) The sentence ’Hamlet is Danish’ is **not** true.

(B) The name ‘Hamlet’ does not refer to any individual.

(C) The sentence ‘Hamlet is Danish’ is true iff: (i) ‘Hamlet’ refers to some individual; and, (ii) the referent of ‘Hamlet’ satisfies the property expressed by ‘is Danish’.

(D) An utterance or use of the phrase “Hamlet is Danish” may (in some contexts) be taken as an (abbreviated) assertion of the sentence ‘In *Hamlet*, Hamlet is Danish’.[21]

(E) The sentence ‘In Hamlet, Hamlet is Danish’ is true.

4 How should we analyse (2)?

The question which then presents itself is: how should we analyse (2)? Since we are committed to the claim that (2) is true, what are its truth-conditions?

4.1 How should we analyse the fiction-operator, ’In *Hamlet*...’?

We may begin by considering the fiction-operator, ‘In *Hamlet*…’. As should be clear on consideration, the operator ‘In *Hamlet*…’ is not truth-functional: we cannot determine the truth-value of a sentence ‘In *Hamlet*, ϕ’ as a function of the truth-value of the embedded sentence ϕ.

[19] Stokke: 2. I borrow this concept of *metafictional* discourse from Stokke: “On its metafictional use [a sentence] is used to say something about what is true in [a] story.”

[20] Moves along these lines abound in the literature. See: Martin and Schotch; Gregory Currie, “Fictional Truth,” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 50, no. 2 (1986): 195-212; Alex Byrne, “Truth in fiction: The story continued,” *Australasian Journal of Philosophy* 71, no. 1 (1993): 24-35; John F. Phillips, “Truth and Inference in Fiction,” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 94, no. 3 (1999): 273-293, https://www.jstor.org/stable/4320938; Diane Proudfoot, “Possible Worlds Semantics and Fiction,” *Journal of Philosophical Logic* 35 (2006): 9-40; Stacie Friend, “The great beetle debate: a study in imagining with names,” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 153, no. 2 (2011): 183-211; Stokke.

[21] Lewis: 38. How are we to decide when an utterance is to be taken as a metafictional assertion, implicitly prefixed by some fiction-operator? As Lewis says, “context, content, and common sense will usually resolve the ambiguity.” Comprehension tests, like that with which we began, are useful for necessitating metafictional discourse.

Lewis's suggestion starts from the point of regarding 'In *Hamlet*...' as a modal operator. This analysis allows Lewis to treat metafictional discourse from within the (already developed) framework of a *possible world semantics*, of which he is a noted advocate. The key tenet of such a framework is that modal discourse — about possibility and necessity — should be understood and interpreted as discourse about possible worlds, and possible individuals.

Lewis subscribes to the view that modal discourse involving the notions of *possibility* and *necessity* is inherently opaque (without clear truth-conditions or standards of validity). He holds that this opaqueness can only be dissolved by giving truth-conditional analyses of modal sentences in terms of quantification over possible worlds.[22] In particular, the modal operators 'necessarily' and 'possibly' should be translated as universal and existential quantifiers, ranging over a domain of possible worlds.

Similarly, Lewis suggests that fiction-operators should be interpreted as "*relative* necessity operators," and analysed as "*restricted* universal quantifiers over possible worlds."[23] More intuitively, the fiction-operator 'In *Hamlet*...' serves to identify some relevant domain of possible worlds; and the truth-value of a sentence of the form 'In *Hamlet*, ϕ' is determined by the truth-value of the embedded sentence φ at each world in the relevant domain.

This analysis outlines an approach for defining truth-conditions for sentences within the scope of fiction-operators. The Lewisian suggestion can be stated as follows:

(F) For any sentence φ 'In Hamlet, ϕ' is true iff ϕ is true at each of some set of possible worlds ("this set being somehow determined by *Hamlet*").[24]

This approach has a couple of obvious advantages. Firstly, there is something intuitively satisfying about identifying truth-in-*Hamlet* with truth at some possible worlds. We often have recourse to talk about 'the world of the fiction', or of storytelling as a 'world-building' exercise. Lewis's approach directly echoes this picture.[25] Secondly, analysing 'In *Hamlet*...' as a restricted universal quantifier over possible worlds allows discourse about truth-in-*Hamlet* to be subsumed into a general possible world semantics, with clear and definite standards of valid inference.

In addition to ¬A – E, we now find ourselves committed to the additional claim:

(G) For any sentence φ, 'In *Hamlet*, ϕ' is true iff ϕ is true at each of some set of possible worlds (which we may denote '$\triangle$').

[22] David Lewis, "Anselm and Actuality," in *Philosophical Papers Volume I* (Oxford: Oxford University Press, 1983), 10. "The standards of validity for modal reasoning have long been unclear; they become clear only when we provide a semantic analysis of modal logic by reference to possible worlds and to possible things therein."

[23] Lewis, "Truth in Fiction:" 39.

[24] Lewis: 38. We ought really to specify a type of modality here, in virtue of which a world is *possible*. It is clear from Lewis's work that the relevant type is *logical* or *metaphysical* modality. As such, when we describe a world as 'possible', we should understand that as meaning '*logically* or *metaphysically* possible'.

[25] In general, fictions do not specify enough to determine a single possible world as 'the world of the fiction'. Worlds are *complete* — they settle every question. As such, there are otherwise identical worlds, differing only in Ophelia's blood type. Which of these is *the* world of *Hamlet*? It seems arbitrary to choose between them.

4.2 How should we analyse the embedded sentence, (1)?

The analysis of the operator 'In *Hamlet*...' given in G, in conjunction with our existing commitment E, imply a further commitment:

(H) The sentence 'Hamlet is Danish' is true at each possible world in the set $\triangle$.

As should be apparent, in order to make sense of this claim, we must think of (1) not as having a single truth-value, but as only having a truth-value *at a world* (or, relative to a state of affairs).[26] For instance, we may say that, even though (1) is false at *the actual world* (denoted '@'), it may be true at *some other possible world*, or at each of some set of worlds.

However, in order to allow the truth-value of (1) to vary between worlds, we must also relativise its *truth-conditions* to a world. As such, as (1) contains a singular term and a predicate, we must also relativise the relations of *reference* and *property-satisfaction* to worlds.

In this vein, drawing on the standard analysis of (1) given in C, we may say that (1) is true **at a world w** if and only if the referent of 'Hamlet' **at w** satisfies the property expressed by 'is Danish' **at w**. This states a further commitment:

(I) The sentence 'Hamlet is Danish' is true at a world w iff: (i) 'Hamlet' refers to some individual at w; and, (ii) the referent of 'Hamlet' at w satisfies the property expressed by 'Danish' at w.

From H and I, follows a further commitment:

(J) For any possible world w in the set $\triangle$: (i) 'Hamlet' refers to some individual at w; and, (ii) the referent of 'Hamlet' at w satisfies the property expressed by 'is Danish' at w.

5 How should we analyse the reference-conditions of 'Hamlet'?

This commitment raises a further question: what are the (world-relative) reference-fixing conditions for 'Hamlet': what conditions must obtain for 'Hamlet' to refer to an individual (at a world)?

[26] Strictly speaking, we should also relativise the truth of (2) to a world, by means of an *accessibility* relation. Lewis does not consider this: he is only concerned (as we will be here) with the semantics of fiction-operators with respect to the actual world.

5.1 Ordinary Proper Names

In many ways, 'Hamlet' looks and behaves like an ordinary proper name. The standard treatment of such 'ordinary' proper names (viz. reference and modality), popularised by Kripke,[27] comprises two elements. First, the reference of an 'ordinary' proper name (on an occasion of use) is determined at the actual world, by means of something like the *causal picture of reference* we considered earlier. Second, when considered with respect to other possible worlds, an 'ordinary' proper name refers at any world to the individual to whom it refers at the actual world: 'ordinary' proper names are *rigid designators* (across possible worlds). (As Kripke demonstrated, this property of rigidity is central to our ability to engage in modal reasoning using proper names.)[28]

This property of rigidity can be spelled out in the following claim:

(K) For any ordinary proper name 'α', any world w, and any individual x: 'α' refers to x at w if and only if 'α' refers to x at @.

As Kaplan observes,[29] a natural corollary of this claim is as follows:

(L) For any ordinary proper name 'α': if 'α' does not refer to any individual at @, then 'α' does not refer to any individual at any world.

As such, if we were to treat 'Hamlet' as an ordinary proper name (*qua* rigid designator), then it follows, from B and L, that 'Hamlet' does not refer to any individual at any possible world. The only way to reconcile this conclusion with J, would be to hold that the set $\triangle$ is empty. However, this conclusion, in conjunction with F, would make every sentence *vacuously* true-in-*Hamlet*.

Seemingly, then, the only way to reconcile our Lewisian analysis of fiction-operators as modal operators, with an interpretation of fictional names as rigid designators, would be to completely obscure the distinction between fictional truths and fictional falsities.[30]

As such, in order to hold on to our existing commitments, we must treat fictional names differently from 'ordinary' proper names (on the standard Kripkean treatment): a move which raises a suspicion of *ad hoc*-ness.

Moreover, realist approaches are, at least in theory, compatible with an interpretation of fictional names as rigid designators. Such approaches, therefore, seem to constitute a more promising option for providing a uniform semantical account of fictional and

[27] Kripke, *Naming and Necessity*

[28] David Lewis, "Counterpart Theory and Quantified Modal Logic," *The Journal of Philosophy* 65, no. 5 (1968): 113-126. This concept of rigidity relies on a primitive notion of trans-world identity for individuals. Lewis is critical of this notion, preferring to explicate talk of transworld identity by means of his *Counterpart Theory*, a primary tenet of which is that no individual can inhabit more than one world. However, in "Truth in Fiction" he adopts the conventional language of transworld identity, which we use here.

[29] Kaplan, 502

[30] It is important to note that this is simply a consequence of Lewis's analysis of fiction-operators as *modal* operators, and of fictional names as being used meaningfully within the scope of these modals. It does not depend on the finer points of Lewis's analyses, and applies equally to any way he might flesh out the restriction on possible worlds, $\triangle$.

'ordinary' names. Such uniformity would, plausibly, be considered a theoretical value. Thomasson,[31] and Adams, et al.[32], both attempt to offer such uniform accounts.

5.2 Descriptivism

Given that, as we have seen, a Lewisian analysis of fiction-operators as modal operators is essentially incompatible with an interpretation of fictional names as rigid-designators, we find ourselves in need of an alternative analysis of fictional names: one which assigns them non-rigid (world-relative) reference-fixing conditions.

One potentially attractive suggestion might be to let the reference of the name 'Hamlet' be fixed (at a world) *satisfactionally*, by the descriptions given of Hamlet in *Hamlet*. Approaches along these lines are given by Martin and Schotch, and Currie,[33] amongst others.

With some degree of idealisation, we can extract from the text of *Hamlet* some set of attributes, relations and deeds ascribed to the character Hamlet. Such a set might include the following: 'is Danish'; 'is a prince'; 'is slain by a poisoned blade', etc. Intuitively, we can regard such a set of descriptions as constructing a specification, or a 'sketch', of the character Hamlet. A very general outline of a descriptivist approach to resolving the reference-conditions of 'Hamlet' might, then, be to say that 'Hamlet' refers to an individual at a world if and only if that individual matches this 'sketch' of Hamlet at that world.

There are numerous ways this broad condition may be fleshed-out. We might say, for instance, that 'Hamlet' refers to an individual at a world if that individual satisfies **all** the descriptions given of Hamlet in Hamlet at that world; or some **majority** of these descriptions; or **best** matches the inferred 'sketch' (of the individuals available at that world). We shall not worry about this here.

A version of this descriptivist suggestion can be stated broadly as follows (with the necessary addition of a *uniqueness* condition):

> **(M)** For any world w, and any individual x: 'Hamlet' refers to x at w if and only if x uniquely satisfies all (or most) of the descriptions given of Hamlet in *Hamlet*.

