Macalester Journal of Philosophy

Volume 20 | Issue 1 Article 13

6-21-2012

# Contact Page

# MACALESTER COLLEGE

## JOURNAL OF
# PHILOSOPHY


VOLUME 20
SPRING 2011

Macalester Journal of Philosophy
Volume 20
Spring 2011

Table of Contents



*Edited by Drew Van Denover and Genevieve Kaess*

# REAL RESPECT: A REJECTION OF RICHARD MILLER'S PATRIOTIC BIAS IN TAX-FINANCED AID

*Gerbrand Hoogvliet*

**Abstract** This paper analyzes Richard W. Miller's argument for favoring compatriots in the allocation of tax-financed aid. It argues that Miller's patriotic bias is derived via an incorrect framing of the problem. It furthermore contends that Miller's notion of equal respect is too uninformative to ground such a patriotic bias. A better definition of respect in terms of human rights is offered. This definition is more informative but fails to uphold the stringent bias Miller argues for.

National borders occupy a curious position in political philosophy and ethics. Their existence and location is often the result of mere historical accident. Yet, despite this arbitrary nature, the nation states defined by these borders are often chosen as the primary actors in theories of international relations. Similarly in ethics, there is a tension between the fact that citizenship seems morally arbitrary, insofar as it is usually bestowed upon persons at birth, and on the other hand the moral obligations that participation in a particular society seem to give rise to. In the context of global poverty national borders take on another moral dimension since they often, as Michael Blake puts it, "divide not simply one jurisdiction from another, but the rich from the poor as well"[1].

[1] Michael Blake, "Distributive justice, state coercion, and autonomy", *Philosophy and Public Affairs* 30, no. 3 (2001), 257.

Given the grim facts of poverty in many parts of the world, the question of whether wealthier nations are morally allowed to favor their own citizens over foreigners in dire need becomes an important one.

Richard Miller, in his contribution to the anthology *The Political Philosophy of Cosmopolitanism*, entitled "Cosmopolitan Respect and Patriotic Concern", provides a universalist defense of such a favoritism. He argues that on the basis of the principle of equal respect for all persons we are in fact obligated to prioritize our compatriots when it comes to tax-financed aid. He argues that a violation of such a patriotic bias would entail disrespectful treatment of our fellow citizens and would lead to an excessive loss of social trust. Given that breaking the principle of equal respect is wrong, violation of the patriotic bias is also wrong. We are thus morally obligated to prioritize compatriots in the administration of such aid.

In this paper I will argue against the position put forward by Richard Miller. I will begin with an exposition of his argument. For the benefit of the reader I will also provide a brief explanation of concepts found in John Rawls's *Justice as Fairness: A Restatement*, that are important to a proper understanding of Miller's position. I will then provide my own critique, focusing firstly on what I hold to be an improper framing of the issue, followed by a more fundamental criticism of the notion of equal respect used by Miller. I will show his definition of equal respect to be uninformative and anemic and will proceed to redefine this concept in a more substantial way by appealing to the philosophical literature on human rights.

**Miller**

In his paper, Miller aims to provide "a universalist justification of the patriotic bias in aid."[2] Universalism here refers to a position similar to cosmopolitanism, which takes human beings as 'the relevant unit of moral concern'. It is mainly defined in contrast to what Miller calls particularism, which is a view maintained by philosophers such as David Miller and Michael Sandel, who ascribe intrinsic value to communities of persons such as nations. For particularists, the defense of patriotism is usually based on some notion that it benefits the community or the nation state. Since Richard Miller rejects a view of nations as intrinsically valuable he cannot make a similar claim. In fact, because he adopts the universalist view of all persons as having equal moral value, he commits himself to the use of universal principle that applies to all persons. This principle is that of equal respect.

In order to establish a patriotic bias, however, he first has to identify what such a bias consists of. He points out that the patriotic bias is really a combination of two biases: an attention bias and a budgetary bias. To establish the attention bias he has to prove that we are justified and indeed obligated to pay more attention to the needs of our compatriots than to the needs of foreigners. The budgetary bias is then the working out of this attention bias in terms of assigning aid and simply means that the majority of our tax-financed aid is indeed spent on compatriots. He recognizes that he has to establish the attention bias before he can claim the budgetary bias.

