
Aporia XXII

Aporia

Undergraduate Journal of the St Andrews Philosophy Society

VOLUME XXII

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 Everyone,

With another tumultuous year of no classes and 5 weeks of strikes (in one semester I mind you) to compound that frustration we are now able to release the 22nd edition of Aporia. Our goal was to have a separate feminist edition for the first time, and although we didn't get the submissions to make this a possibility, we did get enough to publish a feminist appendix, we hope to make this a reality in the coming year.

I would like to thank Olivia Griffin for doing so much for this edition, at times taking the role of head editor even, and taking the reins for the coming year, she's been an indispensable asset and to have her at my side through it all has been, and continues to be, amazing. I would also like to thank Roberto Garcia as the deputy editor as well as Louisa McDonald who headed the first feminist edition as well as all editors and everyone who has enabled this edition to come together.

Special thanks also goes to Kyle Scurville for always being an inspiration to me and assisting me tremendously throughout the academic year. Thank you all who read this and I hope you enjoy the edition!

All the Best,

Nigel A. Mika

Acknowledgements

Nigel Mika
Editor-in-Chief

Luis Roberto Garcia Martinez
Deputy Editor

Louisa McDonald
Feminist Edition Editor

Editors

Rebecca Beynon

Paul Brull

Ciel Burges

Janan Choi

Olivia Griffin

Sophia Pawliw

Heather Reid

Sarah Routley

Anna Smith

Oliver Staples

Lara Thain

Mia-Belle Tierney

Eric Wallace

Olivia Griffin
Cover Art

Luis Roberto Garcia Martinez
Cover Design and Typesetting

Contents

A Phenomenological Approach to the Bayesian Grue Problem

Ibrahim Dagher*

University of California, Davis

It is a common intuition in scientific practice that positive instances confirm. This confirmation, at least purely based on syntactic considerations, is what Nelson Goodman's 'Grue Problem', and more generally the 'New Riddle' of Induction, attempt to defeat. One treatment of the Grue Problem has been made along Bayesian lines, wherein the riddle reduces to a question of probability assignments. In this paper, I consider this so-called Bayesian Grue Problem and evaluate how one might proffer a solution to this problem utilizing what I call a phenomenological approach. I argue that this approach to the problem can be successful on the Bayesian framework.

1 Introduction

It is a common intuition in scientific practice that positive instances confirm. That is, that repeated instantiations of some predicate P lend inductive support to a general hypothesis wherein P is projected. The hope that such syntactic considerations might serve as the basis for an inductive logic is what Goodman[1] sets out to defeat in his so called 'New Riddle' of induction. As such, the New Riddle has received considerable discussion, including treatments of the riddle along Bayesian lines.[2]One such reformulation has been proffered by Sober (1994), which prompts new considerations — such as how different *kinds* of hypotheses differ with respect to their confirmation conditions — and how this might give rise to various manifestations of the riddle. In this paper, I consider the New Riddle cast in the Bayesian framework proposed by Sober, and appraise a 'phenomenological approach' to the riddle. I argue that the approach, as applied to the grue problem, can be successful. I will proceed as follows. In §2, I explicate a Bayesian formulation of the grue problem along the lines Sober[3] outlines. In §3, I discuss some general difficulties Bayesian answers will have to deal with. I outline the phenomenological approach to answering the problem in §4 before concluding in §5.

2 A Bayesian Grue Problem

Consider the predicate 'grue', which applies to any x just in case it is green and examined earlier than some time t or blue and examined at or later than t. Following Sober (1994), we can now begin to concern ourselves with various hypotheses from which the riddle will emerge. First, consider these two hypotheses, which are said to be *generalizations*:

*My name is Ibrahim Dagher, and I'm a second-year philosophy student at UC Davis. I have a special research interest in epistemology, as well as philosophy of law and religion. I enjoy playing basketball and doing the daily Wordle!

1. Nelson Goodman, *Fact, Fiction, and Forecast* (Harvard University Press, 1955).

2. For formulations other than Sober's, see Irving John Good, "Explicativity, Corroboration, and the Relative Odds of Hypotheses," *Synthese* 30, nos. 1-2 (1975): 39–73; Richard C. Jeffrey, *The Logic of Decision* (New York, NY, USA: University of Chicago Press, 1965)

3. Elliott Sober, "Grue!: The New Riddle of Induction," chap. No Model, No Inference: A Bayesian Primer on the Grue Problem, ed. Douglas Stalker (Open Court Publishing Group, 1994), 225–238.

1. **AllGreen:** All emeralds are green.
2. **AllGrue:** All emeralds are grue.

Presumably, AllGreen is a perfectly rational generalization to commit oneself to. However, AllGrue does not seem to be. Thus, the first question of the riddle is this: what asymmetry exists between AllGreen and AllGrue, such that we are justified in our belief in the former *rather* than the latter?

There is also another question to be asked at this point. Consider these two hypotheses, which are instead said to be *predictions*:

1. **NextGreen:** The next emerald to be examined will be green.
2. **NextGrue:** The next emerald to be examined will be grue.

Again, presumably it would only be rational to believe the first prediction, assuming the next emerald will be examined at or later than *t*, so that these predictions are contradictory. We are thus compelled to ask: what asymmetry exists between NextGreen and NextGrue, such that we are justified in our belief in the former rather than the latter? There are two distinct issues at hand: the first is finding some epistemic asymmetry between AllGreen and AllGrue, and the second is finding one between NextGreen and NextGrue.

Finally, I wish to make one more distinction. The specific hypotheses and their respective questions, as formulated herein, are what I take to constitute the grue problem — the problem of finding some epistemic asymmetry between AllGreen and AllGrue, and NextGreen and NextGrue *specifically*. The New Riddle is the problem of characterizing the epistemic relationships between hypotheses of generalizations, predictions, and their respective instantiations *more generally*. This distinction is important because the solution I propose here ought to be considered only a solution to the grue problem, and not the much more general riddle.

2.1 Bayesian Confirmation Conditions

With these questions on the table, we can now move to explaining what the sufficient conditions are for answering these questions. As mentioned above, these conditions shall be cast along Bayesian lines.

First, since the sought conclusion to both of our questions will take the form 'hypothesis H_1 can be assigned a higher *posterior probability* than hypothesis H_2 because...' it is worth explaining what obtaining a posterior probability of a given hypothesis consists of for Bayesians. Where H abbreviates some hypothesis and O abbreviates the set of observations we have made, Bayes' theorem tells us that the posterior probability of H can be calculated by reference to the likelihood and prior probability of H, as well as the probability of O:

$$Pr(H \mid O) = [Pr(O \mid H) * Pr(H)] / Pr(O)$$

The Pr(H | O) is the posterior probability of H––the probability that H is true given the observations we have made. Pr(O | H), on the other hand, is the *likelihood* of H: the probability H confers onto O's obtaining. Lastly, the Pr(H) is what is often termed the *prior* probability of the hypothesis: the probability H enjoys before any observations are made.

Since the nature of both of our questions is comparative, we should wish to reformulate Bayes' theorem into a comparative principle. This is simple enough:

$$\text{(CPs): } Pr(H_1 \mid O) > Pr(H_2 \mid O) \text{ when and only when}$$

$$[Pr(O \mid H_1) * Pr(H_1)] > [Pr(O \mid H_2) * Pr(H_2)]$$

Interestingly, the comparative principle as explicated above is a *synchronic* one. We might also wonder what difference in the probability of H is *incited* by the truth of O. In other words, we may also be interested in a *diachronic* comparative principle. Assuming that the larger the difference between the posterior and prior probabilities of a hypothesis, the greater the confirmation, then:

(CP_d): O confirms H_1 more than H_2 when and only when

$$[Pr(H_1 \mid O) - Pr(H_1)] > [Pr(H_2 \mid O) - Pr(H_2)]$$

Thus, (CP_d) differs from (CP_s). So, two further subdivisions have to be made with respect to the issues at hand: not only must we consider the relevant probabilities of AllGreen compared to AllGrue and NextGreen compared to NextGrue, but each comparison must be considered diachronically and synchronically. Let us diagnose each in turn.

2.2 AllGreen vs AllGrue: A Synchronic Analysis

Suppose 'D' denotes a proposition that contains the relevant past data, namely, 'all emeralds examined have been observed to be green'. To analyze the posterior probabilities of AllGreen and AllGrue synchronically, it is important to begin with what is commonly affirmed: the truth of either AllGreen or AllGrue entails D. Thus, the likelihoods of either hypothesis are exactly 1. This is just the fact that our past data confirms both generalizations.

However, given (CP_s), if two hypotheses are of equivalent likelihoods, the only way in which one could have a higher posterior probability than the other is if one has a higher prior probability than the other. That is:

$$Pr(AllGreen \mid D) > Pr(AllGrue \mid D) \text{ when and only when}$$
$$Pr(AllGreen) > Pr(AllGrue)$$

If this is correct, our condition for preferring AllGreen rather than AllGrue is this: AllGreen enjoys a higher prior probability than AllGrue. More will have to be said about what might qualify — or if anything at all can qualify — as a justified reason for such prior probability assignments.

2.3 AllGreen vs AllGrue: A Diachronic Analysis

Much like the synchronic analysis, the conditions for different posterior probabilities on the diachronic analysis appear to reduce to considerations of prior probability. Using CP_d with Bayes' theorem, we obtain the following for AllGreen:

$$[[Pr(D \mid AllGreen) * Pr(AllGreen)] / Pr(D)] - Pr(AllGreen)$$

And the same for AllGrue:

$$[[Pr(D \mid AllGrue) * Pr(AllGrue)] / Pr(D)] - Pr(AllGrue)$$

As an inequality, this transforms into:

$$[[1 - Pr(D)] * Pr(AllGreen) / Pr(D)] > [[1 - Pr(D)] * Pr(AllGrue) / Pr(D)] \text{ And, on the assumption that } Pr(D) < 1, \text{ we obtain:}$$
$$Pr(AllGreen) > Pr(AllGrue)$$

Thus, on our diachronic analysis the posterior probabilities are higher for the AllGreen hypothesis than the AllGrue hypothesis when and only when the priors are higher *and* our data was not certain.

2.4 NextGreen vs NextGrue: A Diachronic Analysis

For our predictive hypotheses, the conditions under which we can assign comparatively higher posterior probabilities change. I will begin with the diachronic case. At first glance, it might be thought that because our past data confirms and raises the probability of the general hypotheses, it ought to also confirm and raise the probability of the predictive hypotheses. After all, the truth of either general hypothesis *entails* the truth of the respective predictive hypothesis.

However, this is not so. At least, not without significant assumptions about the sampling process involved. If I know that some marbles placed in a bag were randomly sampled from a source with an equivalent ratio of black to blue to red to green marbles, then the fact that every marble I have examined has been red does not confer any further probability on the predictive hypothesis 'the next marble will be red'. The probability remains 0.25. Yet, the fact that every marble I have examined has been red *does* increase the probability that every marble is red, by virtue of the fact that this has eliminated certain hypotheses from the possibility space (namely, all the hypotheses entailing that less than x-many red marbles would be examined, such as the hypothesis that all the marbles are black).

This asymmetry in confirmation arises precisely because of my knowledge of the sampling process. My knowing that the marbles do not have their colors selected, as it were, *collectively*, or by some law-like process, precludes the possibility that all the marbles' being homogeneous in color is anything other than mere happenstance. It is only when this possibility is introduced that one can begin to alter the probability of a predictive hypothesis.[4]

In other words, it is only when the conjunction of the relevant prediction and the data is more probable than the independent occurrence of each that the data confirms the prediction. Evidence confirms a prediction only if the two are *positively correlated*, or dependent, facts. If they are independent, then their conjunction can never be more probable than the occurrence of both of their conjuncts.

With this analysis in hand, we are now prepared to outline the probabilistic conditions on which we ought to prefer NextGreen over NextGrue. First, assume that the next emerald observed will be either green or blue. Next, assume the present moment is *t*, so that NextGreen and NextGrue are contradictory, and logically exhaustive, hypotheses. So, if some condition confirms NextGreen, it will disconfirm NextGrue. Here is the condition:

(C_d): NextGreen is confirmed by data D if and only if

$$Pr(\text{NextGreen} \,\&\, D) > [Pr(\text{NextGreen}) * Pr(D)]$$

Why might we think that the probability of the conjunction of NextGreen and our past data is greater than the independent occurrence of each of these facts? Presumably it is because of an assumption about the nature of emeralds and their color: namely, the color predicate that is ultimately true of emeralds should be true of them *qua* their being emeralds. That is, we assume their color is determined as a *group*. It is not as though each emerald is sampled from a possible space of colors individually and independently of any other emerald. The more pressing question that arises at this point is not that of why we might think the inequality would hold, but rather why we should think *this* inequality holds. Plausibly, NextGrue is also positively associated with the past data in the same way that NextGreen is. Our motivations for thinking that emeralds would collectively be green apply equally well for thinking that emeralds would collectively be grue.

This question will soon be addressed, but the important lesson here is this: NextGreen and NextGrue have slightly different conditions for epistemic asymmetry than do AllGreen and AllGrue. For our past data to confirm the generalizations, we need some reason to prefer a certain assignment of priors. For our past data to confirm the predictions, we need some reason to prefer a certain positive association over another.

4. For more on this relationship, see SoberElliott Sober, "Confirmation and Law-Likeness," *Philosophical Review* 97, no. 1 (1988): 93–98

2.5 NextGreen vs NextGrue: A Synchronic Analysis

Finally, let us consider how our past data might serve to confirm the predictive hypotheses on a synchronic analysis. Holding fixed the aforementioned conditions that made it such that NextGreen and NextGrue were contradictory and logically exhaustive hypotheses, on a synchronic analysis the question of posterior probability assignment boils down to the following:

$$\text{When is } Pr(\text{NextGreen} \mid D) > Pr(\text{NextGrue} \mid D)?$$

Since Pr(NextGrue) = 1 – Pr(NextGreen), this can be expanded to:

$$Pr(\text{NextGreen} \,\&\, D) - [Pr(\text{NextGreen}) * Pr(D)] > [Pr(D) * [1 - 2 * P(\text{NextGreen})]] / 2$$

Simplifying:

$$Pr(\text{NextGreen} \,\&\, D) > [Pr(D) / 2]$$

This becomes:

$$Pr(\text{NextGreen} \mid D) > 0.5$$

Thus, the synchronic case is similar to the diachronic case: insofar as NextGreen is positively associated with D, and the Pr(NextGreen | D) > 0.5, it follows that Pr(NextGreen | D) > Pr(NextGrue | D).

Unsurprisingly, there remains a kind of epistemological indeterminacy under both analyses with respect to which predictive hypothesis ought to be assigned the aforementioned positive association. There is nothing, it appears, in D that could possibly account for that kind of preferential assignment.

3 Difficulties on the Bayesian Framework

Very roughly, answers to the Bayesian grue problem will have to amount to a favorable prior probability assignment to AllGreen rather than AllGrue, and a favorable assignment of positive association between NextGreen and the past data rather than NextGrue. But before attempting to characterize a certain answer as meeting these conditions, there appear to be deeper difficulties with even meeting these conditions at all.

The first difficulty is the well-known *problem of the priors*. That is, what norms ought to dictate the distribution of prior probabilities to any logically exhaustive set of hypotheses under consideration? Is the only requisite norm a requirement on cohering with the axioms of probability (Subjective Bayesianism)? Or is there, in addition to this norm, a norm on which our priors follow some concern for evidential or reason-based indifference (Objective Bayesianism)? Or is it rather that our priors should be such that conditionalizing on a given class of evidence produces posteriors that would be in line with some theoretical virtue like explanatory power, simplicity, or convergence to truth (Future-Oriented Bayesianism)? This is an ongoing debate, and if an answer to the grue problem requires that our past data supports AllGreen *rather than* AllGrue if and only if the prior of the former is greater than the prior of the latter, any attempted solution will have to contend with the question of what kinds of norms ought to dictate our prior probability assignments.

Sober (1994) raises another obstacle for any solution to the Bayesian grue problem. It can be put as follows: either the prior probabilities we are considering are objective or subjective. If they are objective, then they cannot possibly be assigned *a priori*. If they are subjective, then varying prior probability assignments could not possibly amount to an epistemic asymmetry between AllGreen and AllGrue.

However, this dilemma, at least with respect to the grue problem, can be dissolved. It is obvious that the probabilities discussed with respect to AllGreen and AllGrue are not objective chances. Indeed, such a notion appears to be fraught in the context of the question of what colors emeralds instantiate. Rather, I take the probabilities discussed herein to be *credences*. The Pr(AllGreen | D) represents the credence, or degree of belief, one has in the hypothesis AllGreen on the past data. However, *pace* Sober, this fact does not remove the possibility of an epistemic asymmetry. This would only be the case if our answer to the question 'Why believe AllGreen *rather than* AllGrue?' were the descriptive answer 'Because *in fact* my credence in AllGreen given the data is greater than my credence in AllGrue given the data.' But, as I see it, the answer we will take on is actually normative. It is of the form, 'Because there is a (true) norm *N* according to which I ought to have a greater credence in AllGreen given the data rather than AllGrue given the data.' Surely if I ought to have a greater credence in some hypothesis compared to another hypothesis that constitutes a substantive epistemic asymmetry between the two hypotheses.[5]

4 The Phenomenological Approach

Having now remarked on what conditions the Bayesian grue problem requires of answers, and recognizing some general difficulties with any answer, I now wish to briefly sketch a solution that I call the 'phenomenological approach', and discuss how it provides new insights into the conditions and difficulties analyzed above.

Let's begin with noting the following difference between AllGreen and AllGrue. If all emeralds are grue, and there are emeralds examined earlier and later than *t*, then there is a phenomenological asymmetry in the world. That is, before *t* we will have a certain phenomenological experience associated with the observation of emeralds. Then, after t, we will have a noticeably different phenomenological experience when we view emeralds. We might put the point as follows: grue emeralds are perceptually different.

On the other hand, if all emeralds are green, then there is phenomenological symmetry. That is, there is a constant phenomenological experience associated with the observation of emeralds, since green emeralds are perceptually the same.

The phenomenological approach to the grue problem attempts to draw an epistemic asymmetry between AllGreen and AllGrue on the basis of this difference.[6] It is important, however, to recognize what is *not* being claimed as a difference between the two hypotheses. It is not being claimed that grue emeralds do not instantiate the same color across times (or some equivalently grue-ified predicate, 'grulor'). Nor is the claim that grue emeralds *qua* emeralds experience some kind of change in their properties across times. The claim is merely relative to green speaking humans: if there are emeralds examined earlier and later than *t*, and these emeralds are grue, there will be a change in our phenomenological experience.

4.1 Building an Asymmetry

I take the lesson of the New Riddle to be the following: observing that a certain predicate P has consistently applied in the past does not *by itself* warrant the projection of P into the future. This is because there are many predicates P', P"…

5. I recognise that there are many pressing issues concerning the objectivity of norms, as well as the epistemology associated with being justified in their assertion. Unfortunately, because of space, I am unable to provide evaluations of these questions. What I am concerned with is whether there could be *any such norm* applicable to the grue problem.

6. For another attempt at utilizing this fact to solve the grue problem, see Sydney Shoemaker, "Functionalism and Qualia," *Philosophical Studies* 27, no. May (1975): 291–315, https://doi.org/10.1007/BF01225748.

that apply equally well of the same past phenomena, but if they were to be projected about phenomena examined in the future, would contradict P. There must be some other consideration, apart from consistent application to the past, that distinguishes P from P', P"….

It's at this point that the phenomenological approach begins. It notes that rather than projecting any predicates that have been true in the past, we ought to instead project the phenomenological experiences that have been had in the past.