This suggestion, however, has a rather disquieting consequence — the observation of which is due to Kripke.[34] Consider: we can perfectly well imagine discovering that some individual actually existed, who uniquely satisfied all or most of the descriptions given of Hamlet in *Hamlet*. In Lewis's words, this imagined individual "had the attributes, stood in the relations, and did the deeds" ascribed to Hamlet in *Hamlet*. Would this individual then be Hamlet? Would the name 'Hamlet' — *as we use it* — refer to him?

[31] Amie Thomasson, "The Reference of Fictional Names," *Kriterion* 6 (1993): 3-12, http://www.kriterion-journal-of-philosophy.org/kriterion/issues/Kriterion-1993-06/Kriterion-1993-06-03-12-thomasson.pdf.

[32] Fred Adams, Gary Fuller and Robert Stecker, "The Semantics of Fictional Names," *Pacific Philosophical Quarterly* 78 (1997): 128-148.

[33] Gregory Currie, *The Nature of Fiction* (Cambridge: Cambridge University Press, 1990).

[34] Kripke, *Naming and Necessity*.

If we hold a condition like M, and allow the reference of 'Hamlet' to be fixed satisfactionally (at a world), then we should have to answer in the affirmative. Such an individual would, indeed, be Hamlet. Our name 'Hamlet' would refer to him. This, however, may not be a welcome conclusion. Kripke and Lewis both find it intolerable.[35] Nevertheless, there is no absolute consensus on this claim, with Martin and Schotch, for instance, who advocate for something like the satisfactional account just given, opting to bite the bullet in the face of this Kripkean challenge.

5.3 A Lewisian Hybrid

Lewis offers a further alternative to these two treatments. His view is, in essence, a hybrid, holding, with the descriptivist, that fictional names are non-rigid, but following Kripke in denying that matching the 'sketch' of Hamlet is sufficient for being Hamlet (or being the referent of the name 'Hamlet').[36]

His suggestion can best be explained and motivated by considering what one might take to be wrong with the previous descriptivist suggestion. Let us consider the problem again: we can well imagine discovering that some individual actually existed matching the 'sketch' of Hamlet. Would this individual *be* Hamlet?

"Surely not!", says Lewis.[37] But why not? Lewis's reaction, I believe, stems from an implicit assumption of a broadly *causal* picture of reference, whereby names refer to individuals (on an occasion of use) in virtue of being part of a name-using practice originating in that individual being dubbed with the name.

As Kaplan says, however, in the case of fictional names, this is simply not the case.[38] The name 'Hamlet', as we use it, does not refer to any individual, because it does not originate in some individual being dubbed with the name — rather, Shakespeare just made it up. The possibility of a Hamlet-doppelgänger is irrelevant: if our name 'Hamlet' did not originate in their being dubbed with the name, then the name does not refer to them.

We can co-opt the metaphor of the sketch here, to good effect. Just as a literal sketch is not a sketch *of* a particular individual by virtue of that individual bearing a likeness to the sketch, so someone is not the referent of our name 'Hamlet' in virtue of matching the 'sketch'. A sketch is *of* an individual in virtue of bearing some causal, intentional relationship to that individual. Similarly, an individual is the referent of a name, like 'Hamlet', in virtue of some bearing causal, intentional relation to the name.[39]

On these grounds, Lewis, as I read him, would hold that we can imagine discovering that Hamlet really existed; or that the name 'Hamlet', as we use it, *does* refer to an

[35] Kripke, *Naming and Necessity*; Lewis, "Truth in Fiction."

[36] Lewis says relatively little about fictional names as such. My reading of him here draws primarily from his reflections on fictions being "told as known fact."

[37] Lewis: 39.

[38] Kaplan, 505

[39] Kripke uses a similar metaphor in motivating his causal picture.

actual person. Consider: we can imagine discovering that Shakespeare did not write *Hamlet* as a work of fiction, but as a factual biography of a real Danish prince and his tragic demise. We can imagine that some individual actually existed who "had the attributes, stood in the relations, and did the deeds" ascribed in *Hamlet* to Hamlet and that Shakespeare wrote the story *Hamlet* about this individual. We should be inclined to say, then, that our use of the name 'Hamlet' is part of a name-using practice originating in this individual being dubbed with the name — either by Shakespeare, or previous to his usage.

Such a scenario, we assume, is not actual; but it is conceivable, and therefore describes a possible world — a possible world where the story *Hamlet* is told, just as it is at the actual world, but where it is "told as known fact."[40] This, roughly, then, is Lewis's suggestion:

> **(N)** For any world w, and any individual x: 'Hamlet' refers to x at w iff *Hamlet* is truly told *about* x at w.[41]

There is reason to be sceptical of such a hybrid view. Lewis concedes, as his modal analysis requires of him, that fictional names are non-rigid. They are, as such, distinctly unlike 'ordinary' proper names, at least in their modal profile. His rejection of the descriptivist suggestion, however, seems to be rooted in a tacit subscription to the sort of causal picture of reference which Kripke endorses for proper names. What may be a little unclear to us is, if Lewis already distinguishes between fictional and ordinary names with respect to rigidity, why should we credit the analogy with regards their actual reference-fixing conditions?

6 A Further Problem

A further problem emerges here, threatening both the descriptivist and the Lewisian lines. We may observe that, despite their differences in content, the reference-fixing conditions given by M and N share the same formal structure:

> **(O)** For any world w, and any individual x: 'Hamlet' refers to x at w iff x uniquely satisfies some condition ξ at w.

Both conditions are, in essence, satisfactional. However, as such, they are both vulnerable to particular criticisms, analogous to those Kripke makes of classical descriptivism.[42] In defending his rigid designation thesis, Kripke argues that certain coherent patterns of modal reasoning are fundamentally incompatible with a treatment (per classical descriptivism) of ordinary proper names as disguised definite descriptions. For example, we may coherently reason (modally) about Aristotle not writing the *Nicomachean*

[40] Lewis: 40.

[41] Lewis relies here on a notion of trans-world identity for stories — that one and the same story should be told at different worlds as fiction and as known fact. As he admits, he does not give clear criteria for this identity.

[42] Kripke, Naming and Necessity.

Ethics, or about Gödel not authoring the incompleteness theorems. However, if the reference of 'Aristotle' were fixed (partly) by the description 'the author of the *Nicomachean Ethics*', it would follow that the sentence:

(3) Aristotle wrote the *Nicomachean Ethics.*

is a necessary truth: true at all possible worlds. This is plainly inadmissible.

An analogous objection is available here, against satisfactional theories of fictional names: if we let the reference of a fictional name be fixed by a description, then intuitively coherent patterns of "counter-fictional reasoning," as Friend calls it,[43] come out as automatically false.

For example, we may hypothesise about what might have happened, had Hamlet attempted to resolve his feud with Claudius by less clandestine means;[44] we conjecture as to whether Frodo Baggins making more liberal use of the Great Eagles would have expedited his journey to Mordor, or whether it would have disclosed his mission to Sauron.[45] These conjectures are perfectly coherent. In fact, they are a key part of how we engage with fiction. Students in literature classes are not only expected to answer simple comprehension questions, but to be able to engage in genuine discussion about fictional *might-have-beens*.

However, if we hold, along the descriptivist line, that the references of fictional names are fixed by the set of descriptions given of the character in the fiction, then sentences like:

(4) Hamlet's feud drives him to insanity.

should be taken as necessary truths — since, by definition, there can be no possible worlds where the referent of the name 'Hamlet' is not driven to insanity.

By a similar token, if we follow the Lewisian hybrid line, and hold that, at any world w, 'Hamlet' refers to an individual x if and only if *Hamlet* is truly told about x at w, then we should (plausibly) also take (4) as a necessary truth — since *Hamlet* cannot be truly told of an individual who is not driven to insanity.

Seemingly, then, such satisfactional theories of fictional names are fundamentally at odds with our ability to engage in coherent and significant counter-fictional reasoning. Resolutions to this tension may be available, but it is no easy task.

[43] Friend: 189.

[44] James D. Carney, "Fictional Names," *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 32, no. 4 (1977): 384.

[45] Tom Schoonen and Franz Berto, "Reasoning About Fiction," (Preprint version), 2018, https://tomschoonen.com/content/1-research/schoonen-berto-2018-reasoning-about-fiction.pdf.

7 Conclusion

The commitments deriving from the Lewisian analysis of fiction-operators force us to consider the issue of the reference-fixing conditions for the name 'Hamlet'. We dismissed the possibility of treating 'Hamlet' on the model of ordinary proper names, holding instead that fictional names must be non-rigid designators, in order to cohere with our modal analysis of fiction-operators. We considered a possible descriptivist solution, taking the reference of 'Hamlet' to be fixed satisfactionally by the descriptions given of Hamlet in *Hamlet*. However, we saw that taking such a line would lead to a potentially unpalatable conclusion. In light of this, we considered a hybrid of the previous two positions: an alternative which I think best represents the position Lewis takes in "Truth in Fiction". However, we raised concerns that this hybrid might be less a best-of-both option than a confused amalgam.

We also noted a further problem affecting both the descriptivist approach and the Lewisian hybrid view, on the model of Kripke's criticisms of descriptivism: namely, that these approaches seem fundamentally at odds with our ability to engage in seemingly coherent counter-fictional reasoning. How serious is this problem? I leave this question open. The reader may decide, given the balance of commitments required to pursue this anti-realist line, whether this line is still preferable to the realist alternative.

Bibliography

Adams, Fred, Fuller, Gary, and Stecker, Robert. "The Semantics of Fictional Names." *Pacific Philosophical Quarterly* 78 (1997): 128-148.

Badura, Christopher and Berto, Francesco. "Truth in Fiction, Impossible Worlds, and Belief Revision." *Australasian Journal of Philosophy* 97, no. 1 (2019): 178-193.

Byrne, Alex. "Truth in fiction: The story continued." *Australasian Journal of Philosophy* 71, no. 1 (1993): 24-35.

Carney, James D. "Fictional Names." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 32, no. 4 (1977): 383-391.

Currie, Gregory. "Fictional Truth." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 50, no. 2 (1986): 195-212.

———. The Nature of Fiction. Cambridge: Cambridge University Press, 1990.

Frege, Gottlob. "Über Sinn und Bedeutung." Zeitschrift für Philosophie und philosophische Kritik 100 (1892): 25–50. Translated as "On Sense and Reference." In *Translations from the Philosophical Writings of Gottlob Frege,* 56-78. Edited and translated by Max Black and Peter Geach. Oxford: Blackwell, 1980.

Friend, Stacie. “The great beetle debate: a study in imagining with names.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 153, no. 2 (2011): 183-211.

Kaplan, David. “Bob and Carol and Ted and Alice.” In *Approaches to Natural Language,* 490-519. Edited by Jaakko Hintikka. Dordrecht: Reidel, 1973.

Kripke, Saul. *Naming and Necessity*. Hoboken: Wiley-Blackwell, 1981.

———. *Reference and Existence: The John Locke Lectures*. Oxford: Oxford University Press, 2013.

Kroon, Fred and Voltolini, Alberto. “Fictional Entities.” *The Stanford Encyclopedia of Philosophy*. 2018. https://plato.stanford.edu/archives/win2018/entries/fictional-entities.

Lewis, David. “Counterpart Theory and Quantified Modal Logic.” *The Journal of Philosophy* 65, no. 5 (1968): 113-126.

———. “Truth in Fiction.” American Philosophical Quarterly 15, no. 1 (1978): 37-46.

———. “Anselm and Actuality.” In *Philosophical Papers* Volume I, 10-25. Oxford: Oxford University Press, 1983.

———. *Counterfactuals.* Hoboken: Wiley-Blackwell, 2001.

Martin, Robert M. and Schotch, Peter K. “The Meaning of Fictional Names.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 26, no. 5 (1974): 377-388.

Parsons, Terence. “Fregean theories of fictional objects.” *Topoi* 1 (1982): 81-87.

Phillips, John F. “Truth and Inference in Fiction.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 94, no. 3 (1999): 273-293.