---

[2] Richard W. Miller, "Cosmopolitan Respect and Patriotic Concern," in *The Political Philosophy of Cosmopolitanism*, ed. Gillian Brock and Harry Brighouse (Cambridge: Cambridge University Press, 2005), 127.

**Equal Respect**

In establishing the principle of equal respect, Miller makes an appropriate distinction between respect and concern. Whereas most of the literature conflates these two terms, he defines them separately. Concern, for Miller, applies to personal relationships such as between family members, friends etc and signifies a deep level of caring for the well being of others. I think Miller rightly restricts this type of sympathy to those who we are personally acquainted with. As an example, he states that although he owes equal respect to his daughter and the girl across the street, he is not required to have the same level of concern for the latter. I think this is a sensible distinction and it clarifies the task at hand: since concern covers all persons that we stand in a personal relationship to, the principle of respect is the one that will regulate our behavior to strangers domestically and abroad.

The equal respect that we owe to strangers has two main parameters:

1) One avoids moral wrongness just in case one conforms to some set of rules for living by which one could express equal respect for all.[3]

2) A choice is wrong just in case it violates every set of shared rules of conduct to which everyone could be freely and rationally committed without anyone's violating his or her own self-respect.[4]

[3] Ibid., 132
[4] Ibid.

The phrasing of these parameters is somewhat confusing, but in a nutshell they provide two conditions under which equal respect is violated. Under the first rule, it is morally wrong to choose a method of administering tax-financed aid that does not show equal respect for all. The second parameter claims that it is wrong to choose a way of distributing aid in a way that some persons could not self-respectfully accept. To use an example, if you and I were to start a lawn mowing business and I suggested that, even though we put in the same amount of work, I should get all the money, then that would not be an arrangement that you could self-respectfully accept.

Miller thus separates respect out into a *respect outward* and *respect inward*; respect for others and self-respect. Any administration of tax financed aid thus has to express and satisfy both forms of respect.

**Rawlsian Intermezzo**

At this point I think it will be beneficial to elucidate some concepts from John Rawls that are implicit in much of Miller's further discussion. Although Miller is not defending anything like a Rawlsian position, much of political philosophy is steeped in the tradition started by Rawls and it is therefore useful to have a basic understanding of some of the background concepts informing this discussion.

Rawls conceives of society as "a fair system of cooperation"[5] among free and equal citizens. Fairness is necessary

[5] John Rawls, *Justice as Fairness: A Restatement* (Cambridge: Harvard University Press, 2001), 14.

for Rawls because one does not choose what society one is born into, and exiting a society is extremely difficult if not impossible. Society is thus unlike other forms of association such as local communities, schools, clubs, church congregations etc. where membership can be given up if one is asked to uphold rules and practices that one is unwilling to support. Since no such an exit option exists for the nation state there is a more urgent demand for fairness.

Not only is societal membership largely involuntary, it also exposes persons to the coercive nature of the state. For Rawls "political power is always coercive power applied by the state and its apparatus of enforcement." [6] As citizens we participate in the creation of laws, which the state then enforces in our name. Justification is thus demanded both on the grounds that laws are enacted in our name as well as that laws are enforced upon us.

Given this nature of society and the demands for justification that it gives rise to, Rawls is particularly concerned with the well being of what he calls "the least-advantaged members of society."[7] It is easy to see why this is: given the coercive nature of the state and the near impossibility of exiting society, it is the worst off group that is most likely to feel trapped in a system that they would not voluntarily uphold. This group could certainly be coerced into cooperation, but the ideal of a just society would then have been forfeited. I take Miller's concerns about respect to also be focusing largely on this group, and for similar reasons.

[6] Ibid., 40
[7] Ibid., 43

**Loss of Social Trust**

Returning to Miller's argument, he claims that a failure to prioritize compatriots would entail a violation of the principle of equal respect. This violation comes about in two ways. First, without a patriotic bias, tax-financed aid is distributed in a way that does not express respect to all. Specifically, the least-advantaged members of society are not treated respectfully by their fellow citizens. This goes against the first parameter of equal respect that I stated above. The idea here seems to be that by not paying extra attention to the needs of disadvantaged compatriots we are treating them disrespectfully, which the first parameter holds to be wrong.