Certainly, the principle that we ought to project the phenomenological experiences associated with past phenomena is just as, if not more, intuitive than the principle that we ought to project the predicates of past phenomena.

But now our epistemic asymmetry emerges. For while it is true that a whole host of predicates apply to the same past data, and thus the principle that 'predicates of past phenomena ought to be projected' is false, there is only *one* way in which we have experienced past data. There is only one phenomenological experience associated with the observation of emeralds, and thus the principle 'the phenomenological experience associated with the past data ought to be projected' cannot be defeated by consideration of the various predicates that might all equally apply to the data. And, since only AllGreen is consistent with the application of this principle, it is on this basis that we ought to prefer AllGreen to AllGrue.[7]

Another way of putting the idea is as follows. Goodman's New Riddle tells us that the following inductive schema (alone) is faulty:

1. a_1 is green

2. a_2 is green

3. a_3 is green

.

.

.

C: All *a*'s are green.

The phenomenological approach does not attempt to provide a semantic or epistemic explanation as to why only 'green', and not 'grue', will fit this schema in a truth preserving way. Rather, it proposes the alternative schema, Sp, with the principle that the only predicates that ought to be projected are those that entail a projection of perception:

1. a_1 is green and there is a single phenomenological experience *p* associated with observing a_1

2. a_2 is green and *p* is associated with observing a_2

3. a_3 is green and *p* is associated with observing a_3

.

.

.

Principle: If all *a*'s are green, then *p* is associated with observing all *a*'s C: All *a*'s are green

7. Interestingly, Barry Ward, "Explanation and the New Riddle of Induction," *Philosophical Quarterly* 62, no. 247 (2012): 365–385 uses a similar principle, built instead along explanatory rather than phenomenological lines, to reach this conclusion.

The idea at the heart of S_p is that a predicate ought to be projected for some *a*'s just in case (i) there are many positive instances of the predicate amongst the *a*'s and (ii) if the predicate were to be projected, that would entail projecting the same phenomenological experience that has been true of the past *a*'s.

I mentioned that the first inductive schema suffers from the following objection (which is just the grue problem): a substitution of 'grue' for 'green' in the argument yields a set of true premises, but a (supposedly) false conclusion. This schema alone, then, cannot be all there is to the inductive logic. Some semantic or epistemic considerations of the predicates inserted into the schema are at play, otherwise the schema is not truth preserving.

Might S_p suffer the same fate? Consider substituting 'grue' for 'green'. For each grue emerald it is true that the observation of that emerald has a phenomenological experience associated with it. So premises (1–…) are true. But the last premise, the principle, would be false. For if all emeralds are grue, then certainly *p* is not associated with observing all emeralds. In fact, if all emeralds are grue, then the *p* we have associated with all the past emeralds will not be the same for the emeralds examined later than *t*. The principle is built into the schema to discriminate between predicates whose projections entail different perceptual experiences and those that do not.

4.2 Bayesian Application

With the basic thrust of the phenomenological approach in mind, let us now turn to applying the idea in the context of the four Bayesian subdivisions of the grue problem.

4.2.1 AllGreen and AllGrue: Synchronic and Diachronic

Our Bayesian analysis in §2 took for granted that the likelihoods of AllGreen and AllGrue were the same. This was because both hypotheses entailed the relevant data proposition, D, which was 'all emeralds examined have been observed to be green'. It becomes immediately obvious that, on the phenomenological approach, this is *not* the relevant data proposition. The entire thrust of the solution proposed herein is that the relevant facts to our induction are not that x-many emeralds are green, but rather that x-many emeralds have a phenomenological experience associated with their being green.

So, the data we are considering must be expanded to include the phenomenological facts associated with our observations *and* the fact that the principle in our schema is only true with respect to AllGreen.[8] Call this data proposition D', which precisely spelled out is just this: 'premises (1-…) and the principle of S_p are true for the predicate 'green' and false for 'grue".

Thus we are now interested in Pr(AllGreen | D') and Pr(AllGrue | D'). As before, by Bayes' theorem this gives us:

$$\Pr(\text{AllGreen} \mid D') = [\Pr(D' \mid \text{AllGreen}) * \Pr(\text{AllGreen})] / \Pr(D')$$

$$\Pr(\text{AllGrue} \mid D') = [\Pr(D' \mid \text{AllGrue}) * \Pr(\text{AllGrue})] / \Pr(D')$$

I have already given some reason to think that our phenomenological data, D', confers a higher credence to AllGreen rather than AllGrue. But note why this is the case: it is because the *likelihoods* of the hypotheses are now different, and not the priors. Why think that Pr(D' | AllGreen) > Pr(D' | AllGrue)? First, note that:

$$\Pr(D' \mid \text{AllGreen}) = \Pr(D' \,\&\, \text{AllGreen}) / \Pr(\text{AllGreen})$$

8. It might be contended here that the principle I constructed in the schema, which constitutes the asymmetry between green and grue, is not part of our *data*, but instead should be considered a norm that governs prior probability assignment. I think this is mistaken. It does not follow from any analytic analysis of 'green' or 'grue' that one predicate should be true or false with respect to the principle. It seems possible that our perceptual experience of the world could have been such that we view grue emeralds as the same, and green emeralds as different. That our phenomenology happens to be one way rather than another is a fact we learn a *posteriori* and should be considered part of empirical data.

$$Pr(D' \mid AllGrue) = Pr(D' \,\&\, AllGrue) / Pr(AllGrue)$$

Assuming identical priors, the question becomes: why think that Pr(D' & AllGreen) >Pr(D' & AllGrue)? Here is one argument: D', which is just the truth of the premises of S_p with respect to 'green' and their falsity with respect to 'grue', provides a good basis for inferring AllGreen, but not AllGrue. And since the conjunction of a set of premises and a conclusion that can be justifiably inferred from the premises is more likely than a conjunction of those premises with some arbitrary conclusion (of the same prior) that cannot be justifiably inferred, it follows Pr(D' & AllGreen) > Pr(D' & AllGrue). Indeed, if we have a seemingly justified schema from D' to AllGreen, but not AllGrue, the likelihood of AllGreen will be greater than AllGrue.

Much the same can be said for the diachronic comparison of the two generalizations. Assuming that Pr(D') < 1, by the same argument it follows that AllGreen has a higher likelihood, and therefore posterior probability, than AllGrue.

4.2.2 NextGreen and NextGrue: Diachronic and Synchronic

What about NextGreen and NextGrue? Starting with the diachronic condition:

(C_d): NextGreen is confirmed by data D' if and only if

$$Pr(NextGreen \,\&\, D') > [Pr(NextGreen) * Pr(D')]$$

The central issue that arose was: why think that this inequality holds *rather than* the inequality with NextGrue? There was nothing in our previous data proposition that appeared to break this symmetry. However, utilizing D', we now have a reason to think that only the above inequality holds. This is because the probabilities of the conjunctions can be reduced to a question of posterior probabilities:

$$Pr(NextGreen \,\&\, D') = [Pr(NextGreen \mid D') * Pr(D')]$$

$$Pr(NextGrue \,\&\, D') = [Pr(NextGrue \mid D') * Pr(D')]$$

And, much like earlier, there is good reason to think Pr(NextGreen | D') > Pr(NextGrue | D'). Namely, our phenomenological data confers a higher probability on the next emerald being green than it does for the next emerald being grue.

The same can be said in the synchronic case. D' gives us good reason to think that it is more likely that the next emerald is green rather than not. Note also that the arguments I gave in favor of a higher posterior for AllGreen rather than AllGrue, if successful, can provide reason for thinking the data confirms NextGreen over NextGrue, provided we assume that there is positive association between emerald colors.

5 Concluding Remarks

I have outlined how one might apply a so-called phenomenological approach to the grue problem construed in a Bayesian framework. I would like to now — even more briefly — remark on an interesting feature of this approach.

The interesting lesson seems to be this: the answer provided herein does not rely on justifying a higher prior to AllGreen or AllGrue as independent hypotheses. This is normally the principal difficulty with Bayesian treatments of the grue problem: if both predicates are interdefinable, on what basis might prior probabilities be assigned asymmetrically? By arguing instead that the data be expanded to include what phenomenological experiences applied in the past, all that

had to be justified was a different prior *relationship* the data bore to each hypothesis. Solutions to the grue problem should not aim to locate an asymmetry in Pr(AllGreen) and Pr(AllGrue) at all, but rather in the different likelihoods produced by some expanded data, D*.[9]

I have construed the expanded data along phenomenological lines, but this need not be the case. Whatever asymmetry one finds between AllGreen and AllGrue to be the more pressing concern — be it phenomenological or other — ought to be construed as creating different likelihoods for the hypotheses by virtue of some expanded set of data to consider. Different assignments of priors are unnecessary. Reconsidering our data well appears to be enough.

References

Good, Irving John. "Explicativity, Corroboration, and the Relative Odds of Hypotheses." *Synthese* 30, nos. 1-2 (1975): 39–73.

Goodman, Nelson. *Fact, Fiction, and Forecast*. Harvard University Press, 1955.

Jeffrey, Richard C. *The Logic of Decision*. New York, NY, USA: University of Chicago Press, 1965.

Shoemaker, Sydney. "Functionalism and Qualia." *Philosophical Studies* 27, no. May (1975): 291–315. https://doi.org/10.1007/BF01225748.

Sober, Elliott. "Confirmation and Law-Likeness." *Philosophical Review* 97, no. 1 (1988): 93–98.

———. "Grue!: The New Riddle of Induction." Chap. No Model, No Inference: A Bayesian Primer on the Grue Problem, edited by Douglas Stalker, 225–238. Open Court Publishing Group, 1994.

Ward, Barry. "Explanation and the New Riddle of Induction." *Philosophical Quarterly* 62, no. 247 (2012): 365–385.

9. A similarly spirited remark is found in Fitelson (2008), wherein it is suggested that solutions to the grue problem can make use of the fact that 'the evidence only confirms both hypotheses *depending on the background corpus one starts with*'(emphasis added).

Does Connectionism undermine Fodor's Language of Thought Hypothesis?

Jonathan Fryer*

University of Bristol

In 1975, Fodor hypothesised that thought is structured in much the same way as language.[1] Thoughts have semantics, a combinatorial syntax, and store information symbolically. In the 1980s, Connectionism looked to undermine his view. It suggested that mental information is stored non-symbolically in neural nets; it was considered a "paradigm shift" for cognitive theories.[2] In the 1990s, further work by Chalmers and Rowlands undermined Fodor's Language of Thought Hypothesis.[3][4][5] Modern cognitive research into Deep Learning uses an inherently Connectionist framework.

This paper separates Fodor's hypothesis from his arguments in its support. It argues that Fodor's Language of Thought Hypothesis is still a legitimate theory of cognition. However, it accepts that Fodor's arguments in favour of his hypothesis are fallacious. The paper examines three of Fodor's arguments for a language of thought: the only game in town argument, the argument from systematicity and productivity, and the argument from isomorphism.[6][7][8][9] It shows each to be flawed.

Further, this paper dismisses the dilemma Fodor and Pylyshyn present the Connectionist: that they must either merely implement his Language of Thought Hypothesis or concede that it is an inadequate theory of cognition.[10] he paper uses Chalmers' Backpropagation Model, a system that encodes grammatical information without using symbols, to escape the dilemma.[11]

Throughout, I argue that despite successfully undermining his arguments, Connectionism does not undermine Fodor's Language of Thought Hypothesis. I provide two positive reasons to upholding the Language of Thought Hypothesis. This paper concludes that – at present – neither Connectionism nor Fodor's Language of Thought Hypothesis has undermined the other.

1 Introduction

Connectionism undermines the arguments Fodor provides for a language of thought, but it does not undermine the Language of Thought Hypothesis (LOTH) itself. I distinguish between the LOTH – the thesis that thought is syntactically

*Jonathan Fryer completed his undergraduate studies at Bristol University with a First-Class Honours BA in Philosophy. His philosophical interests include the philosophy of psychology, ethics of all kinds (meta, normative, applied), and philosophy in the digital age. Having taken a year's break to recover from a COVID-filled undergraduate experience, he will begin an MPhil in Philosophical Studies in September at University College London. Jonathan considers himself "rubbish at science", and credits his interest in this area of philosophy to his lecturer at Bristol University, Max Jones.

1. Jerry A. Fodor, *The Language of Thought* (Harvard University Press, 1975).
2. Jerry A. Fodor, "Why There Still has to Be a Language of Thought," in *Psychosemantics* (MIT Press, 1987), 82.
3. D. Chalmers, "Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation" (1990).
4. David J. Chalmers, "Syntactic transformations on distributed representations," in *Connectionist natural language processing* (Springer, 1992), 46–55.
5. Mark Rowlands, "Connectionism and the Language of Thought," *British Journal for the Philosophy of Science* 45, no. 2 (1994): 485–503, https://doi.org/10.1093/bjps/45.2.485.
6. Fodor 1975.
7. Fodor 1987.
8. J. A. Fodor, "The Language of Thought," *Critica* 10, no. 28 (1978): 140–143.
9. Jerry A. Fodor and Zenon W. Pylyshyn, "Connectionism and Cognitive Architecture: A Critical Analysis," *Cognition* 28, nos. 1-2 (1988): 3–71, https://doi.org/10.1016/0010-0277(88)90031-5.
10. Fodor and Pylyshyn 1988.
11. Chalmers 1990.

structured – and the arguments Fodor provides in favour of it. My separation of hypothesis and supporting argument is vital: the arguments support the LOTH, but they are not the LOTH.

I focus on three of Fodor's arguments:

1. The 'only game in town' argument[12]
2. The argument from systematicity and productivity;[13] and
3. The argument from isomorphism.[14]

Section 1 sets out the LOTH and the three arguments. I show that the LOTH can still be true, even if all three arguments are flawed. I then set out the Connectionist challenge in Section 2,[15] using Sanderson's model program that recognises hand-written digits as an example.[16] The existence of coherent Connectionist models undermines the argument that a language of thought is the 'only game in town'. It does not undermine the LOTH.

Fodor responds to the Connectionist challenge in a paper with Pylyshyn, in which he employs (2).[17] I discuss this in Section 3. They argue that symbol manipulation — a property of classical cognitive architecture — is required to explain the nomic necessity of systematicity and productivity in thought. An adequate theory of cognition must be able to explain this. Connectionism, therefore, either merely implements classical architecture or is an inadequate theory of cognition.

For Connectionism to undermine (1) and (2), it must escape this dilemma. In Section 4, I argue that the nomic necessity requirement is unnecessarily stringent. Section 5 discusses how Chalmers undermines (2) by creating a structure-sensitive, non-implementational Connectionist model.[18] This re-establishes Eliminative Connectionism as a legitimate theory of cognition. Connectionism being a legitimate theory of cognition further undermines (1). However, it does not undermine the LOTH.

Moreover, Rowlands shows (3) to be fallacious.[19] Section 6 follows their argument that logically structured representations do not follow from an isomorphism of the causal relations between representations and the logical relations between propositions. Although Chalmers and Rowlands succeed in undermining Fodor's arguments, I argue that they do not undermine the LOTH itself.

In Section 7, I provide two positive reasons for a language of thought, before concluding that the LOTH remains a legitimate cognitive theory.

First, some clarifications. For Connectionism to undermine Fodor's LOTH, the criticism must come from Connectionists; it must be about the nature of mental states and mental processes.[20] Both parties believe that representations exist and are physicalist about brain states — they believe states and processes of the mind to be identical to states and processes of the brain.[21][22] Whilst there are some Connectionists who deny representational states — such as Churchland[23] — the majority of debate assumes their existence. Thus, I will not discuss Eliminativism about representations or anti-realism about mental states. Further, Connectionist models match what we know about the neurological structure

12. Fodor 1975.
13. Fodor, 1987; Fodor and Pylyshyn, 1988.
14. Fodor 1987.
15. Michael Rescorla, "The Language of Thought Hypothesis," in *The Stanford Encyclopedia of Philosophy*, Summer 2019, ed. Edward N. Zalta (Metaphysics Research Lab, Stanford University, 2019).
16. Grant Sanderson, "But what is neural network? Chapter 1, Deep learning," Youtube, 2017, https://youtu.be/aircAruvnKk.
17. Fodor and Pylyshyn 1988.
18. Chalmers 1990, Chalmers 1992.
19. Rowlands 1994.
20. Fodor and Pylyshyn 1988, 3.
21. Fodor 1987, 282.
22. J. J. C. Smart, "The Mind/Brain Identity Theory," in *The Stanford Encyclopedia of Philosophy*, Spring 2017, ed. Edward N. Zalta (Metaphysics Research Lab, Stanford University, 2017).
23. Churchland 1990 as found in Rescorla.

of the brain, but Connectionism on a nonrepresentational level is not relevant.[24] It is possible for a brain to be neurologically Connectionist but implement classical representational architecture.[25] The only relevant Connectionism is at the representational level. Moreover, non-Fodorian LOT theories (such as Schneider's) are not discussed in detail.[26] I include Deep Learning[27] in my definition of Connectionism, but it is not relevant to my argument, so will not be examined in detail.

2 The LOTH and Fodor's Supporting Arguments

Fodor's hypothesis is that mental states are syntactically structured, and that mental processes are syntactical operations on mental states.[28] A state's structure determines its causal role in mental processes. Thinking occurs in a mental language. To have a belief that p is to bear an appropriate relation to a mental representation whose meaning is that p.[29] This mental representation takes the form of a sentence with combinatorial syntax and semantics. For example, to have the thought "I believe that X and Y" is to bear an appropriate relation to a complex mental representation whose meaning is that "X and Y". The complex representation gets its meaning from its constituents and how they are combined: from the meaning of its atomic constituents (X, Y) and from its syntactic parts (the conjunctive, and). As such, thought is combinatorial and structure sensitive.[30]

Distinct from the LOTH are the arguments that Fodor presents to support it. In The Language of Thought (1975), Fodor provides (1), which has widely become known as his 'only game in town' argument. He notes that our only remotely plausible cognitive theories of decision-making, concept learning and perception require a representational system to be coherent.[31] Representation presupposes a medium of representation, and a medium of representation requires symbolisation. Symbolisation requires symbols and thus a LOT.[32]

He supports this conclusion in *Why There Still Has to be a Language of Thought (1987)* with (2): our linguistic capacities are productive and systematic. Language is productive: we can conceivably say infinitely many unique and new thoughts, despite our finite physical resources.[33] This can be explained if thought is combinatorial — we can combine the constituents of sentences in as many ways as we would like. For example, *"I believe that it is very warm"* is distinct from *"I believe that it is very, very warm"* and *"I believe that it is very, very, very… ad infinitum… warm".* Even with the seven words used above, an infinite number of different mental sentences might be constructed. and our ability to understand some sentences means we understand others.

Productivity might be denied, since it requires idealisation – we never actually use any more than a finite part of any mental capacity, so our mental capacities might not necessarily be infinite. Fodor acknowledges this, but thinks idealisation is justified if it leads to independently well confirmed theories.[34] Systematicity, though, does not require idealisation, so this objection is not the focus of this essay.

Language is systematic: the ability to produce and comprehend some thoughts is intrinsically connected to the ability to produce and comprehend many other thoughts. If you understand the sentence *"Mary loves John"*, then you understand the sentence *"John loves Mary"*.[35] Our linguistic capacities are productive and systematic because they have combinatorial structure. Thought is also productive and systematic; this must be because it too has combinatorial structure. I will discuss how he uses (2) to try and undermine Connectionism later.