Priest, Graham. “Sylvan’s Box: a Short Story and Ten Morals.” *Notre Dame journal of Formal Logic* 38, no. 4 (1997): 573-582.

Proudfoot, Diane. “Possible Worlds Semantics and Fiction.” *Journal of Philosophical Logic* 35 (2006): 9-40.

Salmon, Nathan. “Nonexistence.” *Noûs* 32, no. 3 (1998): 277–319.

Schoonen, T. and Berto, F. Preprint. "Reasoning About Fiction.” Amsterdam: University of Amsterdam.
https://tomschoonen.com/content/1-research/schoonen-berto-2018-reasoning-about-fiction.pdf.

Searle, John. “The Logical Status of Fictional Discourse.” *New Literary History* 6, no. 2 (1975): 319-332.

Stokke, Andreas. “Fictional Names and Individual Concepts.” *Synthese* (2020): 1-31.

Thomasson, Amie. “The Reference of Fictional Names.” *Kriterion* 6 (1993): 3-12.

———. “Fiction, Modality and Dependent Abstracta.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 84, no. 2 (1996): 295-320.

———. *Fiction and Metaphysics.* Cambridge: Cambridge University Press, 1999.

———. “Speaking of Fictional Characters.” *Dialectica* 57, no. 2 (2003): 205-223.

Van Inwagen, Peter. “Creatures of Fiction.” *American Philosophical Quarterly* 14, no. 4 (1977): 299–308.

Zalta, Edward N. *Abstract Objects: An Introduction to Axiomatic Metaphysics*. Dordrecht: Reidel, 1983.

“Deepfakes” and the End of the Photographic Age

Patrick Pan

Wolfson College, University of Cambridge

A recent innovation in artificial intelligence, known as Generative Adversarial Networks (GANs), has enabled computers to generate images that are visually indistinguishable from photographs. GAN images, sometimes called “deepfakes”[1], have already been recognized to pose an epistemic threat to society by undermining the capacity of photographs to provide evidence. In this paper, I will investigate both the epistemic status and the potential aesthetic value of GAN images, as well as how the proliferation of GAN images will affect the epistemic and aesthetic value of true photographs. I will affirm the view that GAN images are a potential epistemic threat, but also argue that they are nevertheless a medium with significant potential for artistic expression. To do so, I will draw upon Dawn Wilson’s argument that photographs are ontologically dual and can be considered as both mind-independent “photo-images” and mind-dependent “photo-pictures”. I will extend that argument to GAN images to show that, while they are indeed the outputs of mind-independent computer algorithms that do not provide information about real objects in the world, they can also be skilfully generated in a way that can embody artistic intentions. Consequently, I will argue that if GAN images and photographs become indistinguishable, then photographs will come to occupy a role in society similar to that of paintings today, in that they will lose their epistemic authority but continue to be valued aesthetically.

1 Introduction

The goal of this paper is to investigate the philosophical implications of Generative Adversarial Networks (GANs), an innovation in artificial intelligence that enables computers to generate images that are visually indistinguishable from photographs. GAN images are known as “deepfakes” in the popular press, which has already recognized them to have immense social and political implications. However, little has been said with regards to their potential aesthetic value, or their potential impact on the aesthetic status of true photographs. I will aim to answer two questions. First, what are the epistemic and aesthetic statuses of GAN images? Second, how will the epistemic and aesthetic statuses of true photographs change if they become fully indistinguishable from GAN images?

In Section 2, I will analyse the contemporary debate around the epistemic and aesthetic value of photographs to provide a basis for similar arguments regarding GAN images. In that section, I will introduce two sceptical arguments that discount the aesthetic value of photographs and introduce Wilson’s ontology of photographs to reject those sceptical arguments. In Section 3, I provide a substantive account of the GAN image process

[1] The term “deepfake” is often to refer to images (or videos) of people created for deceptive or salacious purposes. In this paper I will use the term ‘GAN image’ to refer more broadly to all photorealistic images generated by GANs, regardless of their purpose or content.

in order to ground an epistemic and aesthetic analysis in the following two sections. In Section 4, I will argue that GAN images indeed pose a threat to the epistemic value of photographs. In Section 5, I will attempt to adapt the sceptical arguments introduced in Section 2 to target GAN images and utilize a modified form of Wilson's arguments to demonstrate that GAN images have a potential for artistic expression on par to that of photographs. I conclude in Section 6 that the impending shift in the role of photography due to the proliferation of GANs is not unprecedented, as it will likely be similar to the effect that the invention of photography itself had upon the previously dominant artform of painting.

2 The Duality of Photographs

Photographs are commonly thought to occupy a dual role in society. As epistemic tools, photographs have an unparalleled ability to reliably record what some objects in the world looked like at some time. Simultaneously, photography is one of the most widely used aesthetic media in the world. Photographs often fulfil either of these two roles: this is evident through a glance at the average smartphone camera roll, which might contain pictures of important documents (epistemic tool) as well as colourful sunsets (artistic expression). Crucially, some photographs fulfil *both* roles. For instance, a portrait of a former pet may be valued not only for its ability to serve as a reminder of how the pet once looked, but also for its aesthetic properties.

However, examining the assumptions that underlie this common view illuminates a potential tension between these two roles. Photographs are usually considered to be epistemically valuable because they are "mind-independent": the appearance of a photograph is causally determined by the state of the physical world, independent of any agent's intentions. On this view, photographs are more valuable epistemic tools than paintings or drawings, which are produced through the intentions of an artist. On the other hand, if we accept the widespread[2] claim that art must be intentionally produced (for instance, as formulated by Nick Zangwill[3]), the possibility that photography can be art implies that photographs must possess some degree of "mind-dependence". The dual epistemic and aesthetic roles of photography appear to pose the contradiction that photographs simultaneously possess mind-independence and mind-dependence.

2.1 Arguments Against Photographic Duality

Some philosophers, such as Roger Scruton, sidestep this apparent contradiction by simply denying that photographs have a dual nature. Given that the causal dependence of photographs on certain physical features of the world is easily observable, it seems much plausible to dispense with mind-dependence and claim that photographs are solely mind-independent. This singular thesis of mind-independence can be utilized in multiple dis-

[2] Paisley Livingston, *Art and Intention: A Philosophical Study* (Oxford: Oxford University Press, 2005), 35–40.

[3] Nick Zangwill, "The Creative Theory of Art," *American Philosophical Quarterly*, no. 32 (1995): 307–23.

tinct sceptical arguments purporting to show the impossibility of aesthetic expression in photographs.

One such sceptical argument posits that photographs too closely resemble reality to deserve any aesthetic interest for their own sake. On this view, which I call "transparency scepticism", photographs are, by nature, transparent depictions of reality and deserve no aesthetic interest in themselves. The transparent relationship between photographs and what they depict thus simultaneously justifies photography's epistemic value and dooms its aesthetic value. The most provocative such argument comes from Roger Scruton, who controversially claims that "with an ideal photograph it is neither necessary nor even possible that the photographer's intention should enter as a serious factor in determining how the picture is seen".[4] According to Scruton, the ideal photograph is "recognized at once for what it is — not as an *interpretation* of reality but as a *presentation* of how something looked".[5] Scruton's ideal photograph is the epitome of an epistemic tool, as it enables one to obtain reliable information about the appearance of an object. However, because the nature of the ideal photograph is to faithfully reproduce reality, there is no need to consider the intentions of the photographer. On Scruton's transparency sceptic view, ideal photographs are entirely mind-independent and are merely "a means to the end of seeing its subject".[6]

A second argument, which I call "mechanical scepticism", posits that the creation of a photograph is dominated by mechanical and non-human processes that interfere with the transmission of a human intention. Mechanical scepticism dates as far back as 1865 in a highly dogmatic form claiming that photography lacks "something beyond mere mechanism" and cannot be "the expression of man's delight in God's work".[7] More recently, it has been utilized in a less extreme form by Nigel Warburton, who argues that photographers must personally certify that their photographs actually fulfil their artistic intentions in order to guarantee their aesthetic value. Warburton broadly defines the process of certification as a "conferral of status" and argues that means of certification include "signing, stamping, exhibiting, [or] printing the image in a certain context"[8]. According to Warburton, such certification is the only way "in which photographers overcome the expressive limitations of a process that is largely automated",[9] and photographs lacking certification "can never be reliable indicators of a photographer's intentions".[10] Warburton's mechanical scepticism characterizes photography as a medium whose potential artistic value is undermined by the fact that it contains mind-independent automatic processes, and Warburton argues that this issue can only be mitigated through an additional mind-dependent mark.

Transparency scepticism and mechanical scepticism proceed from the same assumption about the photographic process: the claim that photographs are created through a mechanical, mind-independent process. However, they differ crucially in that the former targets the *content* of a photograph, and that the latter targets the *causal* history of

[4] Roger Scruton, "Photography and Representation," *Critical Inquiry* 7, no. 3 (1981): 588.
[5] Scruton, "Photography and Representation", 588. Italics mine.
[6] Scruton, "Photography and Representation," 590.
[7] 7 "Art and Photography," *The New Path* 2, no. 12 (1865): 198–199.
[8] Nigel Warburton, "Authentic Photographs," *The British Journal of Aesthetics* 37, no. 2 (1997): 134.
[9] Warburton, "Authentic Photographs," 135.
[10] Warburton, "Authentic Photographs," 135.

a photograph. A sufficient rebuttal of the idea that photographs are intrinsically mind-independent would weaken both transparency and mechanical scepticism, as well as other views that deny the ability of photographs to earn our aesthetic interest.

2.2 Photo-Image and Photo-Picture

Dawn Wilson (publishing as Phillips) poses a solution to the apparent contradiction by proposing that photographs have a dual ontology and exist as both "photo-image" and "photo-picture". According to Wilson, the photo-image refers to the visual appearance of a photograph, and "visual properties of the photo-image supervene on… those properties caused by the photographic event [and] material production process".[11] The photo-image is the photograph considered in a purely material sense: its properties are causally determined by mechanical, mind-independent objects and processes. By contrast, a photo-picture "has intentional content as [a product] of human design",[12] and the "properties of the photo-picture also supervene on the intentions of the artist".[13] According to Wilson, skilful manipulation of the photographic process allows a photographer to create a photograph that fulfils the purpose of a photo-picture. Wilson writes:

> The skilled photographer can form intentions to create a visual image that will have particular properties. The photographer is not simply at the mercy of the photographic process; but instead uses photographed objects, along with the camera apparatus, in accordance with a skilled understanding of the photographic process, to create photo-images that have those particular visual properties. In this way a photograph can fulfil the intentions of a photographer as much as a painting can fulfil the intentions of a painter.[14]

Wilson's characterization of skilful photography poses an effective rebuttal to both transparency and mechanical scepticisms. In the case of transparency scepticism, Wilson concedes that "when we take an interest in a *photo-image,* we may be concerned with [the photographed] objects".[15] This allows that Scruton's argument to be correct only if we consider the photograph *qua* photo-image. Yet by viewing the process of producing a photograph *qua* photo-picture, i.e., through the lens of a skilled photographer's intentional actions surrounding the actual photographic event, Wilson brings to light that an artist's intention, far from being invisible in the causal history of a photograph, in fact leaves an indelible mark upon it. All this is to suggest that there is in fact capacity for an artist's intention, and therefore aesthetic value, in a photograph.

Meanwhile, in the case of mechanical scepticism, Wilson's appreciation of photographers' skilful manipulation of the photographic process suggests that the sophisticated mechanical processes involved in photography simply make available a greater set

[11] Dawn M. Phillips, "Fixing the Image: Rethinking the 'Mind-Independence' of Photographs," *Postgraduate Journal of Aesthetics* 6, no. 2 (2009): 13.

[12] Phillips, "Fixing the Image," 5.

[13] Phillips, "Fixing the Image," 20.

[14] Phillips, "Fixing the Image," 18.