The second way in which a breach of equal respect comes about is through the inability of the least-advantaged group in society to self-respectfully accept such an allocation of tax-financed aid. Put differently, the least well off members of society could not choose a use of tax-financed aid that did not prioritize them and at the same time maintain their self-respect. The sacrifice required of them would be too large, the inequalities faced too stark. Since an allocation is imposed on them that they could not self-respectfully accept, parameter 2 of equal respect is violated and the allocation is thus wrong.

It is important to note here that the priority that Miller requires is a very strong one:

> [P]riority does not totally exclude support for foreign aid in the presence of relevant domestic burdens. Still, until domestic political arrangements have done as much as they can [...] to eliminate serious burdens of domestic inequality of life-prospects, there should be no

> significant sacrifice of this goal in order to help disadvantaged foreigners.[8]

To put the consequences of this patriotic bias in context, Miller presents us with three persons who present the three main stakeholders in the outcome of this discussion. Kevin is a corporate lawyer living in a rich suburb of New York. Carla lives in the South Bronx and earns a meager living cleaning other people's apartments. Khalid, finally, collects scrap metal and lives in a slum in Dacca, Bangladesh. Miller maintains that the patriotic bias and its consequences can be self-respectfully accepted by all three. As we stated above, Carla, as a member of the least-advantaged group in society, can self-respectfully accept a situation in which she is prioritized to the extent that Miller suggests in the statement above. Kevin also upholds the principle of equal respect since he is treating Carla in a respectful manner. Khalid, according to Miller, can also self-respectfully accept the patriotic bias that Kevin and Carla adhere to since he understands that both value the social trust that would be lost without such a bias. Kevin and Carla are also assumed to be treating Khalid respectfully, although Miller does not go into detail as to why that would be the case.

Naturally such a bias is a very convenient view for rich societies to hold since it reduces their obligations to foreign aid significantly. As Thomas Nagel points out in "The Problem of Global Justice", however, the fact that a theory is convenient doesn't make it false.[9]

---

[8] Miller, "Cosmopolitan Respect and Patriotic Concern", 134

[9] Thomas Nagel, "The Problem of Global Justice," *Philosophy and Public Affairs* 33, no. 2 (2005): 126.

There is, however, another reason to be suspicious about Miller's patriotic bias as based on the principle of equal respect. Note that changes in Khalid's level of deprivation do not change the bias. Miller chooses to think of him as a scrap metal collector in Bangladesh, but we could just as easily imagine him as living in a refugee camp in Chad, or working 70 hours a week in a coal mine in Brazil, and Miller's bias would remain unaffected. Also note that Khalid does not feature anywhere in Miller's argument prior to the establishment of the patriotic bias. The fact that Khalid's circumstances are not being taken into account at all makes it at the very least unlikely that he is being shown equal respect.

Deciding on the extent of a patriotic bias that is supposed to show equal respect to all can hardly be done without looking at the needs of foreigners, especially given the severity of global poverty. Although the facts of global poverty cannot, in and of themselves, decide the debate about patriotic bias, they can help pull it into focus. Thomas Pogge estimates that in the 15 years following the Cold War, 270 million people died from poverty related causes, an average of 18 million a year.[10] Against the backdrop of these grim facts, a theory that does not take into account the needs of the global poor can hardly claim to express equal respect for all.

In the next section I will present two criticisms of Miller's argument. The first focuses on a framing issue that I think skews the debate and misrepresents the trade-offs involved in reallocation

[10] Pogge, Thomas W. M. "From *A Cosmopolitan Perspective on the Global Economic Order*." In *The Political Philosophy of Cosmopolitanism*, edited by Gillian Brock and Harry Brighouse, (Cambridge: Cambridge University Press, 2005), 92.

of tax-financed aid. The second criticism is far more fundamental and proves that the principle of equal respect used by Miller is uninformative and stands in need of a better definition. I will consequently suggest a more informative definition grounded in contemporary political philosophy of human rights.

**Framing**

My claim here is that Miller gets the strong bias that he wants by the way he frames the reallocation of tax-financed aid. In short, my contention is that Miller implicit assumes the amount of tax-financed aid to be fixed, or determined at a point prior to the patriotic bias discussion. By doing this, any imagined change to the allocation of this aid becomes a zero-sum game between Carla and Khalid. The amount of aid is set, so any aid to Khalid will have to come out of tax money reserved for Carla. This places undue tension on the allocation decision as we are forced to choose between two persons clearly in need. Certainly, in absolute terms Khalid is worse off than Carla, but on the other hand Carla is forced to participate in a society with people like Kevin, which raises concerns of fairness domestically. The radically unequal income distribution in the United States only further aids Miller's argument.