24. Cameron Buckner and James Garson, "Connectionism," in *The Stanford Encyclopedia of Philosophy*, Fall 2019, ed. Edward N. Zalta (Metaphysics Research Lab, Stanford University, 2019).
25. Fodor and Pylyshyn 1988, 6.
26. Susan Schneider, *The Language of Thought: A New Philosophical Direction* (MIT Press, 2011), Section 7.
27. Buckner and Garson.
28. Jerry A. Fodor, "Connectionism and Cognitive Architecture," Youtube, 2018, 10:49-11:02, https://youtu.be/vyrn1JWgqFA.
29. Rescorla.
30. Fodor and Pylyshyn 1988, 8.
31. Fodor 1987, 31 (decision-making); 36 (concept-learning); 51 (perception).
32. Fodor 1975, 55.
33. Fodor 1987, 292.
34. Fodor 1987, 293.
35. Fodor 1987, 294.

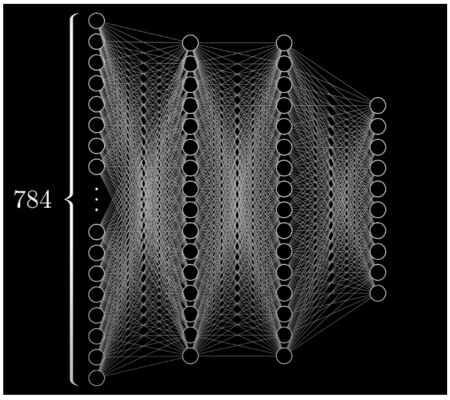

Figure 1: A Neural Network

Fodor presents (3) in Propositional Attitudes (1978) by pointing to the isomorphism of causal relations between representations and logical relations between propositions.[36] One can map the causal relations between representations onto a set of logical relations between propositions without losing the meaning of the representations. The propositions index the representations. From this isomorphism, he concludes that the representations must have logical form.

Fodor uses these three arguments to establish his LOTH. However, they are distinct from it – even if the arguments turn out to be invalid and/or unsound (as they do), this does not affect the truth of the LOTH. Analogously, I might present a fallacious argument that 'proves' that grass is green; the issues with my argument do not alter the fact that grass is green. I will show that Connectionism undermines (1)-(3), but it has not undermined the LOTH.

3 The Connectionist Challenge

Connectionism is not one idea or hypothesis, but a vast range of ideas. There is, though, a general form of Connectionism that threatens to undermine Fodor's. It offers an alternate theory of cognitive processing, using a different account of representation, mental states, and mental processes. Thus, it threatens Fodor's hypothesis that mental states are syntactically structured.

Connectionism claims that cognitive functioning can be explained by collections of units in a neural network. *Figure 1* provides a simple example of a neural network designed to learn to recognise hand-written digits.[37]

There are three types of units (or neurons): input units, hidden units, and output units. The units are organised into layers.[38] *Figure 1* has an input layer of 784 neurons, two hidden layers of 16 neurons and an output layer of 10 neurons.

The 784 neurons in the input layer (1st left to right in *Figure 2*) correspond to the 784 pixels on a 28x28 computer screen. When an image of a hand-written digit is on the screen, each pixel lights up in a certain way. Each neuron represents the greyscale value of its corresponding pixel as a number between 0 and 1. This is the neuron's activation value. The output layer has 10 neurons, representing the digits 0-9. The activation value of these neurons represents how much the system 'thinks' that a given image corresponds with a given digit.[39] Every neuron in the first hidden layer is connected to all 784 neurons from the input layer. Every neuron in the second hidden layer is connected to all 16 in the first, and each in the output layer connected to all 16 in the second.

36. As found in Rowlands 2019, 492.
37. Sanderson 2017a 03:47, based upon Nielsen 2015, 13.
38. Buckner and Garson.
39. Michael A. Nielsen, *Neural networks and deep learning*, vol. 25 (Determination press San Francisco, CA, USA, 2015), 14.

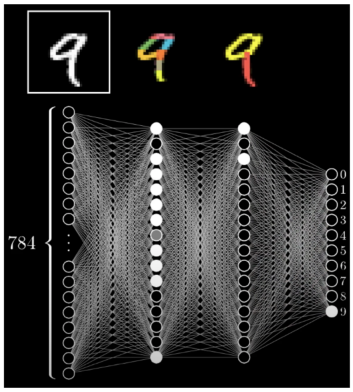

Figure 2: The Layers

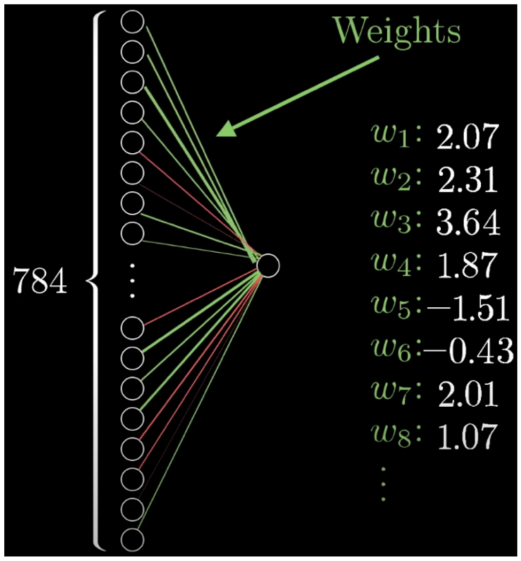

Figure 3: Weighted Connections

The 16 neurons in each of the hidden layers represent a subcomponent of a hand-written digit.[40] For example, the number 9 is a loop on top of a line; the number 8 is one loop on top of another. The rightmost hidden layer (3rd) represents these subcomponents. The leftmost hidden layer (2nd) represents subcomponents of these subcomponents. For example, a loop is composed of various small lines. *Figure 2* demonstrates what each layer might represent for an image of the number 9.[41]

Each neuron can be thought of as a function, taking the outputs of all the previous neurons, and producing a number between 0 and 1. In the first hidden layer, the activation value of each neuron is determined by the activation values of all 784 input units. As well as the activation value of each input unit, we also assign a value to the connection between x and each of the 784 neurons from the first layer, as shown in *Figure 3*.[42] This value is called the weight of a neuron's connection.

The network learns to recognise hand-written digits by finding the right weights and biases to produce the greatest activation value at the output unit for the correct digit. This process is loosely analogous to biological networks,

40. Nielsen, 11.
41. Grant Sanderson, "What is backpropagation really doing? Chapter 3, Deep learning," Youtube, 2017, 07:40, https://youtu.be/Ilg3gGewQ5U.
42. Nielsen, 14.

where neurons firing causes other neurons to fire.[43] Network learning methods generally fall under two categories: supervised and unsupervised.[44] The difference between the two is that supervised learning has an element of human oversight. Chalmers' model uses backpropagation, a supervised learning method. I will explain this process in Section 5, when discussing Chalmers' model.

If Connectionism is a legitimate theory of cognition, (1) seems undermined. Upon the relevant Connectionist models, representations are not operations over symbols. Information is stored non-symbolically in the weights of connections between the units in a neural net. There is another game in town. It designs systems that exhibit intelligent behaviour without retrieving or operating upon structured symbolic expressions.[45]

If Eliminative Connectionism is shown to be the correct theory of cognition, then the LOTH would be undermined. Eliminative Connectionism is Connectionism whose models do not implement a LOT. There are also Connectionist models that operate upon symbols.[46] These models implement the LOTH upon Connectionist hardware. Thus, even though Eliminative Connectionism undermines (1), the existence of Implementational Connectionism means the LOTH is not undermined. The existence of another possible explanation does not disprove the LOTH.

Nevertheless, in the 1980s, Connectionism looked to be a "paradigm shift" for cognitive theories – the LOTH seemed under threat.[47] Fodor and Pylyshyn responded to this threat.[48]

4 Fodor and Pylyshyn's Response: The Connectionist's Dilemma

F and P adapt the argument from systematicity and productivity (2) into an argument against Connectionism.[49] They argue that a theory of mental computation is explanatorily adequate only if it explains the nomic necessity of systematicity and productivity in thought. Symbol manipulation is the only way to explain the nomic necessity of systematicity and productivity. Whilst there might be Connectionist models that are systematic and productive (for example, ones that implement classical/LOT architectures), Connectionism does not require these qualities. This is an issue because systematicity is a necessary quality of thought. Thus, the Connectionist is presented with a dilemma: to endorse symbol manipulation, making Connectionism nothing but a way to implement a LOT, or to reject symbol manipulation. In rejecting it, they would be unable to explain the nomic necessity of systematicity and productivity; Connectionism would be an inadequate theory of cognition.

Accepting an implementational role is a real option for Connectionists in terms of accurately accounting for the nature of mental states and processes. However, Connectionism cannot undermine Fodor's LOTH if it implements it. Implementational theories – for example, Marcus, who argues for neural networks that implement symbol manipulation[50] – are not useful for undermining the LOTH. If the Connectionist cannot escape this dilemma, (1) is no longer undermined – the LOTH would be the only coherent option. Eliminative Connectionism must escape the charge of inadequacy by accounting for productivity and systematicity.

43. Sanderson 2017a, 04:32-05:00
44. Buckner and Garson.
45. Fodor and Pylyshyn 1988, 2.
46. Chalmers 1990, 341.
47. Fodor 1987, 82.
48. Fodor and Pylyshyn, 1988.
49. Reconstructed in Rescorla.
50. Gary F. Marcus, *The Algebraic Mind: Integrating Connectionism and Cognitive Science* (MIT Press, 2001).

5 Escaping the Dilemma

The dilemma presented by F and P is unnecessarily stringent by requiring systematicity and productivity as a nomic necessity. There are classical architectures that lack systematicity and productivity.[51] Therefore systematicity and productivity cannot be a nomic necessity for classical architectures. By their own requirement, then, the LOT would be an inadequate theory of cognition. Thus, it seems fair to drop the requirement that productivity and systematicity be necessary. Connectionist models still need to be productive and systematic, though.

F and P's paper roused numerous responses (Clark 1989; Smolensky 1987, 1990; van Gelder 1990; Elman 1990; Pollack 1990).[52] For this question, the only useful response to F and P is to produce an Eliminativist Connectionist model that explains systematicity and productivity. Therefore, I focus on Chalmers, who produces a non-classical model of cognition that operates structure-sensitive processes.[53] Others have also attempted this – notably, Pollack (1988, 1990) and Smolensky (1987, 1990).[54] However, they provide productive, systematic models by extracting the original constituents of a representation. This effectively renders their models implementational and, as such, not useful for undermining the LOTH.[55]

The success of Chalmers' model would undermine (1). It would also undermine (2) in its employment as a counterargument against Connectionism.[56]

6 Chalmers' Eliminativist Connectionist Model

Chalmers argues against F and P's claim that Connectionist models cannot support systematic operations in a non-classical way. He claims that F and P misrepresent the Connectionist endeavour and underestimate the difference between localist and distributed representations.[57]

When describing Connectionism, F and P provide an example of a localist Connectionist model where atomic symbols are represented by single nodes, connected by associative links.[58] They briefly assert that it would change nothing if these nodes were replaced by a distributed pattern of activation.[59] Chalmers, though, demonstrates that there is a clear and significant difference between localist and distributed representation. In a localist model, nodes represent atomic symbols. They are connected with associative links. F and P representing Connectionism in this way is problematic – many Connectionists define themselves on attempting to do away with the atomic symbol in theories of meaning.[60] Distributed models do not use atomic symbols. They have groups of separately functioning nodes that have functional properties far beyond that of an isolated unit. Representation does not occur at the level of the node, but at a much higher level. Information is stored in the activation values of nodes and weights of connections. At that higher level, patterns of activation between nodes combine compositionally and autonomously to produce distributed, malleable representations. Thus, a small difference in the activity of a subset of nodes can cause substantial differences in later processing.[61]

F and P claim that Eliminativist Connectionist models cannot account for productivity or systematicity. In other words, they cannot carry out structure-sensitive operations. Responding to this criticism is paramount for Connectionism to be able to undermine Fodor's LOTH. Chalmers responds by undergoing a series of experiments, in which

51. Buckner and Garson.
52. Found in Chalmers 1990, 340, and Terence E. Horgan and John L. Tienson, *Connectionism and the Philosophy of Mind* (Kluwer Academic Publishers, 1991)
53. Chalmers 1990, 1992.
54. Found in Chalmers 1990, 344.
55. Chalmers 1990, 340.
56. (2) will be fully undermined in section 5
57. Chalmers 1990, 340.
58. Fodor and Pylyshyn 1998, 10.
59. Fodor and Pylyshyn 1998, 15.
60. Chalmers 1990, 343.
61. Chalmers 1990, 343.

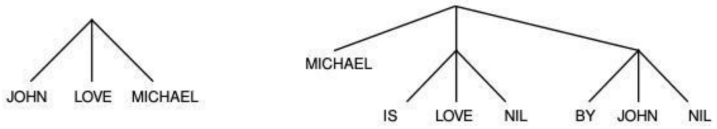

Figure 4: Sentence Representations

he demonstrated that structure-sensitive operations are possible upon distributed representations.[62]

The experiments looked to use distributed representations to transform sentences from their active to their passive forms. Chalmers combined five different names and verbs to produce 250 different sentences, all of similar syntactic form to the active "John loves Michael", or the passive "Michael is loved by John". He used Pollack's Recursive Auto-Associative Memory (RAAM),[63] a system that recursively encodes symbolic tree structured representations of sentences in distributed form.[64] *Figure 4* shows how the sentences were given syntactic structure.

RAAM uses backpropagation to create patterns for each of the internal nodes of the trees. Backpropagation is a form of supervised learning. Whilst untrained, a neural net might produce activation values at the output nodes that are wildly inaccurate. Backpropagation is an algorithm that computes a list of changes required to the weights and biases to produce the correct results.[65] It compares the net's outputs to the 'correct' outputs provided by a training data set and works backwards, seeing how the weights and biases of the connections from hidden and input layers have led to the 'incorrect' values of the output nodes. Over many training cycles, backpropagation fine-tunes the weights and biases of the connections between nodes, until the network produces the 'correct' outputs. As the theory goes, during this process the network generalises the syntactic rules of the operations it learns.

Chalmers' RAAM assigned each word a primitive localist representation and learned to represent all 250 sentences. Once encoded in the RAAM, the representation is considered distributed. 150 of these encodings were randomly selected to train the Transformation Network, another backpropagation network. The Network was to take an encoded distributed representation of an active sentence as input and transform it into the appropriate encoded distributed representation of a passive sentence as output. The RAAM then decoded the represented output sentences. To truly see whether the network was structure-sensitive, it needed to successfully operate on sentences outside of its training data set – this would test whether the network had generalised the syntactical rules it was being fed.

Thus, after the Transformation Network was trained, the RAAM encoded the other 50 active sentences, fed them through the Transformation Network, and then decoded the Transformation Network's output pattern. In all 50 cases, the output pattern decoded to the correct passivized sentence; generalisation rate was 100 per cent. As noted by Chalmers (1990), this shows that distributed representations formed by RAAM can effectively facilitate structure-sensitive operations in a non-classical way.[66] If Chalmers' model can account for structure-sensitive operations, then it can explain why understanding the sentence "Mary loves John" entails understanding "John loves Mary". It can account for systematicity and, it follows, productivity. Thus, if Chalmers' model is satisfactory, then F and P's dilemma is escaped: eliminativist Connectionism is explanatory of systematicity and productivity. (1) and (2) are undermined.

The effectiveness of RAAM and Backpropagation Models, though, is questionable. Buckner and Garson note that such models fall short when they are applied to truly novel sentences.[67] Marcus argues that multilayer perceptron approaches that backpropagation cannot capture the flexibility and power of everyday reasoning.[68]

Debate regarding the fine-tunings and legitimacy of RAAM and Backpropagation Models is beyond the scope of this essay. What is important, though, is not that Chalmers' model is perfect, but that it is another game in town.

62. Chalmers 1990, 1992.
63. Chalmers 1990, 5.
64. Seth Rait, "DRAAM: Deep (Recursive Auto-Associative Memory) And Applied eMbeddings," Undergraduate Honors Thesis (Undergraduate Honors Thesis, Brandeis University, 2018), 12.
65. Sanderson, "But what is neural network? Chapter 1, Deep learning."
66. Chalmers 1992.
67. Buckner and Garson.
68. Marcus, 169.

There are many other Eliminativist Connectionist models that have also been purported to explain systematicity and productivity of thought. Moreover, new research into Deep Learning has opened up the opportunity for further discovery down the line. Rait (2018) argues that a new Deep Learning RAAM (DRAAM) could provide novel insights into cognitive processing, since the main issue with RAAM at Pollack's time of writing was its technical limitations.[69] Loula, Barni and Lake (2018) report that their nets qualified as demonstrating strong semantic systematicity.[70] Whether any of these models are without criticism is less important for this essay – the nature of mental representations is still an undetermined issue, but Eliminativist Connectionism remains a game in town.

(1) is certainly undermined. (2) might still stand. Whilst Eliminative Connectionism is a legitimate prospect, systematicity and productivity might be evidence of the syntactic structure of thought. This still does not undermine Fodor's LOTH. Connectionism provides an accurate neurological account of the brain which should be adopted, but this does not preclude the possibility of Connectionist models implementing the LOTH. The LOTH is one of the two current theories of high-level cognition that might be true. I now turn to Rowlands, who attempts to directly undermine Fodor's reasoning behind the LOTH.[71] I argue that Rowlands' successfully shows (2) and (3) to be fallacious. However, he also fails to undermine the LOTH itself.

7 Rowlands' Critique of the LOTH

The LOTH makes two distinct claims:

(C1) Mental representations are structured entities

(C2) Mental representations have the structure of propositions or sentences.[72]

Rowlands argues that all the arguments that Fodor provides to support and motivate the LOTH are based upon a fallacious conflation of (C1) and (C2). Fodor assumes that the arguments prove (C2) when they only prove (C1). This fallacy is clear in the argument from productivity and systematicity (2), and the argument from isomorphism (3). Rowlands focusses on (3).

As mentioned earlier, (3) points to the isomorphism of causal relations between representations and logical relations between propositions.[73] One can map the causal relations between representations onto a set of logical relations between propositions without losing the structure of the representations. The propositions index the representations. Fodor argues that this structure-preserving mapping must be specifiable in terms of the logical form of propositions and concludes that the objects of these attitude must have logical form. Rowlands shows this inference to be a fallacy using the analogy of a painting[74]:

A painting has many features. Conceivably, all of its features can be put into a structure-preserving representation theorem that maps the features of the picture onto a set of propositions. This set of propositions would preserve the structure of the painting. We could therefore say that the propositions index the features of the picture, in just the same way that the propositions index the causal relations between representations. Nobody, though, would imply that a painting has logical, syntactical structure, merely because we can map its features on a set of propositions. It is possible to make any system isomorphic with another if you find the right mathematical function. Isomorphism does suggest (C1) – the representations do have structure. However, it does not suggest (C2) – isomorphism does not entail logic structure.

(2) commits the same fallacy.[75] It points to certain features of natural language that are mirrored in thought (productivity and systematicity) and argues that this mirroring must be because mental representations have the same

69. Rait, 2.
70. As found in Buckner and Garson.
71. Rowlands.
72. Rowlands, 489.
73. Rowlands, 493-494.
74. Rowlands, 491.
75. Rowlands, 494.

structure as propositions. Fodor is aware of this: after re-stating (2) in *Why There Still Has to Be A Language of Thought*, he notes that one might accuse him of affirming of the consequent; he waves it off as inference to the best explanation.[76] I argue that this dismissal of fallacy would be reasonable if (1) still stood. If the LOTH was the only game in town, then noting that thought has a shared property with language and concluding that this is because they have the same structure would be a reasonable inference to make. However, it has been shown that the LOTH is not the only game in town. Thus, arguments (2) and (3) unjustly make this inference. Fodor's arguments are demonstrably undermined.