[15] Phillips, "Fixing the Image," 19. Italics mine.

of tools with which photographers can realize their intentions. Historically, technological advances have broadened the expressive range of photographs. For instance, prior to the invention of colour photography, photographers had only light and dark to establish contrast in a photograph, but today they can also utilize warm and cool colours to effect further contrast. By understanding the photographic process as a means to the end of an artist, Wilson dismisses Warburton's concern that automatism limits creative expression. On the contrary, mechanical sophistication only empowers an artist's expression.

Wilson's description of the photographic process is particularly compelling because it permits photography to fulfil both epistemic and aesthetic roles without compromise. Instead of defining mind-independence and mind-dependence as directly opposed, Wilson conceives of the mind-independence of a photo-image and the mind-dependence of a photo-picture as independent traits. I believe that Wilson's account of photography satisfactorily justifies the dual roles of photography, and that it is useful for analysing photographs in terms of their epistemic and aesthetic value. To determine how Wilson's argument relates to GAN images, we must now investigate the process by which GAN images are created.

3 The GAN Image Process

In this section, I will briefly describe the process by which a GAN is used to generate an image. While the GAN image process involves a great deal of algorithmic computation, it also requires a significant amount of human operation that centres on three factors. The first is the selection of the "training set", a database of images which the GAN learns to imitate. The second is the choice of a "seed number", which the GAN mathematically transforms into an image resembling one from the training set. The third is the final step of curating images produced by the GAN.

A GAN consists of a competing pair of algorithms, a "generator" and a "discriminator". The generator is tasked with turning random numbers called "seeds" into images that resemble those from the training set. The discriminator then receives a mix of images created by the generator and images from the training set and attempts to determine the origin of each image. Both algorithms receive the discriminator's results and use them to improve through trial and error: the generator learns to operate more closely to when it succeeded in deceiving the discriminator, while the discriminator learns from its mistakes. Over time, the generator learns to turn any seed number into an image that the discriminator would guess is part of the training set.[16]

Notably, if the training set depicts one kind of object, then the GAN will produce images that appear to depict the same kind of object. For instance, a GAN trained with photographs of human faces will generate images resembling human faces. As proof of the photorealism of GAN images, consider the two faces in figure 1 and attempt to determine which was computer-generated.

[16] "Overview of GAN Structure |Generative Adversarial Networks," Google Developers, accessed Jan 4, 2020, developers.google.com/machine-learning/gan/gan_structure.

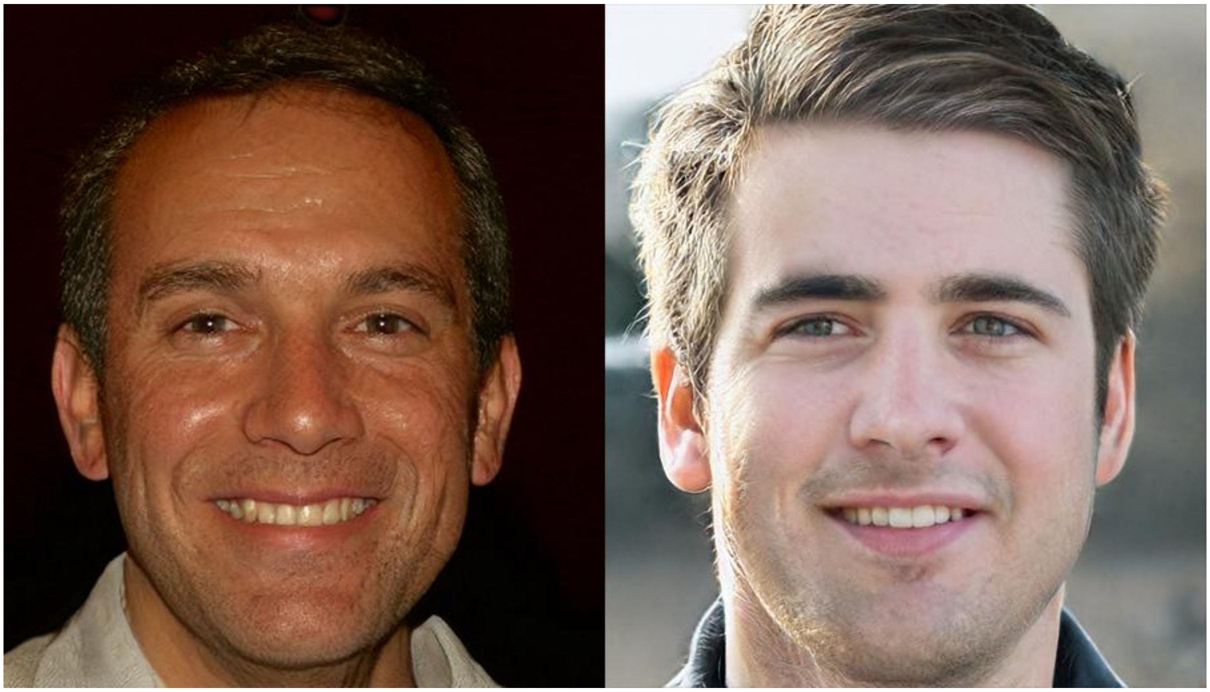

Figure 1: One face was generated by the GAN powering thispersondoesnotexist.com[17]

After training, the generator can be provided an arbitrary seed number to convert into an image. Using other AI techniques, it is possible to select a seed that further influences the appearance of the output. One particularly interesting application turns linguistic descriptions into seeds that generate images matching the description. Figure 2 contains images generated from descriptions of birds by various algorithms developed from 2016 to 2018 (lower rows show newer results).[18] While the results of even the latest algorithms are not quite as realistic as the above faces, it is likely that the quality of these images will continue to increase in time to become equally photorealistic. These results illustrate that the seed number, which can be deliberately selected, has a strong influence upon the appearance of the generated image.

In total, the GAN image process offers three opportunities to manipulate the appearance of the final image. First, the selection of an appropriate training set circumscribes the possible appearance of all generated images. Second, the choice of a seed number allows one to influence the appearance of the image within the boundaries set by the training set. Finally, the curation of GAN-generated images allows selection of the results that best fit certain criteria, which can range from simple photorealism, to correspondence with a description, to even specific visual qualities that fulfil an artistic intention.

[18] 18 Han Zhang et al., "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, no. 8 (2019): 1954, fig. 3.

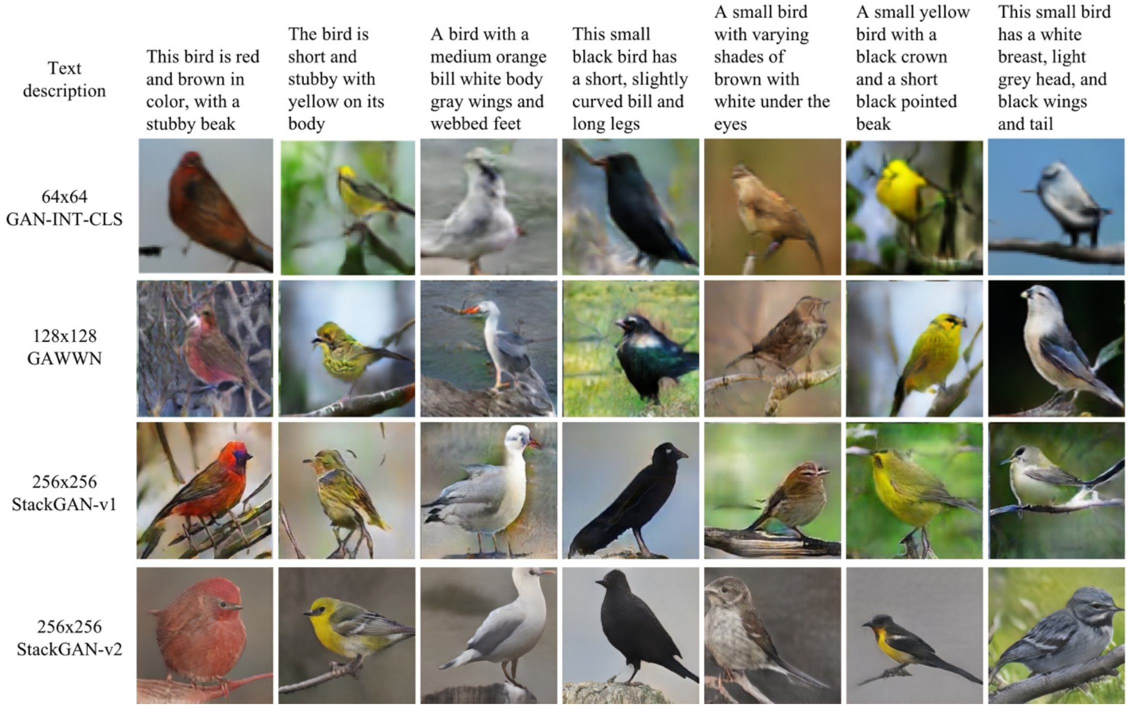
Text description
This bird is red and brown in color, with a stubby beak
The bird is short and stubby with yellow on its body
A bird with a medium orange bill white body gray wings and webbed feet
This small black bird has a short, slightly curved bill and long legs
A small bird with varying shades of brown with white under the eyes
A small yellow bird with a black crown and a short black pointed beak
This small bird has a white breast, light grey head, and black wings and tail
64x64 GAN-INT-CLS
128x128 GAWWN
256x256 StackGAN-v1
256x256 StackGAN-v2

Figure 2: Results from various GANs trained to generate images of birds according to image descriptions. Reproduced from Han Zhang et al. with permission.

4 The Epistemic Value of Generated Images

4.1 Probabilistic Visual Information

As discussed in Section 2, photographs are commonly understood to be epistemically valuable in virtue of their mind-independence. Here, the term "mind-independence" refers specifically to a mind-independent counterfactual relation to the portrayed object. For instance, if a photograph is taken of a person's face, then a change in the person' facial expression necessitates a like change in the photograph — the photograph will show a smile if and only if the person is smiling. This dependence is invariant with respect to any mind, as the photograph will show a smile regardless of the mental state of the photographer or viewer. Indeed, this relation persists even if the photograph is taken by a computer-controlled security camera. However, this idea about mind-independence is merely an intuition, not a rigorous epistemic argument.

In a pair of papers,[19] [20] Cohen and Meskin analyse the process by which we purport to gain knowledge from photographs. On their account, photographs are not sources of knowledge (defined as justified true belief), but of "information", which Cohen and Me-

[19] Jonathan Cohen and Aaron Meskin, "On the Epistemic Value of Photographs," *The Journal of Aesthetics and Art Criticism* 62, no. 2, (2004): 197–210.

[20] Aaron Meskin and Jonathan Cohen, "Photographs as Evidence," in *Photography and Philosophy*, ed. Scott Walden (Malden: Blackwell 2008) 28 Jan. 2009: 70–90.

skin define as a "probabilistic, counterfactual-supporting, connection between independent variables".[21] Their definition of information is two-pronged: A carries information about B only if, firstly, A is likely to be very close in value to B (probabilistic connection), and secondly, a change in B will result in a change in A (counterfactual). Having defined information, Cohen and Meskin assert that photographs "typically provide information about many of the visually detectable properties of the objects they depict".[22] Thus, if a photograph carries information, then its appearance resembles the appearance of the object it depicts, and any change in the depictum causes a corresponding change in the photograph. Additionally, Cohen and Meskin clarify that "informational links are constituted independently of any subject's beliefs or mental states"[23], which introduces the condition of mind-independence. This claim is similar to the mind-independence thesis with which we are already familiar, but it differs significantly in that it operates with regards to *information,* not knowledge or belief.

Images generated by a computer do not necessarily have a similar mind-independent counterfactual correlation with the appearance of their depicta. While the images generated by a GAN algorithm resemble those of the training set, this resemblance does not support counterfactuals relating to specific concrete objects. For instance, the images at thispersondoesnotexist.com depict faces that vary in age, gender, skin tone, and other features, yet they do not provide information about any actual person. GAN images cannot carry information as photographs do about their depicta.