8 Arguments in Favour of a LOTH

The purpose of this section is not to present a conclusive argument in favour of Fodor's LOTH, nor one against Connectionism. Instead, it looks to present two positive reasons for maintaining the language of thought as a legitimate cognitive theory.[77]

8.1 Brains and Neural Networks learn differently

The way neural networks learn is only loosely analogous to biological networks in the brain; there are significant differences between the two. The example in Section 2 used the MNIST data set, a data set of scanned images of handwritten digits created by the United States' National Institute of Standards and Technology.[78] The MNIST data set contains 60,000 images as its training data, and 10,000 images as test data.[79] Other networks use billions of training examples.[80] Human minds do not learn in this way – we do not consult billions of pieces of training data before being able to recognise hand-written digits. It would not physically be possible for brains to do so.

Chomsky's Poverty of the Stimulus argument states that children learn with far less data than something like a neural net requires.[81] Auxiliary verbs are a classic example of this: they are highly syntactically complex, and their employment in language often defies generalisation.[82] There are 1x1022 combinations of English auxiliary verbs, yet only 99 grammatically possible combinations.[83] A human brain could not run through a data set that large to learn the complex rules of auxiliary verbs, as 1x1022 is roughly one hundred billion times the number of neurons in the human brain.[84] Nevertheless, children consistently differentiate between auxiliary and lexical verbs without issue.[85] Despite achieving the same competence as a human brain in specific tasks, neural networks compute in a fundamentally different way.

It is worth noting that Chomsky's Poverty of the Stimulus argument is philosophically controversial, and the debate surrounding it is well beyond the scope of this essay.[86] That debate is further muddled by Fodor's belief that the language of thought is innate.[87] Nevertheless, an important point has been raised. Much of the "paradigm shift"[88] away from the LOT occurred because Connectionism was able to produce models that mirrored biological processes; there are significant differences between human learning and machine learning that Connectionism is yet to reconcile.

76. Fodor 1987, 293.
77. For a series of criticisms of Connectionism, see Marcus.
78. Nielsen, 15.
79. Nielsen, 16.
80. Nielsen, 2.
81. Noam Chomsky, *Poverty of Stimulus: Unfinished Business*, March, 5.
82. Stephen Laurence and Eric Margolis, "The Poverty of the Stimulus Argument," *British Journal for the Philosophy of Science* 52, no. 2 (2001): 226, https://doi.org/10.1093/bjps/52.2.217.
83. Stromswold as found in Laurence and Margolis, 224.
84. Laurence and Margolis, 224.
85. Laurence and Margolis.
86. Laurence and Margolis.
87. Laurence and Margolis, 240.
88. Schneider, *The Language of Thought: A New Philosophical Direction*, 82.

8.2 The LOTH is still pursued

Whilst this essay is not an examination of non-Fodorian language of thought theories, it is worth mentioning that modern studies continue to build upon the belief that thought is logically structured. A language of thought accounts for the productivity and systematicity of thought in simple terms. Symbol-manipulation still provides the best explanation for high-level cognitive phenomena.[89] Further, modern LOT theories, such as Schneider's, offer the LOT a "philosophical overhaul" to separate Fodor's hypothesis from his fallacious arguments, keeping symbol manipulation at the forefront of cognitive theories.[90]

Implementational Connectionism – Connectionism that endorses symbol manipulation and implements classical architectures – remains a genuine field of philosophical inquiry.[91] Schneider also notes that information-processing psychology predominantly operates with symbol-processing models.[92] The LOT is still considered a relevant theory of cognition and is being used to further our understanding of the mind and brain.

9 Conclusion: Fodor's LOTH is not undermined

I have discussed three of Fodor's arguments in favour of the LOTH. Chalmers and Rowlands have convincingly demonstrated each to be unconvincing. However, none of this undermines Fodor's LOTH. Chalmers and Rowlands have showed that Fodor's arguments do not prove that thought is logically structured. This is not the same as showing that thought cannot be logically structured. There are positive reasons to uphold the LOTH.

This question is unfairly stacked against Fodor – his work is singular and complete, whereas Connectionism is a dynamic field of advancing scientific understanding. Thus, there will inevitably come a point where Fodor's writing seems dated and out of touch with modern science. Already, Connectionism acts as a strong non-representational framework within which to build theories of high-level cognition. However, Fodor's hypothesis is yet to be disproved.

It is entirely possible that some future scientific discovery – perhaps in *Deep Learning* models – proves Eliminativist Connectionism to be the only satisfactory theory of cognition and representation. Equally, some breakthrough might champion Implementational Connectionism. Neither has happened yet. Despite the flaws in his arguments, Fodor's LOTH is still a game in town; the LOTH is still being pursued. Thus, Connectionism does not undermine Fodor's Language of Thought Hypothesis. Equally, Fodor's *Language of Thought Hypothesis* does not undermine Eliminative Connectionism. At least not yet.

References

Buckner, Cameron, and James Garson. "Connectionism." In *The Stanford Encyclopedia of Philosophy*, Fall 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2019.

Chalmers, D. "Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation." 1990.

Chalmers, David J. "Syntactic transformations on distributed representations." In *Connectionist natural language processing*, 46–55. Springer, 1992.

Chomsky, Noam. *Poverty of Stimulus: Unfinished Business*, March.

Fodor, J. A. "The Language of Thought." *Critica* 10, no. 28 (1978): 140–143.

89. Marcus, 170.
90. Schneider, 5.
91. Chalmers 344 gives Pollack (1988, 1990) and Smolensky (1987, 1990) as examples
92. Schneider, 4.

Fodor, Jerry A. "Connectionism and Cognitive Architecture." Youtube, 2018. https://youtu.be/vyrn1JWgqFA.

———. *The Language of Thought*. Harvard University Press, 1975.

———. "Why There Still has to Be a Language of Thought." In *Psychosemantics*. MIT Press, 1987.

Fodor, Jerry A., and Zenon W. Pylyshyn. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28, nos. 1-2 (1988): 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.

Horgan, Terence E., and John L. Tienson. *Connectionism and the Philosophy of Mind*. Kluwer Academic Publishers, 1991.

Laurence, Stephen, and Eric Margolis. "The Poverty of the Stimulus Argument." *British Journal for the Philosophy of Science* 52, no. 2 (2001): 217–276. https://doi.org/10.1093/bjps/52.2.217.

Marcus, Gary F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press, 2001.

Nielsen, Michael A. *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA, USA, 2015.

Rait, Seth. "DRAAM: Deep (Recursive Auto-Associative Memory) And Applied eMbeddings." Undergraduate Honors Thesis. Undergraduate Honors Thesis, Brandeis University, 2018.

Rescorla, Michael. "The Language of Thought Hypothesis." In *The Stanford Encyclopedia of Philosophy*, Summer 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2019.

Rowlands, Mark. "Connectionism and the Language of Thought." *British Journal for the Philosophy of Science* 45, no. 2 (1994): 485–503. https://doi.org/10.1093/bjps/45.2.485.

Sanderson, Grant. "But what is neural network? Chapter 1, Deep learning." Youtube, 2017. https://youtu.be/aircAruvnKk.

———. "What is backpropagation really doing? Chapter 3, Deep learning." Youtube, 2017. https://youtu.be/Ilg3gGewQ5U.

Schneider, Susan. *The Language of Thought: A New Philosophical Direction*. MIT Press, 2011.

Smart, J. J. C. "The Mind/Brain Identity Theory." In *The Stanford Encyclopedia of Philosophy*, Spring 2017, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2017.

What Makes Thoughts about Specific Things?

Hugo Heagren*

University of Cambridge

Some thoughts have 'objects'—things those thoughts are about. Answers to questions about the relation between thoughts and their objects often appeal to a distinction between singular and general thoughts. Singular thoughts are supposed to have somehow more particular or specific objects, general thoughts less so. I argue that no such distinction exists, and that though one could be constructed this would not be philosophically useful. §1 surveys views on the nature of the singular/general distinction. §2 lists three problems with this distinction. Consideration of these problems leads to a finer-grained distinction between singular and general concepts in §3, and I motivate this with examples and methods of argument from literature in §4. In the last section I consider a possible objection. I argue that though there is a sense in which it is technically correct, it does not achieve anything philosophically useful.

1 Introduction

Some thoughts are about things, especially existing things in the world, like books, the Moon and my father. What is it for a thought to be about a thing, and how — if at all — does the thought relate to what it is about? Answers to these questions have tended to distinguish different types of thought according to the sorts of things they are about, considering each type separately

The distinction between singular and general thoughts is often employed in this way. Taylor, for example, argues that "without an account of the inner form of singular thoughts we will be at a loss to understand how singular thought . . . achieve[s] semantic contact with objects".[1] Similarly, Bach writes that 'we need an account of how thoughts are about their objects'[2] and immediately gives an account of singular thought. For him, an account of singular thought is required for an explanation of how thought relates to the world. This is a good strategy. Certain relations only apply to certain kinds of relata. It makes sense to clarify the nature of the relata before asking about the relation.

Singular thoughts are supposed to have somehow more particular or specific objects, general thoughts less so. I argue no such distinction exists between the thoughts these philosophers consider, and constructing one is philosophically useless. I propose a finer-grained distinction between singular and general concepts, motivated by examples and methods of argument ready to hand in the literature.

§1 surveys views on the nature of the singular/general distinction. §2 lists three problems with this distinction. First, that it cannot determine whether certain thoughts (even apparently exemplar singular thoughts) are singular or general. Second, that the singular/plural distinction, which distinguishes thoughts in a similar way to singularity/generality,

*I was awarded a philosophy degree from the University of Southampton in 2021, and will graduate this summer from the MPhil Philosophy at Cambridge. I work mostly in philosophy of language, perception and mind, especially on issues at their intersection concerning how humans access and interact with their surroundings. After my MPhil, I hope to continue in academia with a PhD.

1. Kenneth Taylor, "On Singularity," in *New Essays on Singular Thought*, ed. Robin Jeshion (Oxford University Press, 2010), 77.
2. Kent Bach, "Getting a Thing Into a Thought," in *New Essays on Singular Thought*, ed. Robin Jeshion (Oxford University Press, 2010), 39.

cannot adequately distinguish plural from superplural thoughts. Finally, that singularity of thought is often explained in terms of reference, but it is implausible that thoughts refer. Considering the last problem leads naturally to an alternate view, sketched in §3. This solves the other problems and is supported by (occasionally con- fused) approaches to the issue in the literature, discussed in §4. §5 considers a possible objection that an account of singular thought could be reconstructed from my account. While possible, I argue this would be arbitrary and of little theoretical use.

2 Singular Thought

Singular thought is understood in two ways in the literature, corresponding with Sainsbury's labels *internal* and *external* singularity.[3] A thought is externally singular if and only if 'there [exists] an object which the thought is about'. This is relational. Thoughts about the Moon are externally singular because it exists. Thoughts about Vulcan (the planet Le Verrier thought he had discovered) cannot be, because it does not exist.

Internally singular thoughts are those which "recruit resources of a kind appropriate to external singularity"[4]. For example: Jack wants a sloop called *The Mary Jane*. The Mary Jane never exists, but Jack imagines her in enough detail that any real boat fulfilling his description would be The Mary Jane. Sainsbury says that in thinking about The Mary Jane, Jack uses a concept of the same type he would if The Mary Jane existed, and he were thinking about that thing in particular. Jack thinks *as if* he is thinking about a particular existing boat — though he is not. Sainsbury thinks that internal singularity is the common internal form of externally singular thought. Being internally singular is merely a matter of the internal form of a thought, and all externally singular thoughts have this form. However, he also says there are thoughts of this form which are not externally singular — *merely* internally singular. Internally singular thoughts are thoughts which are 'as if' they are about particular existing things.

Some theories identify singular thought with externally singular thought. McDowell, for example, defines singular thought as "not . . . available to be thought or expressed if the relevant object, or objects, did not exist".[5] Jeshion's view is similar.[6]

At least for my purposes, singular thought is a theoretical or taxonomic classification, useful in investigating how thoughts relate to their objects. If singular thought were simply *defined* in terms of how it so relates, it would be theoretically useless. Those working with such theories are using the same terminology I am, but for a different task. They are investigating the nature of externally singular thought — for example, whether all externally singular thoughts are internally singular. I am concerned with the (alleged) nature of internally singular thought — I argue that the distinction between internally singular and general thoughts does not exist. Even if it were true that all externally singular thought is internally singular and vice versa, these are still distinct inquiries. Thus this essay concerns only internal singularity.

Crane's definition is similar to Sainsbury's internal singularity. Following Quine's distinction of singular and general terms by their grammatical role, singular thoughts are those with the cognitive role associated with thoughts which refer to just one object.[7] This includes thoughts which actually do so refer, and thoughts which only seem to, as in Sainsbury's account. Crane gives an example of singular thought:

1. *that man stole my wallet*[8]

He contrasts this with the general:

3. R. M. Sainsbury, "Intentionality Without Exotica," in *New Essays on Singular Thought*, ed. n Robin Jeshio (2010), 300.

4. Sainsbury, 300.

5. John McDowell, "Truth-Value Gaps," in *Logic, Methodology and Philosophy of Science VI: Proceedings of the Sixth International Congress of Logic, Methodology, and Philosophy of Science*, ed. L. J. Cohen (North Holland Publishing Co, 1982), 304.

6. Robin Jeshion, "Introduction to New Essays on Singular Thought," in *New Essays on Singular Thought*, ed. Robin Jeshion (Oxford University Press, 2010), 1.

7. Tim Crane and Jody Azzouni, "Singular Thought," *Aristotelian Society Supplementary Volume* 85, no. 1 (2011): 22, https://doi.org/10.1111/j.1467-8349.2011.00194.x.

8. Italics indicate the sentence expressing a thought. *'That man stole my wallet'* abbreviates 'the thought naturally expressed by the sentence 'that man stole my wallet"

2. *someone stole my wallet*[9]

Azzouni agrees with Crane.[10] His examples are *Bertrand Russell was born in 1872* (singular) and *Anyone who invented the theory of descriptions, co-wrote a work in mathematical logic, and was the only one to do these things, was born in 1872* (general).

In general, then, singular thoughts are thoughts which *seem* to refer to a particular, or *would* so refer if the relevant object existed. Crane sums this up: singular thoughts are where one "[has] some specific object or objects in mind"[11] Crane and Azzouni's examples are typical: most exemplar singular thoughts are expressed in full sentences. Thoughts like 1 or 2 are paradigm cases.

3 Problems for the Standard Distinction

Trying to distinguish between singular and general thoughts is problematic in at least three ways:

3.1 Multiple Subjects

First, even apparently paradigmatic singular thoughts have multiple subjects, and it is unclear which fixes the thought's status as singular or general.

Crane presents 1 s a clear example of singular thought. "[W]hen I think that *man stole my wallet*, I am 'aiming' in thought at just one object" [12]. This is true: I am thinking about *that man*, and if I uttered 'that man stole my wallet' I would intend to refer to him.

As well as *that man,* Example 1 is also about *my wallet* — also a single, particular object. This might look problematic if singularity is a matter of being about only one object, and Example 1 is about two. Not so. Crane and others recognise that singularity is not a matter of the total number of things thought about, but of their each being thought of "specifically", "particularly", or "individually."[13] 1 is not problematic because all of its subjects are singular.

But then it is unclear whether Example 2 is singular. Being about *someone*, it appears general, but being about *my wallet*, it appears singular. But the distinction is meant to be exclusive: no thought is both singular and general.

Similarly, Crane suggests elsewhere that a singular thought is just one which is typically expressed with a sentence containing a singular term[14]. If so, then 2 is singular, as it would contain 'my wallet', but Crane presents it as clearly general.

So Crane's classification of 1 as singular and 2 as general only makes sense when considering just the grammatical subjects of the sentences used to express them. The general problem is that an apparently singular thought will often be expressed by a sentence with singular and general terms. It is not clear which term is relevant to the distinction.

Crane might respond that 'someone' is the (grammatical) subject of 'someone stole my wallet', and only the subject of the sentence expressing a thought is relevant to singularity. But this does not solve the problem. It is still unclear whether Example 2 is singular of general. The same thought as Example 2 could equally be expressed by 'my wallet was stolen by someone', the subject of which is 'my wallet'. So a single thought can appear singular expressed one way,

9. Crane, "Singular Thought", 23.
10. Jody Azzouni, "Singular Thoughts (Objects-Directed Thoughts)," *Aristotelian Society Supplementary Volume* 85, no. 1 (2011): 45.
11. Tim Crane, *The Objects of Thought* (Oxford: Oxford University Press, 2013), 141.
12. Crane 2011, p. 23
13. Crane 2013, p. 141; Azzouni 2011.
14. Crane 2013, p. 138.

general another, even if singularity is determined by the subject of the expressing sentence. To defend the singular/general distinction amongst thoughts, Crane could accept that one of these expressions of the thought is canonical such that only its subject determines the thought's singularity. This seems unlikely to me — surely 'someone stole my wallet' and 'my wallet was stolen by someone' express the same thought?

Furthermore, Crane himself seems to support a view like that I have argued — that none of the objects of a thought is privileged in determining its singularity. He appeals to a very loose notion of 'aboutness' in his account of singularity [15]. Singular thought is sometimes just said to be thought about a particular thing or things[16]. But a thought like *someone stole my wallet* is clearly about someone *and* my wallet. It seems arbitrary to choose one of the objects of the thought as primary.

Finally, even if the grammatical subject of the sentence expressing a thought is privileged, the singularity of the subject itself might be unclear. *There is a citizen of France who is bald* is about *a* citizen of France (no particular one), so appears general. But surely it is also about France (in the loose sense of aboutness described above), with reference to which the general citizen is conceptualised. France is a particular, so it appears singular.

In Examples 1 and 2, it was unclear whether the sentence was singular or general, because one term suggested singularity and the other generality. Here it is unclear whether the (constructed) term itself is singular or general. Even if the subject of a sentence expressing a thought is privileged, the singular/general distinction still cannot deliver a determinate answer for the status of some thoughts.

3.2 Superplurals

The second problem concerns plurality. In explaining what it is for a thought to be about a certain thing or things, a distinction is often recognised between singular and plural thoughts.

Thus far, 'singular' has been used as opposed to 'general', when a specific thing rather than a type of thing is being talked about. Here, 'singular' is used as opposed to 'plural' to indicate talking about one rather than many things. Context should sufficiently distinguish these uses.

Like the singular/general distinction, this is a distinction between the form of the objects of thoughts: singular thoughts are about one thing, plural thoughts are about many[17]. If it is thoughts which are singular or general, then it is thoughts which are singular or plural too. Thus, an argument that there is no singular/plural distinction between thoughts will at least suggest that there is no singular/general distinction either, as it will cast doubt on distinguishing between thoughts by the form of their objects. For authors who include singularity (as opposed to plurality) in their definitions of singular thought, this will pose a more direct problem[18].

A similar unclear-subject problem arises for plural thought. A thought like *I know those men* appears singular, because there is only one of me (*I*), and plural because there are many of *those men.* But there is a more specific problem for the singular/plural distinction. A thought like *these people play against each other and those people play against each other* is surely plural, because these people and those people are pluralities. By the same token, *these people and those people play against each other* is also plural. But the sentence used to express the second thought is superplural — it is about a plurality of pluralities (of objects). There are superplural sentences in English, and superplural examples are often composed of plural items.[19] There are also apparently plural thoughts, expressed using plural terms[20]. Given that plural terms are reflected in plural thoughts, it is likely that superplural terms are reflected in superplural thoughts. A framework which distinguishes singular from plural thoughts should be able to distinguish ordinary plurals from superplurals.