4.2 Salient Image Categories and Epistemic Value

Cohen and Meskin concede that the capacity of a medium to carry information is distinct from our *belief* that the medium actually carries information. To underscore the importance of this distinction, Cohen and Meskin propose the example of courtroom illustrations and veridical portrait paintings, which carry information about their subjects because they probabilistically and counterfactually resemble their depicta. Nevertheless, our beliefs about them differ, and "we do not accord the same epistemic status to realistic portrait paintings as we accord to photographs".[24]

To explain our *beliefs* that photographs carry information, Cohen and Meskin examine the background social practices involved when we interpret images. According to Cohen and Meskin, when we encounter a token photograph, we "typically categorize that token as an instance of the type of photographs" and deem it epistemically reliable.[25] By contrast, when we encounter even the most realistic portrait painting, we "typically do not categorize that token as an instance of the type of veridical portrait paintings" but instead "an instance of the type of portrait paintings" and deem it epistemically unreliable.[26] The difference between these examples consists in that the "type [of photographs]

[21] Cohen and Meskin, "Epistemic Value," 7.
[22] Meskin and Cohen, "Evidence," 3.
[23] Meskin and Cohen, "Evidence," 3.
[24] Cohen and Meskin, "Epistemic Value", 15.
[25] Cohen and Meskin, "Epistemic Value", 16.
[26] Cohen and Meskin, "Epistemic Value", 16.

is salient for subjects in a sense that these other types [portrait paintings] are not".[27] Most of us believe that photographs, paintings, and drawings are salient types of images — only the first of which is a reliable carrier of information — but that veridical portrait paintings and courtroom drawings are not salient types, despite the information they actually carry. Cohen and Meskin remark that "both the saliency ordering among representational types and the generally-held background beliefs about these types are, presumably contingent".[28] According to this view, if our beliefs change and the salience of photographs as a type of image diminishes, then it is possible that we would judge token photographs to fall under a salient type of image that does not reliably carry visual information and photographs would lose their epistemic value.

If GAN images become so photorealistic as to become indistinguishable from true photographs, then true photographs will cease to be a salient category of image. Just as veridical portrait paintings and non-veridical portrait paintings both fall under the salient category of portrait paintings, GAN images and digital photographs would likely both fall under the category of the "photorealistic image" as a whole. Because the unified category of photorealistic image contains both information-carrying photographs and information-devoid GAN images, the category cannot be said to reliably carry information. Although photographs would retain their ability to carry information — there could certainly be no change to the content of pre-existing photos — on the level of belief, they would be judged to be simply "digital images" with no epistemic value. Thus, the proliferation of photorealistic GAN images may entirely strip photographs of their epistemic status.

5 The Aesthetic Value of GAN Images

5.1 Aesthetic Scepticism and Technologial Advancements

Recall from Section 2 the arguments I termed transparency scepticism and mechanical scepticism. Despite their archaic roots, the two sceptical arguments might seem to be increasingly justified by historical and modern advances in photographic technology. In the case of transparency scepticism, one might claim that the argument has been strengthened by improvements in the resolution and colour fidelity of digital cameras. If the argument of transparency scepticism applies to a scratchy, blurry, black-and-white photograph that significantly distorts reality, then it most definitely applies to a 24-megapixel image from a cutting-edge digital camera.

In a similar vein, the argument of mechanical scepticism is also strengthened by technological advancements. Pressing the shutter button on an iPhone is far more "automatic" than spending twenty minutes exposing a delicate daguerreotype plate to light and curing it with noxious chemicals. A mechanical sceptic might therefore argue that the human element in photography has continuously diminished over time to a comparatively

[27] Cohen and Meskin, "Epistemic Value", 16.
[28] Cohen and Meskin, "Epistemic Value", 19.

infinitesimal amount in the present. would follow that the aesthetic value of photography has decreased in parallel.

Given that both sceptical arguments discussed are strengthened by recent developments in photography, it is possible that the arguments will also apply convincingly to GAN images, which represent yet another step forward in image technology. If these arguments do indeed successfully diminish the artistic value of GAN images, and GAN images become indistinguishable from genuine photographs, then it is possible that all digital images, genuine photographs included, will be relegated to an inferior aesthetic status.

5.2 Are GAN Images Mind-Independent?

If GAN images are entirely mind-independent, then the sceptical arguments against the aesthetic value of photography may also apply equally to GAN images. In Section 4, we showed that GAN images have a mind-independent relation to their training set and seed. However, as we have seen in Wilson's analysis of photography, this is not to say that GAN images are *entirely* mind-independent. Wilson justifies her view that a photograph *qua* photo-picture is mind-dependent by characterizing photography as a process that involves intentional decisions made by the skilled photographer throughout the photographic process.

The same can be said for GAN images. As described in Section 3, during the GAN image process the image creator makes three choices: the selection of a training set (general appearance), the choice of a seed (specific appearance), and the curation of output images. These three opportunities for deliberate manipulation of the GAN image process have analogous opportunities in the photographic process. The selection of the training set, which determines the general appearance of generated images, is similar to a photographer's selection of a scene or object to photograph. Just as the choice of a training set consisting of images of human faces will cause the final image to resemble a human face, a photographer's choice to photograph a person will cause their resulting photograph to depict a person. Next, the seed, which determines the appearance of the image within the boundaries of the training set, is analogous to a photographer's selection of a vantage point and camera settings, such as focal length, shutter speed, and exposure. While the seed for a non-intentional GAN image can be random, so too can a non-intentional photograph be taken on automatic settings from an arbitrary vantage point. But skilled photographers control these factors to influence the appearance of their final image, and I argue that creators of GAN images can do the same by selecting a seed. Finally, the curation of GAN images is identical to a photographer's curation of their best work from a session — a wedding photographer may capture a thousand photographs but only determine a few dozen to be satisfactory. If we accept Wilson's argument that the photographic process is sufficiently mind-dependent to fulfil artistic intentions, then these parallels with the GAN image process suggest the same about the latter.

5.3 Dispelling Aesthetic Scepticism Regarding GAN Images

Let us consider the arguments of transparency and mechanical scepticism to determine whether they are applicable to GAN images, and then attempt to adapt Wilson's stance to rebut them if they do. Transparency scepticism, especially in Scruton's formulation, assumes the premise that, in their "ideal form", photographs are reproductions, not interpretations, of reality. In the terminology of Section 4, photographs are meant to carry visual information. Even if we accept this essentialist view of photographs and grant that the information-carrying "ideal form" of photography precludes artistic expression, it is also clear that this argument simply does not apply to GAN images. GAN images, as shown in Section 4, do not carry visual information and do not resemble any existing object. As such, transparency scepticism falls flat against GAN images and provides no basis for denying their potential aesthetic value.

Having dispensed with transparency scepticism against GAN images, let us now turn to mechanical scepticism. The mechanical sceptic argument against photography holds that the presence of mechanistic processes in photography reduces the ability of photographs to reliably transmit the original intentions of the photographer. This argument applies GAN images with little to no modification — if anything, GAN images can be said to use even more automated processes than does photography. However, Wilson's objection to mechanical scepticism successfully defends the aesthetic value of GAN images with little to no modification as well. As a mechanical sceptic, Warburton might claim that the creator of a GAN image leaves even less of a trace upon their work than a photographer does, and that they must further compensate for that loss of agency through some personal certification. However, as inspired by Wilson's argument, so long as the GAN image process still allows an image creator to determine the content of the resulting GAN image in accordance with an artistic intention, the resulting image, qua picture, can still transmit that intention and bear artistic value. Wilson's view states in short that the various image-making processes—painting, photography, even GANs — exist to help artists realize their intentions, and that artistic intentions are fulfilled through those processes, not limited by them.

Thus, scepticism about the aesthetic value of photography is equally as inapplicable to genuine photographs as it is to GAN images. Just as photography can be mastered by a skilled photographer to produce images of a desired visual appearance that bears an artistic intention, so too can the GAN image process allow for a skilled image creator to fulfil their artistic intentions through the skilful use of that process. Wilson's characterization of photographs as a duality of photo-image and photo-picture can be applied to GAN images; it is possible for a skilled individual to create not just a GAN image, but a GAN *picture*.

GAN images and photographs are two very different kinds of images in terms of their production, but they both are deserving of aesthetic interest in their own unique manner. While other arguments that GAN images are intrinsically aesthetically inferior to other visual media may exist, I argue that the most widely held such views will rely on the same flawed thesis of mind-independence that is used to justify similar arguments against photographs. Given that in the art world photographs are widely believed to be a full-fledged visual artform comparable to painting, I believe that, barring the emergence of totally novel objections, GAN images may soon join photographs and paintings on the walls of the art gallery.

6 Implications: The End of the Photographic Age

In 1958, film theorist André Bazin called the invention of photography "clearly the most important event in the history of the plastic arts".[29] On Bazin's account of the evolution of visual art, prior to the invention of photography "painting was torn between two ambitions: one, primarily aesthetic... the other... to duplicate the world outside". In other words, paintings once occupied the dual epistemic and aesthetic role that photography fulfils today.[30] Bazin claims that the invention of photography "freed Western painting, once and for all, from its obsession with realism".[31] According to Bazin's chronology of painting, the period between the development of perspective — the "original sin of Western painting"[32] — and the invention of photography was a dark age in which painters were torn between epistemic and aesthetic commitments. Afterwards, painters abandoned their imagined obligation to realism, ushering in in a golden age of creativity.

Just as Bazin uses the invention of photography to delineate the "dark age" and "golden age" of painting, the proliferation of GAN images presents another boundary, in this case between two ages of photography. Taking inspiration from Bazin, I propose to call the period of time in which photographs occupied their dual epistemic-aesthetic role in society the "Photographic Age". The Photographic Age began with the invention of photography, and, as I have argued in this paper, it may end with the invention of the GAN image.

As we stand in the last days of the Photographic Age, there remain two practical problems to solve, each corresponding to one of the aspects of photography. First, the loss of photography as a ubiquitous and trusted epistemic tool indeed poses a threat to be mitigated, whether through an effort to prevent the perceptual merging of the categories of photographs and GAN images, or through an attempt to inform the public about the impending crisis of epistemic unreliability. Second, for the sake of artistic innovation, we ought to maintain permissive definitions of aesthetic value such that the GAN image process and photography can both be used to their full potential, rather than adopt dogmatic and exclusionary definitions of art that stymie the creation of new works. Photography has long been a special source of information and a cherished artistic medium, but our best hope for its future in the face of technological advances is to enjoy the new artistic possibilities created by artificial intelligence while avoiding the threats that it poses.

[29] Andre Bazin and Hugh Gray, "The Ontology of the Photographic Image," *Film Quarterly* 13, no. 4 (1960): 9.

[30] Bazin and Grey, "Ontology", 9.

[31] Bazin and Grey, "Ontology", 6.

[32] Bazin and Grey, "Ontology", 7.

Bibliography

“Art and Photography.” *The New Path* 2, no. 12 (1865): 198–99.

Bazin, Andre, and Hugh Gray. “The Ontology of the Photographic Image.” *Film Quarterly* 13, no. 4 (1960): 4–9.

Cohen, Jonathan, and Aaron Meskin. “On the Epistemic Value of Photographs.” *The Journal of Aesthetics and Art Criticism* 62, no. 2 (2004): 197–210.

Google Developers. “Overview of GAN Structure |Generative Adversarial Networks.” Accessed January 4, 2020. https://developers.google.com/machine-learning/gan/gan_structure.

Meskin, Aaron, and Jonathan Cohen. “Photographs as Evidence.” *Photography and Philosophy*, January 28, 2009, 70–90.

Livingston, Paisley. *Art and Intention: A Philosophical Study*. Oxford: Oxford University Press, 2005.

Phillips, Dawn M. “Fixing the Image: Rethinking the ’Mind-Independence’ of Photographs.” *Postgraduate Journal of Aesthetics*, 2009, 6, no. 2 (August 209AD).

Scruton, Roger. “Photography and Representation.” *Critical Inquiry* 7, no. 3 (1981): 577–603.

Warburton, Nigel. “Authentic Photographs.” *The British Journal of Aesthetics* 37, no. 2 (1997): 129–37.

Yang, Yingzhi, and Brenda Goh. “China Seeks to Root out Fake News and Deepfakes with New Online Content Rules.” Edited by Edwina Gibbs. Reuters. Thomson Reuters, November 29, 2019. https://www.reuters.com/article/us-china-technology/china-seeks-to-root-out-fake-news-and-deepfakes-with-new-online-content-rules-idUSKBN1Y30VU.

Zangwill, Nick. "The Creative Theory of Art." *American Philosophical Quarterly* 32, no. 4 (1995): 307-23. http://www.jstor.org/stable/20009834.

Zhang, Han, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. “StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, no. 8 (2019): 1947–62.

On the Difference Between Episodic and Autobiographical Memories

Gabriel Zaccaro

Universidade Federal de Santa Maria

Is there a difference between recollecting episodes from the past and recalling autobiographically? Both in the philosophical and psychological literature, it does not seem that there is a consensus on whether autobiographical memories should be considered as a metaphysically equivalent concept to episodic memories or a different category of memory entirely. In this article, I give reasons to believe that autobiographical memories do not relate to the recollection of past episodes since they do not have an associated subjective experience and the consequential experience of mental time travel. Autobiographical memories, I argue, are presented as a narrative that is constructed propositionally, thus differing substantially from episodic memory in its subjective property, the reference to the self, and the content in which each one is grounded. To do so, first I use data from the psychological literature on amnesic patients as evidence for both the difference in phenomenology and content. And second, I use insights from recent philosophical literature on memory and the self, to defend that what is referred here as "autobiographical memory" is indeed a different kind of memory that differs substantially from episodic memory and slightly from semantic memory.[1]

1 Introduction

Through remembering, we can re-experience our past and this is particularly important for our own identity. But to what extent does the remembrance of past episodes contribute to the internal narrative of our lives? Let us suppose that we are asked to determine whether we are generally happy in our lives. To do so, would we have to remember all our happy past episodes, as opposed to the sad ones? Or, would our "happiness evaluation" result from a preconceived notion of "general happiness" available to us, based on the narrative of our lives? This question can be formulated differently to fit the philosophical debate: When we recall the narrative of our lives, do we use episodic information, semantic information, or both? What this problem seems to refer to is the doubt in whether we can differentiate clearly between episodic and autobiographical memories. Here, I defend the idea that remembering a life narrative, or autobiographical recalling, does not need to be represented through the remembrance of past episodes but can be a representation based on semantic information available to the individual. Furthermore, I intend to show that episodic and autobiographical memories can be distinguished based on their phenomenology. Although episodic remembering is an experience of mental

[1] I would like to thank César Schirmer dos Santos for his comments on previous versions of this paper, as well as Eduardo Vicentini de Medeiros for some insightful discussions about the topic.

time travel, with the necessary awareness that the episode belongs to the subject's personal past (autonoesis), both mental time travel and autonoesis are not present in autobiographical memory. This claim is based on psychological research focused on the amnesic patient K.C. The studies of K.C.'s case suggest that although he cannot remember past episodes and create new ones, he can, to some extent, present semantic knowledge about his autobiography. This was primarily used by Tulving to identify the properties of the subjective experience of remembering.[2] However, Tulving left aside any conclusions about how this autobiographical semantic information could be used to compose K.C's life narrative. What I propose is that the semantic autobiographical knowledge that K.C. has is sufficient to account for a chronologically ordered life narrative.

The distinction proposed here considers both the type of content that constitutes each type of memory and the phenomenological nature of both memories. Concerning the content of which type of memory, I defend, differing from most of the philosophical literature, that autobiographical memory draws its content from semantic information. Concerning their phenomenology, I argue that while episodic memory has a distinct phenomenology characterized by autonoetic consciousness, autobiographical memory, as is expected from semantic information, is not related to autonoetic consciousness and mental time travel but can be more easily associated with noetic consciousness. To support my thesis, I will show in section 2 the definition of autonoesis and its necessity for episodic remembering. Then, in section 3, I will defend that autonoesis is not required for autobiographical recalling and that, therefore, it cannot consist of an experience of mental time travel. And finally, in section 4, I will defend the view in which autobiographical memory is a separate form of declarative memory, differing substantially from episodic memory and slightly from semantic memory.

The main argument of this article can be exposed as follows:

(1) All episodic memory has an autonoetic property (shown by the philosophical and psychological literature on memory).

(2) Some autobiographical memory does not have an autonoetic property (shown by the studies of the amnesic patient K.C).

(3) If two entities are identical, then they must present identical properties. (Principle of indiscernibility of identicals).

(4) Episodic memory and autobiographical memory differ in the autonoetic property. (1,2)

(5) ∴ Episodic memory is different from autobiographical memory. (3,4)

[2] Tulving, E., "Memory and Consciousness.," *Canadian Psychology/Psychologie Canadienne* 26, no. 1 (January 1985), 5-6.

Let: episodic memory = E, autobiographical memory = A, autonoetic property = T, any property = P.

(1) $\forall(x)[E(x) \rightarrow T(x)]$

(2) $\$(y)[A(y) \wedge \sim T(y)]$

(3) $\forall(x)\forall(y)[x = y \rightarrow \forall P(P(x) \equiv P(y))]$

(4) $\forall(x)\forall(y)[E(x) \wedge A(y)) \rightarrow \forall(T)(T(x) \equiv\sim T(y))]$ (1, 2)

(5) $\therefore \forall(x)\forall(y)[(E(x) \wedge A(y)) \wedge \forall T(T(x) \equiv\sim T(y)) \rightarrow\sim (x = y)]$ (3, 4)

2 The Necessity of Autonoetic Consciousness for Episodic Memory

The use of the terms “episodic memory” and “autobiographical memory” can vary widely within the philosophical and psychological debate. From their conceptual identification to their conceptual distinction, these terms have been used variably, and this can be a problem. And that is because there is no consensus on whether we should use the term “autobiographical” as meaning a mental representation of a narrative nature or an episodic nature. If we intend to widen our understanding of human memory and how it is related to a life narrative, solving this problem is of great importance. So, for the sake of clarification, in this section, I will show the definitions of two main categories of memory, episodic and semantic, and show that autonoetic consciousness is necessary for episodic memory as viewed as an experience of mental time travel. In the next section, I will deal with autobiographical memory, its characterization, and present reasons why it is not related to autonoetic consciousness or mental time travel.

2.1 Some Definitions on Types of Memories and Their Respective Consciousnesses

Episodic memory is defined as a present mental representation of past experiences with perceptual and temporal information that is accompanied by a state of consciousness, the so-called *autonoetic consciousness* or *autonoesis.*[3] Episodic memory relates specifically to past *episodes* from the individual’s life, and its content is *perceptual,* meaning that it is a re-experience of the given past episode. In this re-experience, the subject can relive the episode with a significant amount of sensorial (visual, olfactory, auditory, etc.) quality. To remember episodically is “to consciously re-experience past experiences”.[4] On

[3] Tulving, "Memory and Consciousness.”, 3.

[4] Tulving, E., “Episodic Memory: From Mind to Brain”, *Annual Review of Psychology* 53, no. 1 (February 2002), 6.

the other hand, semantic memory deals with propositional information related to general knowledge of the world.[5] Semantic memory is necessary for language use because it deals with verbal symbols, their meanings, relations among them, and rules for their manipulation.[6]

To put it more clearly, the difference between episodic and semantic memory is that the former is presented perceptually to the subject, and the latter is presented propositionally.[7] In this sense, to remember episodically is to remember what it was like for you to experience your last birthday party, for instance, and to recall semantically is to recall knowledge learned in the past, such as, for example, “2+2=4” or that “Germany is located within Europe”. You could, nonetheless, remember propositions about your birthday party like “I remembered it happened on the 4th of August” but this information is not presented as an experience of reliving this episode. Likewise, you could remember episodically the day that you learned that “2+2=4” but when you remember the propositional information, the experience of the episode does not come to mind.

Episodic memories are a present re-experience of the perceptual information of an event, be it visual, olfactory, auditory, palatal, or tactile. But, along with the experience of perceptual contents, episodic memories are accompanied by a specific type of consciousness, named autonoesis. Autonoesis is defined as the awareness of one’s experiences in a subjective timeline, and being a necessary component of episodic memory, it assures that when the individual remembers episodically, he is aware of the remembered episode’s existence in his past.[8] On the phenomenal aspect of this kind of consciousness, Tulving says that: “The awareness and its feeling-tone are intimately familiar to every normal human being. One seldom mistakes remembering for any other kind of experience — perceiving, imagining, dreaming, daydreaming, or just thinking about things one knows about the world.”.[9] Thereby, autonoesis is important for episodic remembering because (1) it is what allows the subject to be aware of the subjective time in which events happened,[10] and (2) is what individuates episodic memory from other forms of memory. On the other hand, semantic memory is related to *noetic consciousness*. Noetic consciousness dictates that the subject is conscious of the knowledge he possesses and can cognitively operate upon them, permitting thus its declaration utilizing symbolic knowledge. The difference between noetic and autonoetic consciousness is that in the former, the subject has no awareness of the qualitative temporal character of the content.

It is also important to note that there is a difference between knowing that an episode is from the past and being aware that the episode is from your past. In the former, being past is an attributed property of the representation of the event. For instance, in semantic memory, I can know that the proposition “Napoleon lost the battle of Waterloo in 1815” is past, but only because I infer it from other propositions such as “I am living in the year 2021” and “The year 1815 is past in relation to the year 2021”.[11] Whereas

[5] Tulving, “Memory and Consciousness.”, 3.

[6] Tulving, “Memory and Consciousness.”, 3.

[7] Michaelian, K., *"Mental Time Travel: Episodic Memory and Our Knowledge of the Personal Past"*, (MIT Press, 2016), 35.

[8] Tulving, “Memory and Consciousness.”, 3.

[9] Tulving, E., “What Is Episodic Memory?,” *Current Directions in Psychological Science*, no. 3 (1993), 68.

[10] Tulving, E., “Episodic Memory: From Mind to Brain.” *Annual Review of Psychology*, 53 (2002), 2.

[11] Klein, S.B., “Autonoesis and Belief in a Personal Past: An Evolutionary Theory of Episodic Memory Indices,” *Review of Philosophy and Psychology* 5, no. 3 (2014), 437.

in the latter, which is related to autonoetic consciousness, being past is a phenomenologically intrinsic property of the event that the subject represents. If you remember your last birthday party, the pastness of the event is phenomenologically embedded in the representation, so you have an experience of *what it is like to experience the episode as a part of your past.*

2.2 Episodic Memory as an Autonoetic Mental Time Travel Experience

Thus we arrive at the idea of episodic memory as *mental time travel*. When an individual remembers episodically, she can place herself in a subjective timeline, in the case of episodic memory, in the past, and can consciously re-experience the remembered event. We call this capacity to *project oneself* at a specific point in this subjective timeline, mental time travel (MTT).[12] And it is through MTT that the individual can access the experiential information that is stored in the episodic memory system. Tulving supports the idea that the ability to mentally travel to the past is strictly related to the ability to imagine or pre-experience possible future scenarios, *i.e.*, the ability to mentally time travel to the future.[13] He inferred through the case of patient N.N., later identified as patient K.C.,[14] that there must be a significant correlation between memory deficits and the inability to imagine the future. He suggests that problems in one's autonoetic capabilities may affect both the awareness of the past, as well as the awareness of the future, insofar as K.C.'s behavior indicates that he lives in a "permanent present" (more on that in section 3).

This brought to light another view about the relationship between memory and imagination, which considers both as sharing the same fundamental mental capacity, namely, one of MTT. In the current debate about the difference between memory and imagination, we have both the continuist view, which holds that memory and imagination are brain processes of the same kind, and the discontinuist view which holds that memory and imagination are different types of neural processes. Based on the same beliefs as Tulving, continuists separated MTT abilities according to their temporal orientation. The one related to episodic memory has a past temporal orientation and is called past-oriented mental time travel (PMTT). And the one related to imagined future scenarios has a future temporal orientation and is called future-oriented mental time travel (FMTT).[15] This distinction is quite important when we aim to differentiate the temporal orientation of memories and imagination. However, as I intend to discuss only concepts that are included under the term "memory", I will refer to MTT here as meaning past-oriented mental time travel.