But a simple distinction between singular and plural thoughts will be incapable of this, as the status of some

15. Crane 2013, 7.
16. Crane 2013, 141.
17. Crane 2013, 159; Azzouni.
18. Crane 2013.
19. Øystein Linnebo and David Nicolas, "Superplurals in English," *Analysis* 68, no. 3 (2008): 193.
20. Azzouni.

superplural terms will be unclear. *Those people and these people play against each other* appears superplural, because its expression includes the superplural term 'these people and those people'. But it also appears merely plural, because its expression includes the merely plural terms 'these people' and 'those people.' The apparent distinction seems to show that it is both superplural and merely plural, which is unacceptable. This problem has a similar form to the 'citizen of France' case: the status of one term in the expression of a thought is unclear. Distinguishing singularity from plurality at the level of whole thoughts thus fails to accurately distinguish plurals from superplurals.

This objection is against the singular/plural distinction, so it is not a knockdown argument against the singular/general distinction. However, as I wrote above, the two distinctions are very similar: they both distinguish between the form of the objects of thoughts ('if it is thoughts which are singular or general, then it is thoughts which are singular or plural too'). Showing that the standard singular/plural distinction fails thus casts doubt on the legitimacy of any such distinction, including that between singular and general terms. If thoughts are *not* singular or plural in the way previously supported, it is likely they are not singular or general as previously supported either.

3.3 Reference

The final problem is that singularity is generally accounted for in terms of reference, but thoughts are not the type of things which refer. Crane (2011), Azzouni (2011) and Sainsbury (2010) all invoke some notion of a thought referring in their accounts of singular thought. Crane notes that 'thought' can mean a mental episode of thinking, or the propositional or representational content *of* such an episode[21]. But neither of these things refers in the way a singular term like a name does.

It seems very unlikely that episodes of any kind refer. Reference is generally understood as a property or role of semantic items like names and descriptions, none of which are episodic.[22] It would seem very strange to ask what the referent was of a temporally extended episode. What, for example, could the event of my birth refer to? Extending this to mental episodes, it seems very unlikely that thought-episodes refer (except in a derivative sense discussed below).

It is more plausible that the content of a very simple thought — just holding an object 'in mind' — could refer like a name does. But examples like 1 which Crane and others give are not like this. Saying 'that man stole my wallet' is not like saying 'Matthew' or 'France.' Again, what would 'that man stole my wallet' (as a whole) refer to? It is much less plausible that the complex content of thoughts like these refers (as a whole) to anything.

4 Concepts

These problems suggest a natural solution. A defender of the idea that thoughts (purport to) refer might respond that there is a sense in which some episodic events *do* refer, and that thought episodes refer in the same way. A speech — as given at a conference or meeting — is a good example. Speeches are episodic, but there is still a sense in which a speech can be said to refer. For example, if someone gave a talk about American philosophers mentioning Jody Azzouni, one might say correctly that the speech referred to Azzouni. But the speech itself does not have a referent like a name does. It cannot be said to refer in the same sense that the name 'Jody Azzouni' refers to Jody Azzouni. Instead, the speech must include a term which *does* refer to Azzouni in the standard way — such as his name, or a description like 'the author of *Ontology without Borders*'.

So there is a sense in which the speech refers to Azzouni, but this is derivative of the fact that it contains a part which refers to him in the standard sense. A similar response could be made regarding the reference of thoughts: as a whole, a thought's content itself does not refer, but can be said to refer in virtue of containing some part (or composite of parts, like a description) which do refer. A thought episode then refers just if its content refers.

21. Crane 2011, 22.
22. Gareth Evans, *The Varieties of Reference*, ed. John McDowell (Oxford: Oxford University Press, 1982), 1.

But reference in this sense is grounded in reference in the standard, semantic sense. For the speech to refer to the issue, the speech must contain as a part a term (a name or description) which refers to the issue in the same way a name does. It is a necessary condition of the speech referring that there is such a part which refers. It is these parts which either do or do not refer in the standard sense.

Given that singularity is so often cashed out in terms of reference, *these referential parts* should be distinguished as singular or general instead of whole thoughts, since the parts are doing all the theoretical work. A speaker only refers to a thing in so far as she uses terms which refer to it. Similarly, a thought only purports to refer to a thing in so far as it includes parts which purport to refer. This solves the third problem.

Recognising this, the problem of the unclear subject now dissolves. There is no question of whether 1 is singular or general (or singular or plural). Whole thoughts are not the sort of thing which are singular or general, their parts are.

Accordingly, singularity and generality can still be accounted for in Examples 1 and 2. *That man* and *my wallet* are singular, *someone* is general.

The same approach — decomposing, and judging the parts as singular or general — can be applied to some of the parts themselves. *There is a citizen of France who is bald* would be expressed using the term 'a citizen of France.' This is general, and citizen of France could make it true. But it also includes *France* as a part, which is singular. So some singular or general parts of thoughts are composed of parts wich are themselves singular or general.

Similarly, some plural-like terms are composed of parts which are themselves plural. *Those people* is plural, and *these people and those people* is superplural (though the constituent terms *these people* and *those people* are merely plural). This does not solve the issue of identifying superplural thoughts. A solution would be a way for the distinction between singular and plural thoughts to account for superplurality. Rather, this removes the problem entirely: thoughts are not singular or plural, concepts are, and their singular/plural/superplural status is always clear.

I think an appropriate term for these composable parts of thoughts is *concepts.* There are singular and general concepts, and singular and plural ones. In this I partly follow Sainsbury (2010), as the next section explains.

5 Prior Support

There is support for a view like this in the literature. Crane suggests that a distinction between singular and general thoughts follows Quine's distinction between singular and general terms[23]. Singular terms like names "purport to refer to just one object".[24] Crane then writes "It is widely accepted that just as there are general and singular terms, there are general and singular thoughts."[25]

But this analogy is misplaced. The semantic analogue of a thought is not a term, but a sentence. Sentences, not terms, are used to express thoughts. One does not express a thought with 'Sam' or 'ten', but with 'Sam has ten fingers.' Even if this is not true of all thoughts, apparently clear examples of singular thoughts (Examples 1 and 2) *are* expressed with sentences, so it presumably true of at least these. Crane's move from terms to thoughts is an unjustified slide.

Sentences are the analogues of thoughts and it is a certain class of sentences' parts (terms) which are singular or general. Sentences themselves are not singular or general. Carrying the singular/general distinction from language to thought, thoughts — like sentences — are not singular or general, a certain class of their parts there.

This is borne out by the ways in which Crane and Sainsbury actually account for singular thought. Crane begins by discussing what makes psychological episodes singular. Assuming that these episodes are representational, he briefly discusses the reference of names, then introduces 'mental files' as having similar referential properties. Crane only gives an

23. Crane 2011, 21-22.
24. Willard Van Orman Quine, *Word and Object* (Cambridge, MA, USA: MIT Press, 1960), 96.
25. Crane 2011, 22.

account of the nature of singularity with respect to these files, not thought episodes.[26] Each mental file is a representation of a thing, paradigmatically an existing object. Singular files are those which only make sense as representative of one object. General files make sense as representative of more than one. He seems to take this as sufficient for an account of the singularity of representational mental episodes.[27]

Crane is not explicit about the relationship between these files and thoughts (singular or otherwise). However, it is implausible that he means these files to *be* representational thoughts. If a file represents an object, what file would be associated with the thought *that man stole my wallet*? Crane clearly intends for a file on *that man* to be associated with it — this is why he thinks the thought is singular. A file on *my wallet* should be associated with it too, and possibly one of the event of the stealing. But these are not themselves the thought, they are parts of it, or are involved with how the thought comes about (this must be true as Crane thinks thoughts are episodic[28], but files are persistent[29]). So the details of Crane's account entail a distinction not between thoughts, but between files — the representational *parts* of thoughts. It is merely a terminological difference that Crane calls these files, and I call them concepts (as does Sainsbury).

Sainsbury has a similar account. His definition of singularity is in terms of concepts: *individual concepts* are those 'fit for using to think about individual things.' These concepts are then 'used in' thoughts[30]. Sainsbury does not say what this use consists in, or how individual concepts relate to singular thought, but the 'use' of these concepts is clearly a necessary condition — for him only thoughts which use individual concepts are singular. Again, this locates the reality of the singular/general distinction not in the thoughts themselves, but in the concepts used in them. What matters is whether the concepts used by a thought are individual (again, Sainsbury's individual concepts and my singular concepts are only terminologically distinct).

6 Reconstructing Singular Thought

It might be objected that an account of singular thoughts can be reconstructed from my account of singularity. Such a view might look like this: some concepts are general and some singular. The contents of thoughts are (perhaps structured or compositional) composites of concepts, much like sentences are structured composites of terms (and other words). Those thoughts with content including at least one singular concept are singular thoughts, and all the others are general.

But this does not reveal anything important about the nature of singular thought — it is merely an arbitrary labelling of a certain group of thoughts, which have something in common. One could just as well advance a theory on which singular thoughts are those which *only* contain singular concepts. This will deliver different results from the first theory, but there is no methodological reason to choose between them, and neither label seems philosophically more useful than simply describing 'thoughts using at least one singular concept' or 'thoughts which use only singular concepts.' By definition, both identify singular thoughts, but in both the distinction which is doing all the theoretical work is still between singular and general concepts. There is nothing contradictory or problematic about such theories, but I do not think there is anything particularly useful either. One might just as well introduce a label for the dust jackets on books of philosophy (but not other books), then investigate the relationship between such dust jackets and philosophy students' essays. Really one would be investigating the relationship between the philosophy books and the essays, but this would be masked by the label for the dust jackets. Similarly, one could introduce a label 'singular thought' for all thoughts which use singular concepts (or which only use singular concepts), and investigate how these relate to objects in the world. But using such a label would only obscure the fact that any results of this investigation really concern a relationship between certain *concepts* and objects, not the thoughts which use the concepts. The label would be redundant at best and confusing at worst.

A slightly different approach to reconstruction might claim that a sentence like Example 2 is general with respect to *someone* and singular with respect to *my wallet*. But this is just a different way of phrasing my point: singularity

26. Azzouni footnote 5 makes a similar point.
27. Crane 2011, 36-37.
28. Crane 2011, 22.
29. Crane 2011, 38.
30. Sainsbury 301-302.

is not a matter of thoughts, it is a matter of concepts: for every concept (*someone*, *my wallet*, etc.) there will be a separate fact about singularity or generality (and about singularity or plurality). Obscuring this distinction with less clear, indirect phrasing does not seem useful.

Thus the questions which the singular/general distinction was originally meant to help answer (what is it for a thought to be about a thing? How do thoughts relate to what their objects are?) can be reframed in terms of concepts. Much of the general discussion surrounding singular and general thoughts will apply *mutatis mutandis* to this distinction. This alone will not answer the questions, of course, but it will tell us something about the nature of thought, and how thoughts relate to the things they are about.

7 Conclusion

I have shown that the standard distinction between singular and general (and singular and plural) fails for many thoughts, including those often advanced as clear examples of singularity. The way it fails, and the approaches of some writers to singularity in thought and semantics suggest a better account: singularity, generality and plurality of concepts which are used in thoughts. Though this distinction can be used to reconstruct an account of singularity and generality for thoughts (indeed, several accounts), no such account is philosophically useful.

References

Azzouni, Jody. "Singular Thoughts (Objects-Directed Thoughts)." *Aristotelian Society Supplementary Volume* 85, no. 1 (2011): 45–61.

Bach, Kent. "Getting a Thing Into a Thought." In *New Essays on Singular Thought*, edited by Robin Jeshion, 39. Oxford University Press, 2010.

Crane, Tim. *The Objects of Thought*. Oxford: Oxford University Press, 2013.

Crane, Tim, and Jody Azzouni. "Singular Thought." *Aristotelian Society Supplementary Volume* 85, no. 1 (2011): 21–43. https://doi.org/10.1111/j.1467-8349.2011.00194.x.

Evans, Gareth. *The Varieties of Reference*. Edited by John McDowell. Oxford: Oxford University Press, 1982.

Jeshion, Robin. "Introduction to New Essays on Singular Thought." In *New Essays on Singular Thought*, edited by Robin Jeshion, 1–35. Oxford University Press, 2010.

Linnebo, Øystein, and David Nicolas. "Superplurals in English." *Analysis* 68, no. 3 (2008): 186–197.

McDowell, John. "Truth-Value Gaps." In *Logic, Methodology and Philosophy of Science VI: Proceedings of the Sixth International Congress of Logic, Methodology, and Philosophy of Science*, edited by L. J. Cohen. North Holland Publishing Co, 1982.

Quine, Willard Van Orman. *Word and Object*. Cambridge, MA, USA: MIT Press, 1960.

Sainsbury, R. M. "Intentionality Without Exotica." In *New Essays on Singular Thought*, edited by n Robin Jeshio. 2010.

Taylor, Kenneth. "On Singularity." In *New Essays on Singular Thought*, edited by Robin Jeshion. Oxford University Press, 2010.

An Unfortunate Outcome of Banning Statistical Support for Belief

James Shearer*

University of St Andrews

The concept of Banning Statistical Support for Belief (BSSB) is often used to respond to the lottery paradox. This paper will claim that th the rational status is response is inadequate due to BSSB having strong consequences on everyday beliefs. The paper will first make a case for avoiding BSSB due to its impact on everyday beliefs. Second, the paper will discuss motivations for adopting BSSB. I will then discuss distinctions between credence and beliefs as attitudes, aiming to prove that since many of our beliefs are statistically based, BSSB is highly impactful. I will finish by addressing objections and clarifying why it is important to care about the consequences BSSB poses.

1 Introduction

Banning Statistical Support for Belief (BSSB) is a popular way of responding to the lottery paradox[1]; however, this paper will argue that it has significant consequences for the rational status of everyday beliefs that have not been adequately appreciated. The argument is as follows: a vast number of our beliefs have a statistical basing. Thus, BSSB will entail that many of our beliefs are not rational; instead, we should hold high credences attitudes. By making light of the wide-ranging impact of endorsing BSSB as traditionally stated, I aim to give a reason to motivate avoiding BSSB as a strategy if possible. In §2, I will establish the motivations for adopting BSSB. In §3, I will draw a distinction between credence and belief as attitudes. §4 will seek to prove that many of our beliefs are statistically based, and so BSSB has wide-ranging consequences. Finally, §5 will consider objections to my position while §6 will suggest some reasons as to why we might care about this consequence.

2 Why Ban Statistical Support for Belief

To assert BSSB is to argue that believing P based on the purely statistical evidence ("P will occur 95% of the time" for instance) is irrational.[2] This move was initially motivated by Nelkin as a response to the lottery paradox. There are two popular variants of the lottery paradox in the literature; we will be concerned with the rationality version. We can lay out the paradox formally as follows:

1. It is rational for the agent to believe that their ticket (t1) will lose.

*I am a fourth year student at the University of St Andrews. My areas of study are metaethics, reasoning, and imagination. My summers growing up were spent lobster fishing; they've given me a reprieve while I try out this philosophy thing.

1. Laid out formally in §2
2. Dana K. Nelkin, "The Lottery Paradox, Knowledge, and Rationality," *Philosophical Review* 109, no. 3 (2000): 373–409.

2. If it is rational for the agent to believe (t1) will lose, then it is rational for them to believe that (t2) will lose and so on for every ticket in the lottery.
3. It is rational for the agent to believe (t1) will lose, and (t2) will lose and so on for each ticket in the lottery (from 1&2).
4. Given that the agent does not think the lottery is rigged it is rational for them to believe that either (t1) will win or (t2) will win and so on for each ticket in the lottery.
5. It is inconsistent for the agent to believe the conjunction in (3) and that one of the tickets will win.
6. This inconsistency is apparent to the agent.
7. It is rational for the agent to believe things that they are aware are inconsistent (from 3,4,5,6).
8. It is not rational to believe things which are inconsistent and be aware of the inconsistency.
9. **Conclusion:** 1,2,4,5,6 or 8 are false (by reductio)

Of the potential candidates for rejection (1) seems the most plausible. (2) holds because the agent would need some particular reason to doubt their ticket over others, and that is lacking. (4) seems reasonable given how the case is set up as doubting that the agent can believe that one ticket will win, despite that being the point of the lottery, would presumably entail a broad and undesirable scepticism. (5) is an uncontroversial logical truth and (6) is reasonable if we grant that any rational lottery ticket holder who believes they will lose would be able to reason themselves into recognising this inconsistency by reflection. Denying (8) would mean denying consistency requirements for rational beliefs that are clearly inconsistent, which is not intuitively appealing, giving us a reason to seek an alternative.

Given that denying (1) seems the most desirable, we need an explanation of why the agent is not rational in believing that their ticket will lose. Nelkin notes that the statistical evidence used to justify (1) is a peculiar kind of justification for a belief because it does not entail a causal connection between the belief and the facts that make the belief true. For a belief such as "the furniture is still in the room I just left" the evidence one has ("there have been no peculiar sounds") causally connects the belief to the truth. If the furniture had somehow been moved, one would not expect to have the evidence (you would have heard peculiar sounds) and so would not have that belief. However, in the lottery case, the agent's evidence (losing is statistically likely) is not connected to the truth of the matter in the sense that whether the ticket is a winner has no bearing on whether losing is statistically likely.

Another way of understanding the dissimilarity between statistical and non-statistical evidence is to note how we act differently in light of being wrong depending on the type of evidence. You might think rationality aims to guide one to the truth; when one fails to believe true propositions, they either failed in being rational or were missing some evidence. In the furniture example, had it turned out that the furniture had been moved, then you would seek an explanation for why you lacked evidence to indicate that this was the case. But this does not happen in the statistical case; when you believe P due to statistical evidence and then learn not-P, there is no obvious candidate for evidence that you should look for to explain your false belief.

Nelkin takes this odd nature of statistical evidence to explain why it cannot make the agent's belief that they will lose rational. However, the question then emerges as to what the agent is rational in believing and to this Nelkin suggests "My ticket will probably lose"[3]. I take it to be more accurate to say that the agent is not rational in forming any belief, they are instead rational in forming a high credence that their ticket will lose. In the next section, we will consider how belief and credence are distinct attitudes.

3 The Credence Belief Distinction

E.G. Jackson, amongst other scholars, has made a distinction between two types of propositional attitudes: beliefs and credences. Beliefs are coarse-grained, you can believe P, disbelieve P, or remain undecided between P and not-P. Credences

3. Nelkin, 400.

are fine-grained, usually stated as a number 0-1 where having 1-credence(P) represents being maximally confident in P and 0-credence(P) being maximally confident in not-P.[4] This shows why BSSB entails that the agent in the lottery case is rational in holding a credence but not a belief; credence attitudes can capture the degreed aspect of the attitude that Nelkin ascribes to them. How exactly these two attitudes relate is an open question, a complete account of which is beyond the scope of this paper. This section will be concerned with the more modest goal of demonstrating that there are good reasons to think that credences and beliefs are distinct attitudes. To do this, I will cast doubt on both the belief-first and credence-first views.

3.1 Belief-First Views

Belief-first views argue that credences can be reduced to beliefs; that is, having a credence is just a matter of having a particular belief. Typically, they take a form similar to the following:

Belief-First: For S to have credence of n in p just is for S to believe (Mp), where M in an epistemic modal and M and n correspond to each other.

Epistemic modals are terms that describe the likelihood of an outcome, and they can be precise (the chance of p is 50/50) or imprecise (p will probably happen). An important thing to note about belief-first in its conventional form is that it is content-enhancing, as it relates some credence attitude with content P to a belief with the more complicated content of Mp.