Thereby, as far as the distinction between autonoesis and mental time travel goes, we could put it a little bit more explicitly by saying that autonoesis is the *awareness* of the

[12] Wheeler, Stuss, and Tulving, "Toward a Theory of Episodic Memory: The Frontal Lobes and Autonoetic Consciousness," *Psychological Bulletin* 121, no. 3 (1997), 331.

[13] Tulving, "Memory and Consciousness.", 5.

[14] This disambiguation can be seen in Tulving et al., "Priming of Semantic Autobiographical Knowledge: A Case Study of Retrograde Amnesia," *Brain and Cognition* 8, no. 1 (1988), 7.

[15] Perrin and Michaelian, "Memory as Mental Time Travel," *The Routledge Handbook of Philosophy of Memory* (Routledge, 2019), 228.

existence of oneself in a subjective timeline, and MTT is the *action* of mentally putting oneself at some point of this timeline. As Tulving puts it: “Mental time travel allows one, as an “owner” of episodic memory (“self”), through the medium of autonoetic awareness, to remember one’s own previous “thought-about” experiences (...)”.[16] Therefore, autonoesis is necessary for MTT for it is the medium through which one can travel through the subjective timeline that one is conscious of. Without autonoesis, there would be no MTT and as a consequence, no episodic memory. And that is because a subject that is not aware of his existence through time, is not able to project the self in this subjective timeline and experience the perceptual contents that are stored in the episodic memory system. Most continuist and discontinuist-based views, that accept that episodic memory is an experience of MTT, assume the necessity of autonoesis for MTT and consequently for episodic memory.[17]

In this way, I defend that autonoesis is necessary for episodic memory, as viewed as an experience of mental time travel. And that is because it is impossible by definition for someone to have an episodic memory without (1) having an awareness of his existence through time (autonoesis), and (2) mentally traveling to a specific past episode (mental time travel to the past). In Tulving’s words: “Autonoetic awareness (or autonoesis) is required for remembering. No autonoesis, no mental time travel”.[18] For those reasons, as far as memory is concerned, MTT is only possible through episodic memory. It is the only type of memory that conveys the personal, sensorial, and emotional information that involves the self’s immersion in his own past experiences. This is important because if we want to use concepts such as episodic memory and autobiographical memory and be able to distinguish between them, we must consider their differences, and the necessity of a subjective experience for episodic memory is one of them.

3 Difference in Phenomenology

In this section, I intend to show through the reports of the case of the amnesic patient K.C. that autobiographical memory is a type of memory that is experienced narratively and that it differs from episodic memory, which is presented perceptually. In this view, someone who autobiographically recalls can construct a life narrative that is ordered chronologically, and that can be given verbally. This narrative also encompasses greater periods of the person’s life, unlike episodic memory, which contains only short episodes. The second and main point is that autobiographical memory differs significantly from episodic memory since the latter comprises a subjectivity component, which the former does not, and that the latter can be viewed as a form of MTT, as the former cannot.

[16] Tulving, E., “Episodic Memory and Autonoesis: Uniquely Human?" In *The Missing Link in Cognition* (Oxford University Press, 2005), 9.

[17] Robins, S., “Defending Discontinuism, Naturally,” *Review of Philosophy and Psychology* 11, no. 2 (2020), 471.

[18] Tulving, “Episodic Memory: From Mind to Brain.”, 2.

3.1 The Case of Patient K.C.

After a motorcycle accident that led to a serious case of both anterograde and retrograde amnesia, K.C. lost the ability to remember episodically or form new episodic memories. With the progression of the study of his case, psychologists concluded that K.C.'s lesions to the medial temporal lobe of his brain were the cause of his severe case of amnesia. However, those studies showed that although his episodic memory capabilities were seriously impaired, his capacity to recall semantically related information was maintained.[19] In other words, although K.C. could not remember either his distant or more recent past, he could know several facts about the world and facts that occurred to him in his past.[20] This reinforced the idea that semantic and episodic memories are processed in different areas of the brain since the lesions to K.C.'s brain affected specifically his episodic memory capabilities. K.C. lacks autonoesis, which in turn results in a lack of subjective awareness in time, and consequently in an inability to access memories of the past and to think about the future (MTT).[21]

But despite that, K.C. could know general facts about the world and his past, meaning that he knew about facts that pertained to his life narrative. This preservation of knowledge relating to his personal experiences is what is important to note here. K.C. was not able to remember episodically, *i.e.* bring back to mind perceptual information of past experiences, but he had factual information about his past and was able to construct a verbally presented life narrative. K.C. knew "what year the family moved into the house where they live now, the names of the schools he went to or where he spent his summers in his teens".[22] But, although K.C. could know all that information, he could not remember it, in the episodic sense. The difference is that, although he could "remember" which year they moved to the house where they now live and the names of the schools in which he studied, he could not remember, for instance, the specific episode of the day that they moved into the new house, or a specific episode that he experienced in one of those schools. And that means that he could not relive and bring back to mind the joys and sadnesses of past.

3.2 Defining Autobiographical Memory

The main point of confusion seems to be in the term *remember*. After all, when we refer to K.C.'s capabilities of bringing past information to mind, we refer to it as remember, but we should differentiate between "remembering" in a semantic sense, which means that the subject knows facts about the past, and remembering episodically, which means that the subject mentally travels back to the past and re-experiences perceptually that episode once again. K.C. could not remember in the episodic sense, but he could remember in the semantic sense. Tulving mentions the difference in vocabulary that we should note when

[19] Rosenbaum et al., "The Case of K.C.: Contributions of a Memory-Impaired Person to Memory Theory," *Neuropsychologia* 43, no. 7 (2005), 994; Rosenbaum et al., "Amnesia as an Impairment of Detail Generation and Binding: Evidence from Personal, Fictional, and Semantic Narratives in K.C.," *Neuropsychologia* 47, no. 11 (2009), 2185; Tulving, "Memory and Consciousness.", 4.

[20] Tulving, E., "Remembering and Knowing the Past," *American Scientist* 77, no. 4 (1989), 362.

[21] Rosenbaum et al., "The Case of K.C.: Contributions of a Memory-Impaired Person to Memory Theory.", 993.

[22] Tulving, "Memory and Consciousness.", 4.

we refer to the different actions that the individual engages when he employs this or that type of memory.[23] Stating that in the natural use of language we can differentiate between those actions even when using the same term, in academic use, and to avoid ambiguities in the term "remembering", we should refer to memory in the episodic sense as "recollecting" or "remembering" and in the semantic sense as "knowing" or "recalling". Although Tulving states that K.C. has autobiographical knowledge, meaning that he knows facts that pertain to his past existence, he also says that we should distinguish it from "autobiographical memory" used here in the sense of episodic memory.[24] Furthermore, he mentions the relation of this kind of recalling to the self by saying that "It's knowledge of one's life from the point of view of an observer rather than that of a participant.[25] This means that the subject does not participate in the recall in an experiential sense, but he is "looking from the outside" (I shall treat the self-reference issue in section 4). While I agree with Tulving's affirmation that in recalling autobiographically the subject does not experience the memory as an episode, with perceptual information, I do not agree that we should not call that an autobiographical *memory*. That is because, if that were the case, semantic memory, which is also constituted of propositional information, should not be considered memory as well. So, it seems plausible to call *autobiographical memory*, the propositional knowledge in which: (1) the subject has information of the personal past that is presented verbally, and (2) that allows for the knowledge and construction of an extended linear life narrative.

Additionally, concerning the types of consciousnesses implied in distinct kinds of memory, it is a consensus among psychologists and philosophers of memory that episodic memory relates to autonoetic consciousness, and that semantic memory relates to noetic consciousness. But what about autobiographical memory? Just as semantic memory is related to noetic consciousness, there is no reason to assume that autobiographical memory would not be, since it is based mostly on information of semantic nature, and it is presented propositionally. A noetic conscious being is aware of the knowledge he possesses and can act upon objects, concepts, events, and their relationship, without their presence, employing symbolic knowledge.[26] Furthermore, unlike noetic consciousness and autonoetic consciousness, an anoetic conscious being can perceive the environment, represent its perception, and "behaviourly respond to aspects of the present environment.".[27] In this manner, anoetic consciousness is related to procedural memory, that is, the memory for motor skills learned in the past that the individual can use in the present, like remembering how to ride a bicycle. Autobiographical memories are not related to motor skills, neither are they related to re-experiencing the past and, therefore, the type of consciousness they imply cannot be explained either by anoetic consciousness or autonoetic consciousness. Rather, given the similarity in content between autobiographical memories and semantic memories, it seems reasonable to assume that, like semantic memory, they are presented with noetic consciousness. K.C., for instance, is aware of the knowledge that he possesses about himself and can present this knowledge verbally in an extended temporal manner. Thus, I believe that the awareness of his autobiographical semantic knowledge can be sufficiently explained in terms of noetic consciousness.

[23] Tulving, "Remembering and Knowing the Past.", 362-363.
[24] Tulving, "Remembering and Knowing the Past.", 362-363.
[25] Tulving, "Remembering and Knowing the Past.", 362-363
[26] Tulving, "Memory and Consciousness.", 3.
[27] Tulving, "Memory and Consciousness.", 3.

K.C.'s vision of his past is presented as phenomenologically dry as his general knowledge about the world, such as "2+2=4" or that "humans are mammals", but it is nonetheless a representation of his existence. It is not difficult to imagine that he could have, for instance, drawing solely on semantic information, put together a narrative that goes from his childhood to his adulthood, in chronological order. Those memories are autobiographical because they present the happenings of his life, and he is conscious of them. The difference between remembrance of past episodes with an associated phenomenal experience of reliving the past episode as *past*, and a semantical representation of one's past is what is worthwhile noting for present purposes.

One objection that might arise from this view concerns whether K.C. could organize his life narrative in chronological order without episodic memory. Episodic memory is known as the type of memory that enables us to order events of our past. So, the agent only knows that an event E_1 came before an event E_2 because he *experienced* those events in this order. However, in cases of impaired episodic capacities, it is still possible for the agent to order the semantic knowledge that pertains to a life narrative. This can happen if (1) there is temporal-related information embedded in the content,[28] (2) if the agent learns the temporal order of the events,[29] or (3) if he can interpret or infer the temporal relations from the content given to consciousness.[31] Thus, it would be possible for K.C to order chronologically a life narrative based on information that carries explicit temporal information such as "I used to live in this house during *my childhood*", or "I remember moving to this house in the year *1985*". These two examples give different degrees of specificity of the temporal information that can enable an organization of the narrative structure of autobiographical memories. Surely, in the case of neurotypical agents, the process of organization would be much more fluent, because they can use information derived from episodic memories to do so, but that does not rule out the possibility that K.C could have done it either, even if in a more elemental level. Also, just because neurotypical agents may in some cases use episodic memories to help order chronologically events in a life narrative, that does not mean that this episodic information is *necessarily* a part of the life narrative.

Therefore, if we can say that K.C. has a narrative vision of his life because of his propositional knowledge about his past and that this conscious narrative representation is considered autobiographical memory, then we must agree that K.C. can recall autobiographically, although completely incapable of remembering episodically. This contradicts the common idea in the literature, that episodic memory is the same as autobiographical memory.

Furthermore, assuming the phenomenological difference between episodic and autobiographical memories, and given the more temporal extended quality of the latter, it seems unrealistic to assume that the individual should have sensorial, iconic, and complex information of all these life periods. Thus, the assumption that the ability of MTT is present in most of the content of autobiographical memory, as it seems to be suggested by the identification of episodic and autobiographical memories, can be seen as cognitively unrealizable. Episodic memory is related to shorter and self-contained

[28] Tulving, E., "Episodic and Semantic Memory," in *Organization of Memory* (Academic Press 1972), 389-390.