Jackson argues against the belief-first view.[5] Her argument relies on two notions, grasping, and edge cases, which have been defined below:

Grasping: One grasps proposition P when one understands P such that they can form a propositional attitude with P as its contents.

Edge Cases: Edge cases occur when there is a proposition P such that one can grasp P but could not grasp a proposition more complex than P.

If we accept one can form an attitude with a proposition as its contents only if they can grasp P, then the following problem emerges:

1. There, may exist some proposition P such that S grasps P but would not grasp a more complicated proposition than P.
2. Because S can grasp P, they can form a credence with P as its content.
3. Mp is more complicated than P.
4. If S cannot grasp Mp, then they cannot believe Mp.
5. S cannot believe Mp (1,3,4) yet can have a credence in P (2).

Belief-first views entail that one cannot have a credence without having some related belief. Edge cases would show that it is possible to have some credence P and yet not be able to form the related belief entailed by belief-first because it would be too complex to grasp.

4. Elizabeth Jackson, "The Relationship Between Belief and Credence," *Philosophy Compass* 15 (6 2020): 1–3.
5. Elizabeth Jackson, "Why Credences Are Not Beliefs," *Australasian Journal of Philosophy*, 2021, 1–8.

It might be natural at this point to question whether or not edge cases exist. Why accept that there is some maximal point of complexity beyond which we could not understand a given proposition? Let me give a couple quick points in their defence here.

Firstly, complexity is a clear barrier for belief. By this I mean that we can make a proposition harder to believe by making it more complex and otherwise keeping its content the same. Examples of this are rife in philosophy; too often have I been confounded by a passage of Kant only to find myself amenable to the idea once it was properly explained by someone more learned. If we do things properly then the difference between the propositions expressed by Kant and my learned friend is not content but form. Once the form is simpler the content becomes graspable.

Secondly, how we learn indicates that the complexity of a proposition can be a barrier to our learning it. Grasping the Incompleteness Theorems, for instance, is not something that is immediately possible for most of us. Rather we start by learning simpler propositions that can form the basis for more complex propositions, from which, an understanding of the Incompleteness Theorems can emerge. I take this progression from simple to complex propositions to be a sort of pushing of the edges where the process of learning just is an expansion of the complexity of the propositions that we can grasp. That we must learn the simple to grasp the complex indicates an edge case that can be pushed.

All this to say, if we have reason to believe that there are some types of edge cases, then Jackson has posed a compelling problem for the common belief-first view. This gives us some basis for thinking that credences are not merely beliefs and so I will move on to motivating the second half of this section's argument: that beliefs are not merely credences.

3.2 Credence-First Views

Counter to belief-first views, credence-first takes it to be the case that having a belief is in some way reducible to having a particular credence. A conventional example of a credence-first view is the Lockean thesis which posits a normative connection between having a high credence and forming a belief[6].

Lockean Thesis: S ought to believe P iff S has a rational high credence in P.

One immediate reaction that one might have to this view is the question of how one construes "high credence" in a way that is not ad hoc. The level freest from this concern would be credence 1, but this comes with the worry of entailing a widespread scepticism, there are presumably few things in which we can be maximally confident.

Beyond ad hoc concern, there are reasons to be sceptical that beliefs are purely a matter of sufficient credence. Consider cases of naked statistical evidence (as suggested by L. Buchak[7]) where you have two potential culprits, Jake and Barbara, for a crime, and you know that Jake belongs to a demographic that is 10 times more likely to have committed the crime than Barbara. Given this demographic information, you ought to be able to form the very high credence that Jake committed the crime, but it seems no matter how likely Jake's demographic is to have committed the crime, only evidence that directly connected Jake to the crime could justify believing Jake had done it. What this suggests is that there may be concerns that bear on the rationality of belief (moral concerns for instance), but do not bear on credences such that one cannot simply reduce beliefs to having a specific credence.

4 How BSSB Entails That Many Beliefs are Irrational

Having established the motivations behind BSSB and that it entails denying rational belief to the agent and replacing it with rational credence (a distinct attitude), I now turn to the primary concern of this paper, establishing the wide-ranging

6. Jackson 2020a, 6.7
7. Lara Buchak, "Belief, Credence, and Norms," *Philosophical Studies* 169 (2014): 285–311.

consequences of BSSB. This objection builds on a suggestion made by D. Christensen, where it seems several of our beliefs have the following form:[8]

P only if not-Q.

Where our evidence for disbelieving Q is purely statistical. To see how this works, consider the following case. Mary tells you that she is driving to New York for the weekend, and that following Saturday, when someone asks you where Mary is, you tell them that she is in New York. However, because you are aware that sometimes things go wrong, you know that there are factors that would have prevented her from getting there; perhaps she was hit by a truck (Q). These factors not occurring form a necessary condition for Mary being in New York. Note that to rationally believe that Mary is in New York, you must disbelieve Q, it would not be enough to withhold belief. If the person had asked you if Mary had been hit by a truck on the way to New York, it would not be right for you to assert that you are undecided on the matter but believe that she made it. However, our grounds for believing that Mary was not hit by a truck are purely statistical; we have no evidence other than the fact that being hit by a truck while driving is rare.

And this shows how BSSB has unintuitive consequences. Suppose our belief that Mary was not hit by a truck cannot be made rational by inferring from the statistical evidence. In that case, we cannot rationally believe a necessary condition for Mary being in New York, and if we cannot do that, then we cannot rationally believe that Mary is in New York. Reflection will show that many epistemic beliefs have this structure, from Biden being president depending on him not having died, to your fridge still working because it has not suffered a random power surge. BSSB has the consequence of entailing that these beliefs are not rational. As with the lottery paradox, it seems that what we are rational in holding instead is a high credence in these matters.

5 Objections and Rebuttal

This section will address the concern that our belief that Mary is in New York is not based on disbelieving these Q outcomes (those being outcomes where if Q is true P cannot be the case). You might think that, when forming our belief, it never occurs to us that a truck hitting Mary was possible and so we never form a belief about that outcome. Instead, our belief is based on the far simpler inference of "Mary said she would be in New York; therefore, she is in New York" where Mary's testimony represents a form of non-statistical evidence which is therefore free of BSSB concerns. Under this model, statistical evidence need never enter the picture. In response, I note that my point regarding Q outcomes being a necessary condition stands, so if questioned about these possibilities, one does have to commit themselves to disbelieving them to rationally maintain the belief that Mary is in New York. In these cases, your basis for disbelieving Q is presumably still statistical. The alternative would be inferring that not-Q from your belief that P. But given that P is itself based on testimony, this has the counter-intuitive result of suggesting that Mary's testimony can justify your disbelief in Q; despite it seemingly being the case that there is no explanatory connection between the testimony and the possibility of Q. When asked why you do not believe that Mary was hit by a truck, it would not be satisfactory to respond, "because she told me she would be in New York and therefore I believe she is in New York and was not hit by a truck".

One can press this concern about whether statistical evidence ever enters the picture absent of individuals questioning you directly about Q outcomes. In her 2020 paper, Jackson argues for a saliency-based distinction between two types of evidence.[9]

B-Evidence: Evidence for P that does not make salient not-P.

C-Evidence: Evidence for P that does make salient not-P.

8. David Christensen, "Putting Logic in its Place: Formal Constraints on Rational Belief," chap. Deductive Constraints: Problem Cases, Possible Solutions (Oxford University Press, 2004), 33–68.

9. Elizabeth Jackson, "Belief, Credence, and Evidence," *Synthese* 197, no. 11 (2020): 5073–5092, https://doi.org/10.1007/s11229-018-01965-1.

Where salience is simply a matter of whether a possibility is being considered, P is salient when one considers P as a potential outcome.

Given this, I suggest the following objection.[10] C-evidence cannot make belief rational for similar reasons as we had in §2. Namely, when you base a belief on C-evidence and then find out that you were wrong, there is no obvious candidate for fault in the reasoning process. In contrast, if you base your belief on B-evidence, you will seek an explanation as to why your reasoning led you to a faulty belief. In the lottery case, we clearly have C-evidence, but this is not the case for typical epistemic beliefs; a belief like "Mary is in New York" is an example of B-Evidence. Because our evidence in typical epistemic cases never make salient not-P (by making us consider Q possibilities), we have a way of distinguishing said cases from the lottery cases. Additionally, because Q is never made salient in the first place, we have a plausible reason to think that not-Q does not base our epistemic beliefs.

In response, I would like to show why we might be sceptical of how appropriate this salience requirement is. First, note that under this argument if Q possibilities are made salient, then the belief will once again appear irrational under BSSB because of the difficulties of finding non-statistical evidence against Q as explained above. Jackson admits that whether a piece of evidence is type B/C can depend on factors other than the contents of the evidence; her example is how the evidence is presented.[11] A more pressing concern is that the different mental states of two agents can also affect whether evidence is B/C. Imagine another Mary case where Mary has told two different people, Vanya and Elliot, that she is going to drive to New York. The previous night Vanya had watched a documentary on road safety, and when they receive the testimony evidence, their mind is drawn to the possibility that Mary might crash. Because of this, a Q possibility is made salient, and therefore Vanya cannot rationally believe that Mary will be in New York. Elliot, on the other hand, had seen that same documentary five years prior, but because he has not thought about it for some time, his mind is not drawn to the possibility of a crash when he gets Mary's testimony. So for him, the testimony remains B-Evidence, and he can rationally believe that Mary will be in New York. The difference between Vanya and Elliot is not their body of evidence. The difference is simply in how recently they have considered the relevant evidence, and yet this difference is a deciding factor between which of them can rationally believe P under the saliency model. That a factor such as this could be the difference-maker in whether a person is rational is presumably not a desirable consequence if we agree that such factors have little to do with rationality and therefore gives us reason to doubt the saliency view.

As a final remark in defence of my view I would like to point out that your belief in P is seemingly affected by changes in your attitude towards Q. Consider another Mary case where car accidents are relatively common (they occur 50% of the time), in this world; you could not form a belief regarding Q even if BSSB were not a concern as you have no evidence to form a belief either way. Even absent any questions about car accidents or the possibility being made otherwise salient it is not clear that one could rationally believe that Mary had made it New York without being able to rationally disbelieve Q. I take this observation to suggest that your belief P being rational is sensitive to your disbelief in Q even when you form no conscious attitude towards Q. If this is the case, then the BSSB concerns I have been raising may be present regardless of what one has to say about salience.

6 Why Care?

Even if you accept the assertion that BSSB has the wide-ranging consequences that I have described, you may still doubt that it matters; there is a question of why exactly we should care if many of our beliefs are not rational, but a high credence attitude is. §3 indicates why we should think the attitudes are distinct, but the question of why we should seek to have one over the other is a separate issue. As with §3, a full account of this question is beyond the scope of this essay, though here I will briefly point towards two reasons that indicate that the difference between belief and high credence is one that matters.

First is an appeal to our intuitions about what beliefs we take ourselves to have. You might think that it is intuitively wrong for an argument to have the conclusion that we do not rationally believe a great many everyday things,

10. Jackson's paper does not directly make this point; it is concerned with improving on BSSB. However, I take her view to apply to our current context.

11. Jackson 2020b, 5088.

and this is one of the reasons that we are motivated to find ways to reject scepticism. The results of BSSB that I've been arguing for are not nearly the level of widespread scepticism, but the supposed high credence attitudes are not the common sense understanding of the sort of attitude that we usually take ourselves to have to these epistemic propositions.

Second is an appeal to the idea that beliefs simplify reasoning.[12] Suppose one has a vast number of credences that relate to a specific subject. In that case, the calculations that one must do to reason on that subject while maintaining rationality can become very complicated. For instance, if certain credences are high, then the reasoning process can be simplified by appealing to a belief with the same content. This gives us a reason to maintain the commonsense view that we have many beliefs with which to reason.

Before concluding, I would like to acknowledge that if one is an eliminativist about belief (that being the position that belief is not an attitude which we need to include in our ontology), then you are unlikely to care if BSSB has the consequences I suggest. However, eliminativists still have reason to consider the work done in this essay valuable because it shows how BSSB can support eliminativism by revealing that many everyday beliefs are irrational.

7 Conclusion

This essay has argued that BSSB has the consequence of entailing that many of our epistemic beliefs are not rationally held and that instead they should be replaced with a high credence attitude. This conclusion has been reached by suggesting that many beliefs can only be rationally held if one believes that the required necessary conditions are met. These beliefs in the necessary conditions are usually based purely on statistical inference which cannot make beliefs rational according to BSSB.

I have furthered my view as follows. By arguing that many of our everyday beliefs do have this statistical basing by showing how one is rationally compelled to appeal to these statistical facts when questioned about Q possibilities. I then demonstrated that an appeal to saliency introduces problematic features as difference makers, when questioning whether one is rational. And finally, I suggested that believing P is sensitive to your attitude towards Q even when these attitudes are not conscious. While this paper does not directly doubt BSSB, it is expected that the wide-ranging consequences of the move give us a reason to seek a different solution. I have justified this expectation first by showing that credences and beliefs are distinct attitudes and then suggesting that belief plays a pivotal role in simplifying reasoning that credences cannot.

References

Buchak, Lara. "Belief, Credence, and Norms." *Philosophical Studies* 169 (2014): 285–311.

Christensen, David. "Putting Logic in its Place: Formal Constraints on Rational Belief." Chap. Deductive Constraints: Problem Cases, Possible Solutions, 33–68. Oxford University Press, 2004.

Jackson, Elizabeth. "Belief, Credence, and Evidence." *Synthese* 197, no. 11 (2020): 5073–5092. https://doi.org/10.1007/s11229-018-01965-1.

———. "The Relationship Between Belief and Credence." *Philosophy Compass* 15 (6 2020): 1–13.

———. "Why Credences Are Not Beliefs." *Australasian Journal of Philosophy*, 2021.

Nelkin, Dana K. "The Lottery Paradox, Knowledge, and Rationality." *Philosophical Review* 109, no. 3 (2000): 373–409.

Staffel, Julia. "How Do Beliefs Simplify Reasoning?" *Noûs* 53, no. 4 (2019): 937–962.

12. Julia Staffel, "How Do Beliefs Simplify Reasoning?," *Noûs* 53, no. 4 (2019): 937–962.

Aporia

Undergraduate Journal of the St Andrews Philosophy Society

VOLUME XXII:

FEMINIST APPENDIX

The Unhappy Marriage of Feminism and Veganism

Luke Ryan*

University of St Andrews

This paper will argue that, in a manner parallel to Hartmann's description of the relationship between Marxism and Feminism, the marriage between feminism and veganism has also been unhappy. This paper will argue that this is due to feminism dominating over veganism. I will begin by discussing historical criticisms of unions between feminism and Marxism. From here, Adams unification of veganism and feminism will be introduced. After this, it will be argued that attempts by Adams to marry veganism and feminism have led to a diminished and centralised view of animals as victims. Furthermore, I will explain why it is that this type of 'marriage' was bound to fail from the beginning.

1 Introduction

The women's rights movement and animal rights movements have a long history of overlap, as identified by Carol J Adams in *The Sexual Politics of Meat: A Feminist-Vegetarian Critical Theory*.[1] In her book, Adams documents the history of vegetarian feminism from the 19th century to today. Adams also posits the argument that the dominant ideological frameworks which normalise the exploitation of women and non-human animals[2] are one and the same.

Taking inspiration from Hartmann's *The Unhappy Marriage of Marxism and Feminism*,[3] which argues that approaches synthesising Marxism and Feminism have been unsuccessful because the Marxist component of the analysis has predominated over the feminist component, I will argue that the marriage of feminism and veganism[4] has also been unhappy due to feminism dominating over veganism. This paper will first briefly summarize the section of Hartmann's analysis which is crucial to my argument. I will then criticize Adams' formulation of feminist veganism for de-centering animals. Finally, I will attempt to explain why the vegan component has inevitably been subordinated to the feminist component due to the context in which Adams' work was written.

2 Hartmann's Criticism of Zaretsky

Eli Zaretsky[5] argues that when women do housework they are in fact working for the benefit of capital. They are performing the unpaid labour of social reproduction that is necessary for the maintenance of the workforce, and in doing

*I am a 4th year student studying MSci Computer Science at the University of St Andrews. I am more of an activist than a philosopher. Since turning vegan towards the end of 2020, my primary political focus has become animal rights.

1. Carol J. Adams, *The Sexual Politics of Meat: A Feminist-Vegetarian Critical Theory, Twenty-Fifth Anniversary Edition* (Bloomsbury, 2015).
2. Hereafter referred to as animals
3. Heidi I. Hartmann, "The Unhappy Marriage of Marxism and Feminism: Towards a more Progressive Union," *Capital and Class* 3, no. 2 (July 1979): 6–7.
4. Hereafter 'veganism' shall refer specifically to ethical veganism, rather than broader dietary veganism
5. Eli Zaretsky, *Capitalism, The Family, and Personal Life* (Pluto Press, 1974).

so are creating the conditions necessary for the extraction of surplus value from wage labourers (their husbands). For Zaretsky, the oppression of women consists of their exclusion from the class of wage-labourers and capital's exploitation of their reproductive labour.

Hartmann criticises Zaretsky for only considering the relationship between women and capital, and not addressing the relationship between men and women directly. Hartmann argues that this disregards the fact that women doing housework are working for their husbands, and so therefore their husbands are in a position of power over them. This is true even if the husbands are themselves oppressed as workers due to their relationship to capital.

Through his Marxist lens Zaretsky can see that women are oppressed by capital, but he cannot see that they are oppressed by their husbands. Therefore Zaretsky has tried and failed to explain away patriarchy through a purely class oriented economic analysis, and so has made feminism subordinate to Marxism in his analysis.

3 The Sexual Politics of Meat and the Absent Referent

In *The Sexual Politics of Meat*, Adams argues that there is a link between the oppression of women and the oppression of animals. A key theoretical device employed by Adams is the structure of the absent referent, which she borrows from semiotics. Adams explains the concept, as it relates to the consumption of animal bodies, thusly:

> Through butchering, animals become absent referents. Animals in name and body are made absent as animals for meat to exist. Animals' lives precede and enable the existence of meat. If animals are alive they cannot be meat. Thus a dead body replaces the live animal. Without animals there would be no meat eating, yet they are absent from the act of eating meat because they have been transformed into food.[6]

Adams believes that this linguistic structure is used to make the violence behind the eating of animals socially permissible.[7] Adams argues that not only farmed animals, but also women, are marginalised through the structure of the absent referent.[8] She claims that the way that female bodies are objectified and metaphorically deconstructed into constituent parts (e.g. breasts, legs) parallels the way that animals are killed and physically dismembered to become the uncountable substance known as "meat".

Adams considers pornography to be intrinsically violent and linked with rape culture, and makes a pun of the multiple meanings of the word "consume", drawing a connection between the consumption of pornography and the consumption of animals.[9]

Carrie Hamilton takes issue with Adams anti-sex-work, anti-pornography stance and contends that the comparison between the sexualisation of female bodies and the violent dismemberment of animals fails on two accounts.[10] Firstly, it erases the agency of sex workers. Adams' text equates sex work with sexual abuse. Moreover, she does not draw on the perspectives of sex workers within her work, despite its focus on sex work. Secondly, the comparison does not adequately account for the differences between violence against women and violence against animals. Adams trivialises how extreme violence against animals is by comparing it to the violence which is supposedly implicit in sexualised images. Hamilton says that the connection drawn by Adams is "theoretically weak and evidentially wanting".[11]

Within Adams' analysis feminism is completely dominant. She names the structure oppressing animals as patriarchy. She does not describe any ideological or socioeconomic structure specific to animal exploitation, such as speciesism or carnism. Therefore the book presents the oppression of animals as merely an extension of the oppression of women rather than an entity in its own right.