[29][30]

[31] Klein, "Autonoesis and Belief in a Personal Past.", 437-438.

episodes that have emotional or sensorial relevance to the subject. We remember things that are remembering-worthy, meaning that we remember things that have emotional importance to us. So, if one asks if we think that we are living happy lives, even though the question has an emotional character, our response would probably be based on autobiographical memories because the episodes of happiness and sadness are too many to remember and account for.

My argument states that autobiographical memories can exist without episodic information, in cases such as K.C's. This could suggest two different views on the nature of autobiographical remembering. First, it could be considered that semantic information is more fundamental to the life narrative than episodic information. That is to say that in neurotypical agents, there is some episodic information in the content of autobiographical memories, but in a considerably low amount when compared to semantic information. As seen above, it is more plausible to defend that most of the content is of propositional origin, as it would be unrealistic to assume that agents could remember all, or even most of the perceptual information that would pertain to an entire life narrative. Second, on a stricter view, it could be defended that all the content of the autobiographical memories is propositional, with no place for episodic information. That is to defend that although in neurotypical agents narratives can contain some semantical information that *could elicit* an episodic memory, the life narrative itself is not presented perceptually, it is, rather, presented propositionally. In the first position, autobiographical memories would have to be able to support both autonoetic and noetic consciousness. So, for this reason, I defend that the second position is more suitable because it is a simpler account of the type of consciousness of autobiographical memories. But independently of which view we choose, it is still evident that a difference between episodic and autobiographical memories is necessary, insofar as K.C.'s case shows empirical evidence of an autonoetic difference between autobiographical memories and episodic memories.[32]

4 Memory's Reference to the Self

This section aims to discuss whether we should consider K.C's semantic knowledge of his autobiography an autobiographical memory or whether it should be considered memory at all.

The field of investigations of the self and its relation to memory is vast. Although I do not intend here to give a complete account of the subject, it is worth explaining a basic conceptual difference that tries to shed light upon the distinctions between episodic and autobiographical memory and its self-references, that will be useful for the discussion of whether autobiographical memory should be considered a separate kind of memory. To do so, I will borrow Baddeley's distinction of the types of memory in which, on one hand, the self is the experiencer (1), and on another, the self is the object of the experi-

[32] In the literature concerning the Simulation Theory of Memory, the preferable term for the episodic memory system might be "episodic construction system", that encompasses the whole of the imaginative processes that are able to construct mental scenarios based on episodic information. For the purposes of this article, I preferred to treat it more generally, as "episodic memory system", but I think my view could be applicable to a simulational framework.

ence (2).[33] I consider, opposite to him, that episodic memories are more closely related to type (1), and that autobiographical memories are more closely related to type (2). To put it more clearly, it is as if when we are in a mnemic mental state, there are two selves: The present self or the "experiencer self" and the past self or the "experienced self". In remembering, it is as if both selves are aligned or superpositioned, and consequently, the subject has a full qualitative and personal experience of the episode. This is partly stated by Perrin when he says: "Autonoeticity implies the identity of the self whose experience is simulated with the simulating subject".[34] He also states the same in: "First, for the appearance of episodic memory to occur, I must have the belief that I am the subject whose past experience I represent. This identity belief is a condition of the episodic appearance".[35] Conversely, in recalling, the selves are kept separate, meaning that the present self relates to the recall as an observer instead of an experiencer.

In autobiographical memory, as in a mnemic state that is based on semantic information, it does not seem to be the case that there is a superpositioning of both selves. To defend that this is the case is to consequently affirm that autobiographical memories could share the same phenomenological experiences or would at least be capable of substantiating the same phenomenological complexity that episodic memory is capable of. And that is because the only way in which the "experiencer self" can be identified with the "experienced self" is in an autonoetic state. If my argument, in which autobiographical memories are related to semantic information, and its content presented by noetic consciousness is right, then the alignment of both selves would not be reasonable, because they would define autonoetic awareness, which as I tried to show, is non-present in autobiographical memories.

One objection that can emerge from my ideas is whether we should consider a narrative view of our lives as a type of memory. This point has already been made by Klein.[36] He argues that only those kinds of memories that we would describe as episodic (mental states with a past-oriented subjectivity), can be conceived as memories. This means that the autobiographical knowledge that K.C. has which enables him to have information not only about the past but the past that relates to his life, cannot be considered a memory of any kind. While I would agree with Klein by saying that he cannot remember episodically, because he lacks the ability to mentally travel to the past or to even conceive himself in a subjective timeline, I think it is too extreme to not consider it memory. And that is mainly because the problem of the differentiation of those entities can be solved, as Tulving did, by referring to it by different actions. For this reason, I argue that my view shows a degree to which we can still call the knowledge that K.C. has of his past "memory", by referring to it as autobiographical memory, as it possesses two of the main consensual general characteristics, that seem to be important for the recognition of a mental state as a memory, which are: (1) a present reference to the past, meaning that the present information refers to something that already happened, and (2) a reference to a self, meaning that the subject knows (or feels) that the information brought to mind refers to his experience.[37]

[33] Baddeley, A., "What Is Autobiographical Memory?," in *Theoretical Perspectives on Autobiographical Memory* (Dordrecht: Springer Netherlands, 1992), 19.

[34] Denis Perrin, "Asymmetries in Subjective Time," In *Seeing the Future: Theoretical Perspectives on Future-oriented Mental Time Travel* (Oxford University Press, 2016), 46.

[35] Denis Perrin, 52.

[36] Klein, S.B., "What Memory Is," *Wiley Interdisciplinary Reviews: Cognitive Science* 6, no. 1 (2015), 1.

[37] Criteria (i) and (ii) seem to be consensual throughout most of the philosophical tradition on memory. I think that the differ-

At this point, with the empirical evidence shown, to say that autobiographical memories are the same as episodic memory, is, at least, debatable. Taking Klein's approach and saying that autobiographical memory is not memory at all is a possibility, but one that is very costly because it would mean that semantic memory, which similarly to autobiographical memory draws its contents from propositional information, would not be memory also. Although Klein's view is plausible, I believe that a more intermediate view in which both episodic memory and semantic memory are types of memory, or even a broad view, which holds that episodic, semantic, and even procedural memories are indeed forms of memory,[38] is more adequate.

Also, to call an autobiographical memory a semantic memory, because of its relatedness with propositional information, seems at least questionable. Indeed, although both semantic and autobiographical memories pertain to the self, meaning that both are of things that the self knows, autobiographical memory goes further and is also knowledge about things that happened to the self. And for that reason, I argue that they are different. In autobiographical memory, the present self is an observer of the facts that happened to the past self, without their identification, which would equate to episodic memory. In semantic memory, however, the present self is an observer of general facts about the world that he knows, but not about his past self. I believe that to differentiate clearly between semantic, autobiographical, and episodic memories may help us understand the capability of the brain to maintain a verbally available life narrative, even in cases where autonoetic awareness is missing.

5 Conclusion

Considering what has been demonstrated I conclude that even though they are sometimes still treated as a synonymous term throughout philosophical and psychological discussions, there are good reasons to believe that episodic and autobiographical memory should be considered fundamentally different. Considering that the kind of recall that K.C. has is of autobiographical relevance and that it differs in phenomenology, meaning that it lacks the autonoetic component and it is not presented as an experience of MTT to the past, as episodic memory is, then there should be no reason to identify the two. Furthermore, I argued that autobiographical recalling should be considered a separate kind of memory, than for instance, semantic memory, because its content relates to the self in a more meaningful way than the content of semantic memory, as can be demonstrated by insights about the relation of memory and the self.[39] Concerning the content, I show that the information of autobiographical memories, while similar to semantic information since it relates to facts, consists nonetheless, in facts about agent's life happenings

entiation between episodic and autobiographical memories can be defended both in a causal framework as well as a simulational framework. The only requirement seems to be that the theory supports a view in which episodic memories are presented as an experience of MTT. However because the main theories of memory deal with the difference between memory and imagination, and here I am proposing a differentiation between two types of memory, I left other more specific criteria out. To more information on different theories of memory, see Michaelian and Robins, "Beyond the Causal Theory? Fifty Years after Martin and Deutscher," *New Directions in the Philosophy of Memory* (Routledge, 2018), 13–32.

[38] Michaelian, "Opening the Doors of Memory: Is Declarative Memory a Natural Kind?," *Wiley Interdisciplinary Reviews: Cognitive Science* 6, no. 6 (2015), 476.

[39] Baddeley, "What Is Autobiographical Memory?"; Perrin, "Asymmetries in Subjective Time."

that are, however, not presented perceptually as is episodic memory.

Bibliography

Baddeley, Alan. “What Is Autobiographical Memory?” In *Theoretical Perspectives on Autobiographical Memory*, 13–29. Dordrecht: Springer Netherlands, 1992.

Klein, Stanley B. “Autonoesis and Belief in a Personal Past: An Evolutionary Theory of Episodic Memory Indices.” *Review of Philosophy and Psychology* 5, no. 3 (2014): 427–47.

———. “What Memory Is.” Wiley Interdisciplinary Reviews: *Cognitive Science* 6, no. 1 (2015): 1–38.

Michaelian, Kourken. *Mental Time Travel: Episodic Memory and Our Knowledge of the Personal Past*. The MIT Press. The MIT Press, 2016.

———. “Opening the Doors of Memory: Is Declarative Memory a Natural Kind?” *Wiley Interdisciplinary Reviews: Cognitive Science* 6, no. 6 (2015): 475–82.

Michaelian, Kourken, and Sarah K. Robins. “Beyond the Causal Theory? Fifty Years after Martin and Deutscher.” *New Directions in the Philosophy of Memory*, 2018, 13–32.

Perrin, Denis. “Asymmetries in Subjective Time.” *Seeing the Future*, 2016, 39–61.

———, and Kourken Michaelian. “Memory as Mental Time Travel.” *The Routledge Handbook of Philosophy of Memory*, no. January 2017 (2017): 228–39.

Robins, Sarah. “Defending Discontinuism, Naturally.” *Review of Philosophy and Psychology* 11, no. 2 (2020): 469–86.

Rosenbaum, R. Shayna, Asaf Gilboa, Brian Levine, Gordon Winocur, and Morris Moscovitch. “Amnesia as an Impairment of Detail Generation and Binding: Evidence from Personal, Fictional, and Semantic Narratives in K.C.” *Neuropsychologia* 47, no. 11 (2009): 2181–87.

Rosenbaum, R. Shayna, Stefan Köhler, Daniel L. Schacter, Morris Moscovitch, Robyn Westmacott, Sandra E. Black, Fuqiang Gao, and Endel Tulving. “The Case of K.C.: Contributions of a Memory-Impaired Person to Memory Theory.” *Neuropsychologia* 43, no. 7 (2005): 989–1021.

Tulving, Endel. “Episodic and Semantic Memory.” In *Organization of Memory*, 1:381–403, 1972.

———. “Episodic Memory: From Mind to Brain.” *Annual Review of Psychology* 53, no. 1 (February 2002): 1–25.

———. “Memory and Consciousness.” *Canadian Psychology/Psychologie Canadienne*

26, no. 1 (January 1985): 1–12.

———. “Remembering and Knowing the Past.” *American Scientist* 77, no. 4 (1989): 361–67.

———. Tulving, Endel. 2005. "Episodic Memory and Autonoesis: Uniquely Human?" In H. S. Terrace, & J. Metcalfe (Eds.), *The Missing Link in Cognition* (Pp. 4-56). New York, NY: Oxford University Press.” Cognition, 2005, 4–56.

———. “What Is Episodic Memory?” *Current Directions in Psychological Science* 2, no. 3 (June 22, 1993): 67–70.

———, Daniel L. Schacter, Donald R. McLachlan, and Morris Moscovitch. “Priming of Semantic Autobiographical Knowledge: A Case Study of Retrograde Amnesia.” *Brain and Cognition* 8, no. 1 (1988): 3–20.

Wheeler, Mark A., Donald T. Stuss, and Endel Tulving. “Toward a Theory of Episodic Memory: The Frontal Lobes and Autonoetic Consciousness.” *Psychological Bulletin* 121, no. 3 (1997): 331–54.