6. Adams, 66.
7. Adams, 69
8. Adams, 67-69
9. Adams, 88
10. Carrie Hamilton, "Sex, Work, Meat: The Feminist Politics of Veganism," *Feminist Review* 114, no. 1 (January 2016): 112–129.
11. Hamilton, 1.

Adams only analyses animals' relationship to patriarchy, without analysing their relationship to humans (including women). This is analogous to how Zaretsky only analysed the relationship of proletarian women to capital, without analysing their relationship to proletarian men. So just as Zaretsky did not consider the material interests that proletarian men have in upholding the system of patriarchy, Adams does not consider the material interests that women (as humans) have in upholding the system of animal exploitation.

Where Adams does analyse the relationship between women and animals she is discussing a particular minority: vegetarians. The book records the ideas of first wave feminists who adopted a sort of vegetarian cultural feminism. Those feminists believed that an non-violent stance toward both humans and animals is an essential quality of women. Correspondingly, they also believed that men have an inherent disposition to violence.[12]

Unfortunately sections of the book document these arguments and beliefs without evaluating them. The claims about masculine violence and feminine pacifism were made by western feminists and vegetarians with a limited understanding of other cultures, and an outmoded understanding of sex and gender. Universalising , essentialist claims such as these are not compatible with a historical materialist understanding of patriarchy.

The lack of criticism of these viewpoints leads to it appearing that the book endorses the idea of a natural kinship between women and animals. This runs counter to the obvious fact that the vast majority of women (as with all humans) are invested in animal exploitation both ideologically and economically. Women; as members of animal farming communities; as consumers of meat, eggs and dairy; and as wearers of leather, wool and fur; are the direct benefactors of animal exploitation.

It could be argued that women ultimately do not benefit from animal exploitation, because (in Adams view) the commodification of animals in turn undermines womens' rights and bodily autonomy. Nonetheless, it must be seen that non-vegan women directly benefit from animal exploitation. This exactly parallels Hartmann's point about proletarian men's power within patriarchal capitalism:

> In the long run this may be 'false consciousness', since the majority of men could benefit from the abolition of hierarchy within the patriarchy. But in the short run this amounts to control over other people's labor, control which men are unwilling to relinquish voluntarily.[13]

4 The Failure of the Arguments from Female Exploitation

We will now turn to the least academic and most commonly used family of arguments which connect veganism to feminism. These arguments will be most illustrative in understanding why the marriage of feminism and veganism has been unhappy. They are employed by mainstream animal advocacy organisations like PETA[14] and Viva[15] and an argument of this kind is also implicit in *The Sexual Politics of Meat*. The argument is as follows.

The Argument From Female Exploitation:

- **P1.** Feminists oppose females being oppressed on the basis of their sex.
- **P2.** Within animal agriculture, females are oppressed on the basis of their sex[16]

12. Adams, chapter 7.
13. Hartmann, 6-7
14. Tara DiMaio, "International Women's Day: Why You Can't Be a Cheese-Eating Feminist," 2021, accessed January 2, 2022, https://www.peta.org/blog/feminist-womens-day-go-vegan/.
15. Justine Butler, "Do Feminists Drink Dairy?," 2021, accessed January 2, 2022, https://viva.org.uk/health/health-articles/do-feminists-drink-dairy/.
16. An example: cows in the dairy industry and hens in the egg industry have their reproductive systems exploited. They are selectively bred to overproduce milk/eggs at the cost of their health. Furthermore, female animals in various forms of animal agriculture are sexually assaulted by humans to force them to become pregnant.

- **C1.** Feminists should oppose animal agriculture
- **C2.** Feminists should be vegan.

While this argument may or may not be effective in practice, it is theoretically lacking. The argument has a premise that feminists by definition opposes exploitation of females, or oppose systems where females are exploited more severely than males.

However, within a speciesist domain of discourse, when a non-vegan feminist says a statement like "all mothers deserve rights", they are making a statement only about human mothers. Therefore arguments from female exploitation often rely on an ambiguity of definition in words such as female, mother, oppression etc, where the speciesist is employing a narrower definition, and will not necessarily change their position merely because their words have been re-interpreted by an anti-speciesist.

If a feminist opposes all exploitation of all females, there doesn't seem to be any reason that they should not also oppose all exploitation of all males. In this case there is no reason for the argument to be particular to feminists or particular to female animals, and it seems to be fairly contentless. It boils down to "if exploitation of animals is wrong then exploitation of animals is wrong." The only content in the argument then is the description of certain industry practices. The non-vegan feminist can dismiss the argument by simply saying that they don't believe animals deserve rights, or any such equivalent argument.

If the argument is however saying that animal agriculture is bad because it treats females worse than males, then this seems to be a contingent empirical matter that opens up an irrelevant discussion. Assuming that this were true, would animal agriculture then become acceptable if the conditions of males were worsened to become equal with females? Or conversely if the treatment of females was "improved" to the level of the treatment of males? Male chicks are suffocated to death or ground up alive on their first day of life. Male dairy calves used for veal are kidnapped from their mothers, trapped in small crates and then killed at just a few months old. Males in various forms of animal agriculture are sexually manipulated by humans just as females are.

Perhaps the argument from female exploitation is supposed to be convincing because the supposed worse treatment of female animals relative to male animals then causes or reinforces to a worse treatment of human women relative to human men. This line of reasoning returns to anthropocentrism and throws away the moral consideration of animals as the premise for veganism. Therefore it cannot defend animals from sex-blind exploitation and killing.

The reason why humans treat male and female farmed animals differently is not because females are viewed as less morally important than males. Neither male nor female animals are treated with moral importance – they are treated as commodities. Measures which supposedly account for animal welfare are made as business decisions. They either directly lower the cost of production of animal products, or they increase profitability by increasing the marketability of goods through special labelling.[17] This causes a problem for those who are trying to argue that there is an ideological link between our treatment of female animals and our treatment of women.

So why do these activists and groups focus particularly on the sexual and reproductive aspects of animal exploitation? I believe that the real weight of this argument is simply that it brings up a lesser known but extremely egregious aspect animals exploitation. It seems inconsistent to hold that killing and imprisoning an animal is acceptable but sexually violating them is not. However, the discussion of sexual violation is a more visceral example because it is an aspect of animal exploitation that many people do not know about and most people are not desensitised to. We all known that meat is a body part of an animal who was killed.

Therefore my criticisms of the argument do not really undermine its point. The function of the argument is merely to point out, or to remind one of, the realities of animal agriculture — to dereference the absent referent. I hypothesise that many syntheses of veganism and feminism function in this manner.

17. Gary L. Francoine, *Rain Without Thunder: The Ideology of the Animal Rights Movement* (Temple University Press, 1996).

5 Why has the Marriage *Really* Failed?

I will now suspend charitability and attempt to provide a pragmatic explanation as to why works in the vein of *The Sexual Politics of Meat* are created, and why they decenter animals. I hypothesise that the motivation behind these syntheses is to attempt to tie animal rights to a larger movement in order to recruit more vegans. Animal rights activists may believe that those who already fight other forms of oppression are likely to be amenable to the cause of animal rights. Within this context the animal rights theorist must explain the relevance of animal oppression to the feminist theorist, rather than the other way round. This leads to an explanation of the oppression of animals oppression couched in terms of the oppression of women. Therefore it is inevitable that feminism ends up as the dominant frame within the synthesis.

The problem with this approach, from the animal rights perspective, is that non-vegan feminists can criticise the synthesis solely by disputing the strength of connection between the two forms of oppression. Therefore they can debunk a work of this kind without ever addressing the vegan's central question: whether the killing and enslavement of animals for food, clothing etc. is morally justified. The end result is a conversation which is around animals, rather than about animals. For the vegan, violation of animal rights would be wrong even if it did not affect marginalised humans whatsoever. Despite this, activists following Adams' line have tied their advocacy to the existence of a causal link between these different forms of oppression.

For the feminist activist, they have a lot to lose from endorsing veganism and animal rights. Firstly, to become vegan requires changing one's lifestyle. Adams recalls[18] how she heard that some feminists at the Modern Language Association refused to read *The Sexual Politics Of Meat* because they did not want to stop eating animal bodies. Becoming vegan means one must not just change what they eat and wear, but one must also admit that they have participated in an atrocity up until the point of change, making it a doubly bitter pill.

Like the theologian confronted with alien intelligence, or the physicist confronted with Boltzmann brains, the critical theorist confronted with animals as subjects must relitigate the base assumptions of their analysis. The sheer number of animals exploited by humans demands that we reorient our understanding of the "typical observer" of our society. Animals are exploited in such extreme ways, and largely for reasons as trivial as taste pleasure. The fact that even within justice movements this exploitation is largely ignored and/or justified casts non-vegan critical theory in a concerning new light.

6 Conclusion

In conclusion, I have shown that Adams' attempt to marry veganism and feminism has made the feminist lens of analysis dominant, leading to a poor explanation of the material and ideological systems of animal exploitation, and a decentring of animals as victims. I have also explained why I believe that this marriage was bound to be founded on an unequal footing. I follow Hartmann's lead in stressing that any future approach to synthesising the two movements should follow a materialist methodology. However it must avoid being reductive in the sense of explaining, or explaining away, one mode of oppression in terms of another. Generating such a synthesis will require ongoing work in the expanding field of animal studies.

References

Adams, Carol J. *The Sexual Politics of Meat: A Feminist-Vegetarian Critical Theory, Twenty-Fifth Anniversary Edition*. Bloomsbury, 2015.

18. New Books Network, "Carol J. Adams Interview with Mark Molloy," 2021, 54:40, accessed January 2, 2022, https://newbooksnetwork.com/the-sexual-politics-of-meat.

Butler, Justine. "Do Feminists Drink Dairy?," 2021. Accessed January 2, 2022. https://viva.org.uk/health/health-articles/do-feminists-drink-dairy/.

DiMaio, Tara. "International Women's Day: Why You Can't Be a Cheese-Eating Feminist," 2021. Accessed January 2, 2022. https://www.peta.org/blog/feminist-womens-day-go-vegan/.

Francoine, Gary L. *Rain Without Thunder: The Ideology of the Animal Rights Movement*. Temple University Press, 1996.

Hamilton, Carrie. "Sex, Work, Meat: The Feminist Politics of Veganism." *Feminist Review* 114, no. 1 (January 2016): 112–129.

Hartmann, Heidi I. "The Unhappy Marriage of Marxism and Feminism: Towards a more Progressive Union." *Capital and Class* 3, no. 2 (July 1979): 1–33.

Network, New Books. "Carol J. Adams Interview with Mark Molloy," 2021. Accessed January 2, 2022. https://newbooksnetwork.com/the-sexual-politics-of-meat.

Zaretsky, Eli. *Capitalism, The Family, and Personal Life*. Pluto Press, 1974.

Epistemic Injustice in the Age of AI

Martina Sardelli*

University of St Andrews

Artificial Intelligence (AI) is revolutionising our practices of distributing and producing knowledge. Though promising, these technologies also harbour the potential for corruption - a rising problem in this domain is that of injustice committed against women in the epistemic sphere. In our social framework, being regarded as a credible knower has become synonymous with the potential for self-actualisation: the realisation of one's potential. As such, the gender bias perpetrated by some AI systems is harming women in this domain. Additionally, biased software is barring them from accessing hermeneutical resources relevant to the understanding of their lived experience. Though still in its infancy, the problem should be urgently addressed by conceptualising ways in which a fairer AI could be engineered. Egalitarian ideas, specifically focused on equality of opportunity, seem to be promising avenues for future research and thought.

1 Introduction

Artificial Intelligence (AI) is the ability of computers and machines to perform tasks emulating those undertaken by the human mind, e.g. perception and decision- making, among others.[1] The development of these technologies in recent years has made their use intrinsic to the fabric of our daily lives, and AIs are now important components of our search engines, medical diagnostic tools and surveillance technologies.[2][3][4] Though AI software continues to define technological progress, the outputs produced by these systems can also perpetuate bias.[5] This, coupled with the power dynamics underlying gender discrimination, historically sustained by a social, political and economic infrastructure, are at the core of AI's intersection with epistemic injustice.[6] Epistemic injustice itself can be understood as a series of practices whereby knowers are wronged qua knowers, as well as practices which distort or impede the understanding of a knower, doing so from within a framework of established epistemic practices and institutions.[7][8]

In this essay, I endeavour to show how gender bias is perpetuated by AI through the lens of epistemic injustice. I claim that gender bias in AI is a problem which needs to be urgently addressed as it hinders women's capacity for self-determination and their ability to be perceived as credible conveyors and possessors of knowledge. I will do so first by

*My name is Martina, I come from Luxembourg and I am a final-year student at the University of St Andrews, studying Biology and Philosophy. Later in the year I am set to start a master's degree in social science and the internet. As such, I am particularly interested in the integration of AI into mainstream working practices and its repercussions! I also make the meanest focaccia.

1. IBM Cloud Learn Hub, "What is Artificial Intelligence," 2020, accessed April 20, 2021, https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

2. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018), accessed May 24, 2022.

3. Fei Jiang et al., "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology* 2, no. 4 (2017): 230–243.

4. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ed. Sorelle A. Friedler and Christo Wilson, vol. 81, Proceedings of Machine Learning Research (PMLR, 2018), 77–91.

5. Tolga Bolukbasi et al., "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in *Advances in Neural Information Processing Systems*, ed. D. Lee et al., vol. 29 (Curran Associates, Inc., 2016), 1.

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

7. Gaile Pohlhaus Jr., "The Routledge Handbook of Epistemic Injustice," chap. Varieties of Epistemic Injustice, ed. Ian James Kidd, Jose Medina, and Gaile Pohlhaus Jr. (Routledge), 13.

8. Fricker, 1.

presenting instances of gender bias in different AI systems and their detrimental effects on women's ability to transmit credible knowledge. Then, I will explore Fricker's account of epistemic injustice and how AI factors into the equation. A counterargument for AI's involvement in epistemic injustice focuses on AI's accountability in the moral arena, dodging claims of it being a cause of epistemic injustice. To this, I rebut that AI does not carry out epistemic injustice directly, but rather bolsters a social mind-set where the notion and practices of treating women as non-credible epistemic agents is buttressed. Lastly, I will address what a "fair" AI could look like using Rawls' theory of justice as fairness and Binns' notions of egalitarianism as applied to machine systems.

2 Laying the Groundwork: Instance of Gender Bias in AI

Instances of gender bias have been documented extensively in the AI literature from domains as disparate as precision medicine and civil surveillance, with concrete repercussions for women[9][10]. In many cases, gender and sex are viewed as confounding factors or unimportant, and often aren't factored into the training data which goes on to establish the architecture of an algorithm, *e.g.* for AIs doing precision medicine, "sex" is often omitted in training datasets[11].[12] Recently, important instances of sexist biases have been revealed through the inaccuracy of facial recognition software in identifying women, especially black women[13]. False positives generated by these technologies (*i.e.* cases in which a person is misidentified by the system) and their increasing deployment in the context of the criminal justice system threaten to seriously undermine the civil liberties of those affected in profoundly unfair ways[14]. In addition to gender, protected categories such as race, gender, ethnicity and religion are also at risk of being discriminated against. Though the programming of an unbiased AI necessitates factoring all these groups in, this essay will specifically focus on AI's *negative* bias against women. As such, any detailed analysis of racial, class or religious bias in conjunction with AI is beyond its scope.

Understanding the deleterious effects of biased AI in the medical and civil realm is a pressing issue, however, I hold that natural language processing (NLP), content moderation systems and automated résumé filters are particularly interesting to elucidate how artificially intelligent systems are tied to epistemic injustice in the purview of gender. Applications of NLP range broadly from recognition of speech to machine translation.[15] Most often, the algorithms for these systems have been shown to be insensitive to the different nuances of spoken and written language between genders, to the detriment of women. An oft- cited example is Bolukbasi *et al.*'s (2016) paper on biased word embedding NLPs. Word embedding is employed to capture semantic relationships between words, which are represented as vectors in a geometric space. Words whose semantic meanings are similar will have vectors located close to each other in this space[16]. These relationships can be shown using machine- completed analogy puzzles, e.g. "man is to king as woman is to *x*", where, in this instance, "*x*" equals "queen"[17]. Training of word2vec (an embedding technology) on a corpus of Google News (w2vNEWS) texts has shown clear perpetuation of harmful societal gender stereotypes in multiple instances. Bolukbasi *et al.*, claim that the sexist analogies (e.g. "she: diva= he: superstar"[18]) generated by the software trained on this data, seem not only to reflect but amplify pre-existing biases of the dataset on which the algorithm was trained[19]. Another incident where an AI, trained by engineers at MIT to label people and objects in images, went rogue, calling women derogatory terms in reference to sex workers, springs to mind[20].

As well as the perpetuation of harmful sexist and misogynistic stereotypes, the systematic censorship of women

9. Davide Cirillo et al., "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," *NPJ Digital Medicine* 3 (81 2020): 1–11.
10. Buolamwini and Gebru.
11. Cirillo
12. Katyanna Quach, "MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs," 2020, accessed May 6, 2021, https://www.theregister.com/2020/07/01/mit_dataset_removed/.
13. Buolamwini and Gebru
14. Buolamwini and Gebru
15. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson Education, 2021), 1439.
16. Bolukbasi, 1.
17. Bolukbasi, 1.
18. Bolukbasi, 2.
19. Bolukbasi, 2.
20. Quach

through content moderation software is cause for concern.[21] Binns *et al.* found content moderation systems to be less sensitive to female-labelled test data, generating more false positives: female-authored expressions which human moderators in the study did not deem offensive were more likely to be unreasonably classified as such and censored by the machine system[22]. A similar, but more covert, censorship is carried out by AIs which screen résumés: those featuring female job applicants from all-women colleges and including words such as "women's" were systematically targeted by Amazon's automated filtering system to be rejected.[23] Thus, the common, problematic denominator tying these systems emerges: they harm and handicap women epistemically, *i.e.* as credible knowers. In these cases, women are the subject of clichés which directly impeach their ability to communicate knowledge. More indirectly they play on sexist stereotypes, historically culpable for sustaining the perception of women as less rational and/or less capable of imparting reliable knowledge than their male counterparts, as well as excluding them from having the same opportunities on these very grounds. As I will explore in the next section, uncovering these biases is a key first step to comprehend how AI fits into the wider framework of epistemic injustice.

3 The "Stereotype - Prejudice - AI - Epistemic Injustice" Pipeline

Before delving deeper into the insidious ways AI can perpetuate epistemic injustice, it is important to illustrate the concept of injustice through an epistemic lens as well as the power dynamics at play when epistemic exchanges go awry. To lay the groundwork, Fricker defines social power as a form of power exerted structurally or by an agent whereby either has the capacity to affect others' actions in virtue of a specific social situation[24]. Though having a practical dimension, it also embodies what Fricker terms an aspect of "imaginative social co-ordination"[25], that is, shared beliefs about what it means to belong to a certain social identity, *e.g.* what it means to be a "woman". A corollary of social power, repurposed to focus on shared conceptions of social identity, is identity power, which is power exerted in relation to the conceptions of identity borne of our collective social imagination[26]. Epistemic judgements are fundamentally based on credibility judgements: the final aim of an epistemic exchange is to deem whether the knowledge conveyed is reliable or not (*i.e.* non- credible knowledge). Crucially, Fricker points out that heuristics are important to arrive at a final judgement and often involve the use of stereotypes[27]. Though she adopts the term neutrally[28], throughout this essay I will only use the term in its negative sense, just as I will only use "bias" in its negative connotation. As such, I will take stereotypes to indicate unreliable "widely held associations between a given social group and one or more attributes"[29]. Examples of sexist stereotypes include women's temperaments (*e.g.* "women are hysterical" etc.) but also extend to a more physical/practical dimension, *e.g.* brain make-up ("men's brains are wired to be better at physics" etc.). These unreliable "empirical generalisations"[30] typically prey on groups of people belonging to protected categories; whilst this essay focuses on gender, race and class (among others) have also historically been targeted by stereotypes[31]. Heuristic use of stereotypes is what paves the way to negative identity prejudice — that is: a generalisation based on the belonging of an individual to a social group which is impervious to counterevidence "owing to an ethically bad affective investment"[32]

Results produced by biased word-embedding, content moderation and résumé filter systems seem to use similar heuristic shortcuts: producing results which prey on negative stereotypes of women, historically a centrefold of our collective social imagination. Through the stereotypical portrayal of women as "homemakers" and "divas", preventing them from accessing traditionally male-dominated jobs (such as software engineering) and censoring them on platforms of public speech, it would be fair to say AI is spreading negative identity prejudice. In the next section I also argue this

21. Reuben Binns et al., "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation," in *9th International Conference on Social Informatics*, ed. Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Springer International Publishing, 2017), 3.
22. Binns 2017, 6.
23. Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," 2018, accessed May 5, 2021, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
24. Fricker, 14.
25. Fricker, 15
26. Fricker, 15.
27. Fricker, 31.
28. Fricker, 31
29. Fricker, 31
30. Fricker, 32.
31. Fricker, 33.
32. Fricker, 35.

negative identity prejudice deleteriously affects women's epistemic status as they become tokens of suspicion and distrust, preventing them from being treated as reliable and valid possessors and/or conveyors of knowledge. For example, women may be seen as less qualified for software engineering positions, despite the fact they have the same credentials as their male competitors. However, their bid to apply might be assigned lower credibility in virtue of the fact that they are women, owing to their 'lesser aptitude in the scientific domain' or 'their final aim to become homemakers' etc. Having established an initial link between artificially intelligent systems and how they tie into epistemic injustice, I will now expand on how the association of the two has damaging effects on women's potential for growth and self-determination.

4 The Harms of Epistemic Injustice

Thus far, we have elucidated Fricker's notions of identity power, stereotype, negative identity prejudice and examples of AI's involvement in these. However, defining these terms doesn't capture why or in what capacity AI harms women's epistemic standing in our social framework. In this section I aim to further explain why stereotypes and prejudice are detrimental to self-development and self-actualisation as well as illustrating how these occur in tandem with the results produced by AI systems. As mentioned in the previous section, stereotypes and prejudice both operate on a practical social plane as well as an imaginative one, both of which require levels of social coordination to persist. Though harbouring the potential for change, our social imagination also bears the marks of previous prejudices, which can insinuate themselves into our collective social practices, thus becoming systemic[33]. But why *is* prejudice bad for our social discourse and collective imagination? Recall that prejudice operates on stereotypes, which in turn are "empirical generalisations". I believe that these are at the core of the prejudice problem: they hinder the epistemic development of a given social group by failing to treat the people belonging to it as *individuals*, missing out on member-specific truths on account of applying stereotype-heuristic shortcuts when making credibility judgements (whereas those belonging to these groups are likely to present their own idiosyncrasies and skills)[34]. For example, whilst some women may be more emotional than others, generalising that all women are hysterical in virtue of belonging to the social category "woman" will cause their lived experience not to be taken seriously even in cases where they express upset or hurt. This has been known to occur in doctors' offices, for example, where women's pain is often dismissed as melodrama, producing flawed diagnoses.[35] Not only is this harmful to someone personally, but it also presents an obstacle to understanding the truth[36]. This was apparent in the examples I mentioned, whether with biased software denying women certain job opportunities or their speech being censored for appearing too "emotional" on online fora. Here, the exclusion of women from contributing knowledge into the public domain is extremely detrimental not only in virtue of their human dignity but it also impeaches on their freedom of speech. By being unable to voice their knowledge, the silence they are forced into thwarts their chances of *potentially* being considered a credible knower.

Drawing on the Kantian notion that freedom of speech is essential to the authority of reason, Fricker asserts that censorship deprives us of the *capacity* for reason: the instrument of the human mind that has historically differentiated us from animals and become synonymous, if not nearly identical, with the human ability to convey knowledge[37]. Imparting knowledge is also instrumental to set the boundaries and parameters through which our social discourse unfolds. Those who have the opportunity to communicate knowledge, and are attributed with due credibility when they do so, have the power to dictate what constitutes acceptable discourse. This is why the censorship of women by AI content-moderation systems is so harmful: it betrays an insensitivity and indifference to the nuances of the way women express themselves and sets "norms of acceptability"[38] which are consolidated through sexist standards. At its core then, conveying knowledge is a human right: the opportunity to contribute and add to the social discourse is loaded with meaning and human value. When women are wronged in this (epistemic) capacity through biased software, it could be argued that their value and "currency" as human beings is belittled. This has the potential to stunt one's personal development, what Fricker calls a "process of social construction"[39], cramping the avenues in which a given individual can grow personally. In a sense, this

33. Fricker, 87.
34. Fricker, 33.
35. Ian James Kidd and Havi Carel, "Epistemic Injustice and Illness," *Journal of Applied Philosophy* 34, no. 2 (2016): 172–190.
36. Fricker, 44.
37. Fricker, 44.
38. Binns 2017, 2.
39. Fricker, 59

process is centred on one's ability to make sense of one's lived experience, as gaining an understanding of this is a crucial step on the road to acquire knowledge, but how does AI impact this? It does so by excluding women from accessing the tools of social interpretation needed to make sense of some of their lived experiences. The following section will tackle this in further depth.

5 The Hermeneutical Consequences of AI and Epistemic Injustice

Though most instances of epistemic injustice detailed in this essay could be classified as testimonial injustices (*i.e.* attributing a negative, systematic, ethically- culpable credibility deficit to someone[40]), I can also envision AI systems perpetrating what Fricker terms "hermeneutical injustice", the second sub- category of epistemic injustice where the individual affected lacks the resources of social interpretation needed to makes sense of a consequential portion of their social experience[41]. Though it hasn't been in the purview of this essay to illustrate a taxonomy of epistemic injustice, the hermeneutical sphere of the latter presents interesting repercussions when explored through the lens of AI.

Let's trace back to the example of automatic résumé filters: a woman applies for an engineering job at a tech company with some stellar credentials which set her up to be a great candidate for the position she has applied for. She submits her résumé, which gets screened by an automatic résumé filter. This may find multiple instances of the word "women's" on her submission. The machine's model will most likely have been trained on the company's previous hiring data and as the job she has applied to has a tradition of hiring men for the position, the AI is highly likely to reject her application on the grounds of her being a woman. The woman is unsuccessful in her application but doesn't know an AI filtered her résumé. At this point, she is likely to feel bewilderment at being rejected from the job she was qualified for.

If she isn't aware of the machine algorithm used to make this decision and isn't privy to or can't understand the biases which have been encoded into the system, she will struggle to comprehend why she was rejected. I believe that understanding how this AI works and knowing that she was rejected by one are interpretive resources she would need to make head or tail of what Fricker describes as "an experience which it is strongly in her interests to render intelligible"[42]. This has the markings of hermeneutical injustice (though I use a slightly broader definition of "hermeneutical" than Fricker's): the rejected job applicant will probably feel confused, vulnerable and unsure about the integrity of her credentials, and to a certain extent, her identity as a person. The exclusion of a social group from professions in the hermeneutical sphere (hermeneutical marginalisation), is one but many realms in which hermeneutical injustice can track an individual (if the injustice is consistent in multiple different social domains in addition to the hermeneutical, the injustice is said to be systematic[43]).

An important clarification to make is that hermeneutical injustice is not carried out by an *agent* but rather is a lacuna intrinsic to our hermeneutical resources, caused by identity prejudice in the hermeneutical domain[44]. As AI is developing at break-neck speed but is still relegated to the realm of tech specialists, understanding the subtle but insidious ways it affects women's lived experience as a social group is a hermeneutical resource which is not available to the wider public.

Additionally, the problem of opacity in machine learning (ML) algorithms threatens to make this a reality for specialists too – at times, the way ML systems solve problems is not wholly intelligible even to those who have programmed them (also known as "The Black Box Problem").[45] Worryingly, the examples of epistemic injustice perpetrated by AI that I have presented all exclude the individuals they affect ("women") from being considered rational and significant

40. Fricker, 28.
41. Fricker, 148.
42. Fricker, 148.
43. Fricker, 156.
44. Fricker, 169.
45. Carlos Zednik, "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence," *Philosophy & Technology* 34 (2021): 3.

members of society through the determent of their contributions from generating social meaning[46]. As I have outlined, epistemic injustice is persistent across multiple domains which extend beyond the social: preventing someone from creating meaning will also prevent them from creating significance and value, thus, I agree with Fricker's claim that this type of injustice has the potential to stifle the growth of important aspects of one's personal identity[47].

Through the perception of knowledge as a social currency of sorts, we are effectively buying into the idea that, in virtue of embodying a certain identity, some individuals (and their input) are intrinsically more valuable than others.

6 Mountain Out of a Molehill

Though there is empirical backing showing that AI is gender-biased, things get a bit murky when we try to assess the gravitas with which we should regard the outputs generated by these systems. The idea that AI makes consequential decisions affecting the epistemic standing of women *through* epistemic injustice rests on the assumption that negative identity prejudice is at play when an AI produces a sexist result. However, if we refer back to Fricker's definition of negative identity prejudice, one of its necessary features is that it must be ethically bad, or at the very least that there must be an "ethically bad affective investment" at play[48]. This begs the obvious question of whether AI can be ethically culpable or even said to have a morally questionable affective investment, thus, the larger question becomes whether artificially intelligent systems can be regarded as agents responsible for carrying out epistemic injustice against women. If they cannot be held ethically culpable, then perhaps it would be misguided to claim that their outputs embody negative identity prejudice in the way Fricker would have envisioned it. As such, it would be incorrect to say that AI carries out epistemic injustice in accordance with the parameters by which we have defined it in this essay.

Whilst the topic of moral and agential responsibility of AI deserves to be expounded further, this essay is not purporting to claim that software is *generating* or *causing* epistemic injustice, rather, that it is *perpetuating* it. These systems nurture preexisting stereotypes present in our collective social imagination and repackage them to symbolise impartiality through the guise of objective machine-driven outputs for the tasks they are programmed to perform. Though the focus of this work has mainly been on the outputs generated by AIs, these are arguably not at the root of the problem, instead, the issue seems to lie within the operation of training and refining the processes used by the systems to reach these (problematic) outputs. Gender bias in AI isn't a direct cause of epistemic injustice, but rather an *enforcer* — it lays the groundwork for a culture where epistemic injustice is justifiable within the social infrastructure we live in. The epistemic "wrongdoing" itself is perpetrated by those who buy into this and embrace the stereotypes promoted by biased software. Thus, the focus of the central argument I have put forward does not come down to *culpability* but rather *influence*.

More often than not, human judgement will come between the output of the AI machine and executive action; the result produced by an AI will be assessed critically by specialists before crystallising into something concrete. Thus, it may appear as though we are placing too much importance on the results churned out by AIs, despite them not being the direct cause of epistemic injustice. However important identifying causes may be, I believe this line of thinking is unconducive to recognising the real problem: it presupposes causation of the injustice to be the most relevant arbiter of importance for AI's involvement in the sphere of epistemic injustice. Though causation is undoubtedly key, I believe influence – i.e. AI's ability to reinforce, spread sexist stereotypes and censor women on a massive scale and creating a fertile environment for epistemic injustice to thrive, to be just as bad, if not worse, in its ability to justify and promote the spread of injustice, as it does so in a more insidious manner. Predominantly only discernible by a select few with the epistemic resources to do so (*i.e.* AI specialists). Problematically, this promotion of injustice occurs under the pretence of impartiality. As such, having established that AI is a powerful influencer of credibility outcomes for women in the province of epistemic injustice, (with some unpalatable consequences) how do we move forward? In the next paragraph, I propose a good place to start would be to conceptualise what a "fair" AI could look like and what a good way to conceptualise it could be.

46. Fricker, 153, 161.
47. Fricker, 169.
48. Fricker, 36.

7 Building a Fair AI: New Solutions for New Problems

Most of us have an intuitive notion of what fairness entails, nevertheless, setting parameters to program a "fair" AI has proven to be difficult and the subject of much debate. In this section, I aim to expand on Binns'[49] account of fairness for machine systems and tie it to Rawls' concept of justice as fairness[50] as both present a compelling case to examine how to curb the AI's perpetuation of epistemic injustice. Naturally, questions arise when the idea of "fairness" is broached: what should the focus of a fair AI be (*i.e. should it maximise benefits for most or minimise harms for most?* etc.)? And when a focus is established, how could our chosen metric be quantified and practically applied to AI? The second question isn't as pertinent to the focus of this paper, as such it will not be tackled here. To address the first, however, Binns makes a compelling argument by positing that egalitarian norms can elucidate how algorithms are "unfair"[51]. Egalitarianism is an interesting avenue to explore as its doctrine lends itself well to answering "what should the focus of a fair AI be?" in focusing on the question "the equality of what?". For example, in the case of content moderation — should the chance one has of being censored be equalised regardless of gender? Or should equalising outcomes of the censorship be our focus?

The open-ended nature of this "the equality of what?" creates a debate concerning the application of egalitarian mores in "different social contexts"[52] and whether our answer to this question should be tailored according to the domain an AI system is operating in. Rawls engages with this question in the second principle of his thesis of justice as fairness. He believes we all have the right to a basic set of liberties and that these should provide the greatest benefit they can to disadvantaged members of society[53]. In addition, these basic liberties should be enacted by fostering conditions of equality of opportunity[54]. In Rawls' view, the latter hinges on the notion that everyone should have the same educational and economic opportunities regardless of the social "category" they were born into, as this is arbitrary (woman/man, white/black, rich/poor etc.)[55]. These things considered, we can explore a practical implementation of Binns' and Rawls' ideas using automatic résumé filtering software. Modelling equality of opportunity into these systems could avoid sexist outputs, for example, where software penalises résumés on the basis of containing words such as "women's" and "women's chess captain"[56][57]. If AIs were designed with "equality of opportunity" as a guiding principle, the epistemic injustice which undermines women as possessors and conveyors of specialist engineering knowledge in this case could be corrected, at least partially, as it would allow women to access the same resources (in this instance, jobs) as their male counterparts, granting them the chance to contribute to social discourse in a way that is deemed meaningful. Similarly, if we gave women equal opportunity to men in the arena of self-expression (without AI-mediated censorship), we could encourage a milieu receptive to the ways in which women communicate — enabling them to impart knowledge in a meaningful way. Though Fricker's theory of epistemic injustice is built on attributive (*i.e.* with the attribution of credibility) rather than distributive lines, I can envision "equality of opportunity" to be a good metric guiding attribution of justice on a case to case basis. After all, the adjustments which will need to be made to an algorithm's output will differ on the basis of other protected characteristics such as class, race or creed. However, knowing that our end goal is to guarantee that all women to have the opportunity to receive the credibility they deserve *qua* knowers will be central to stemming the perpetuation of epistemic injustice via AI.

Though this may not be perfectly aligned to what Rawls envisioned, as "equality of opportunity" predominantly concerns itself with economic goods, I still believe an interesting connection could be made between his second principle of justice as fairness and Fricker's conception of an "economy of credibility" and an "economy of collective hermeneutical resources". These terms are defined loosely in her work; however, if we envision knowledge as a currency of sorts (to accrue social status, power etc.), then a "credibility" economy or one based on "collective hermeneutical resources" will concern itself with the production, elaboration and consumption of knowledge. It seems to me that when women strive

49. Reuben Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ed. Sorelle A. Friedler and Christo Wilson, vol. 81, Proceedings of Machine Learning Research (PMLR, 2018), 149–159.
50. Leif Wenar, "John Rawls," in *The Stanford Encyclopedia of Philosophy*, Summer 2021, ed. Edward N. Zalta (Metaphysics Research Lab, Stanford University, 2021).
51. Binns 2018, 6.
52. Binns 2018, 6.
53. Wenar.
54. Wenar.
55. Wenar.
56. Wenar.
57. Dastin.

to be regarded credible *qua* knowers or achieve a full understanding of their lived experience they are effectively pursuing an epistemic currency of sorts, which allows their word to have due influence in the collective social discourse. Though coding these principles into an AI might prove to be complex – I believe it could be a viable solution to address the crux of the problem. AI is discriminatory and does not endow women with equal opportunity to create social meaning, doing so through faulty attributions of credibility and lack of epistemic resources available to them. In a competitive credibility economy, women must be able to fairly compete for the same resources as their male counterparts, and I believe a good place to begin would be by giving them equal opportunity to do so.

8 Conclusion

The problem of gender bias in artificially intelligent systems is rapidly gaining recognition. The increasing weaving of these technologies in our social and economic fabric has made it clear that further research of this phenomenon is warranted to address it and its ramifications. This essay proposed one such ramification to be epistemic injustice. I posited that gender-biased AI plays a role in perpetuating it by differentially censoring women on public platforms, sustaining sexist stereotypes which harm their credibility as knowers and preventing them from accessing the same opportunities as men on the basis of biased credibility judgements. Based on Fricker's case for epistemic injustice I have endeavoured to show that being able to impart knowledge is crucial for a person's self-development and that through their biased outputs, machine systems play a role in preventing women from creating valuable social meaning. This could be seen as an inflated view as, after all, AI cannot be said to commit an epistemic injustice as (per Fricker's definition) there is a debate surrounding its moral accountability. However, though AI may not be directly causing epistemic injustice it is creating an environment where it appears permissible and even normalised to do so, which, as I have argued, is highly problematic. In conclusion, I believe that to move forward we must build fairer AIs. I have argued that through Binns' and Rawls' ideas this would entail using an egalitarian framework to encode equality of opportunity into machine systems. And hopefully, loosen the chokehold of sexist stereotypes in our daily practices of receiving and producing knowledge.

References

Binns, Reuben. "Fairness in Machine Learning: Lessons from Political Philosophy." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:149–159. Proceedings of Machine Learning Research. PMLR, 2018.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." In *9th International Conference on Social Informatics*, edited by Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, 405–415. Springer International Publishing, 2017.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, 29:1–9. Curran Associates, Inc., 2016.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. PMLR, 2018.

Cirillo, Davide, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Anna Azzurra Gigante, et al. "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare." *NPJ Digital Medicine* 3 (81 2020): 1–11.

Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women," 2018. Accessed May 5, 2021. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

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

Hub, IBM Cloud Learn. "What is Artificial Intelligence," 2020. Accessed April 20, 2021. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. "Artificial intelligence in healthcare: past, present and future." *Stroke and Vascular Neurology* 2, no. 4 (2017): 230–243.

Kidd, Ian James, and Havi Carel. "Epistemic Injustice and Illness." *Journal of Applied Philosophy* 34, no. 2 (2016): 172–190.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. Accessed May 24, 2022.

Pohlhaus Jr., Gaile. "The Routledge Handbook of Epistemic Injustice." Chap. Varieties of Epistemic Injustice, edited by Ian James Kidd, Jose Medina, and Gaile Pohlhaus Jr., 13–26. Routledge.

Quach, Katyanna. "MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs," 2020. Accessed May 6, 2021. https://www.theregister.com/2020/07/01/mit_dataset_removed/.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2021.

Wenar, Leif. "John Rawls." In *The Stanford Encyclopedia of Philosophy*, Summer 2021, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University, 2021.

Zednik, Carlos. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34 (2021): 265–288.