My point is that this is an incorrect framing of the question. If we are really concerned with equal respect for all, we should not take tax aid as given, but rather as a function of the needs of Carla and Khalid and what is owed to them on account of this respect. If, for the sake of argument, we take Kevin as the sole tax payer, then the tax rate imposed on him should be set at a level at which both Carla and Khalid can self-respectfully accept the amount of aid

they receive. Framing the question in this way, I think Miller may still be justified in claiming that more is owed to compatriots on account of the coercive nature of the state. However, the amount owed to Khalid is likely to be much higher than what he has in mind. Thinking about the reallocation of aid in this way also makes more sense if we view it from Khalid's perspective. He is more likely to think of himself as being owed some type of aid by Kevin rather than by Carla, since Kevin is in a position to improve Khalid's life significantly, at little cost to himself.

This then raises the question of how much domestic and foreign aid would be sufficient for the satisfaction of the principle of equal respect and whether Kevin could self-respectfully accept such a tax burden. This is where the limitations of Miller's account become clearly visible, because the definition of equal respect that he uses is completely uninformative on this matter. It seems to me that Khalid could not self-respectfully accept the bias proposed by Miller, but how much would foreign aid have to increase for that to change? And if we found this amount, how could we tell if the tax burden required is one that Kevin could self-respectfully accept?

**Equal Respect Revisited**

The uninformative nature of the equal respect principle stems from the fact that Miller defines it in terms of respect. If we look again at the two parameters, we notice that they largely constitute an elucidation of the concept of equal respect. Miller effectively break it down into two components: respect-towards and self-respect. Parameters one and two deal with those respectively. However, the meaning and import of these

components remains unhappily vague as can be seen in the discussion at the end of the previous section.

I think current thought in political philosophy can provide us with more informative concepts of what equal respect entails. The one I shall focus on here is the recent work in philosophy of human rights, although Amarty Sen and Martha Nussbaum's work on the human capabilities approach is also a strong candidate.

**Human Rights as Equal Respect**

International human rights practice is commonly seen as motivated by the need to protect human dignity in some form or other. Although this idea of dignity is rather vague, a clear connection can be seen with the idea of respect. What we mean by equal respect is that we treat other persons as having a certain amount of equal intrinsic value. We regard them as worthy of moral consideration.

Recent works in the philosophy of human rights have expounded this idea of dignity and tried to give it more substance. They have established strong philosophical frameworks for thinking about the goal and content of human rights. The account given by James Nickel in *Making Sense of Human Rights* focuses on vital human interests that human rights are designed to protect. As such, human rights can be seen as necessary conditions for living a minimally good life. James Griffin's account in *On Human Rights* envisions them as protecting a person's liberty, autonomy, and basic standard of living.[11] Again, human rights are used to protect what we see as central to human life.

[11] James Griffin, *On Human Rights* (Oxford: Oxford University Press, 2008), 51.

I think that these accounts can help lend content to the concept of equal respect. Since human rights are necessary conditions for a minimally good life, violating them can rightly be seen as disrespecting the holder of that right. Human rights thus set a minimum standard for what equal respect for all persons requires: namely a guarantee not to violate human rights and a strong duty to help uphold and enforce them whenever one is in a position to do so at relatively low cost to oneself.

Applying this human rights definition of equal respect to Miller's account yields a very different outcome. For one, the patriotic bias can no longer be established by only considering the domestic case. Instead, equal respect demands an effort to guarantee the observance of human right for all persons both domestically and abroad.

Certainly I have only sketched an outline here of what such an approach to the allocation of tax-financed aid would entail. Further development of the idea of 'human rights as a standard for equal respect' is necessary in order to work out its exact practical implications. The duties of different well-off societies to help the global poor in having their human rights protected need to be coordinated and a reasonable limit needs to be placed on the burden that such duties can impose on these societies.

Nevertheless, it appears clear from the outset that any patriotic bias that claims to show equal respect on my definition of that term, would be quite different from the one argued for by Miller. It almost certainly calls for a greater transfer of aid from the per-capita rich countries to those in need. It does not preclude the existence of a patriotic bias in tax-financed aid, and in fact arguments for such a bias are probably justified. It does mean that

demands for equal respect will take precedence over any considerations of patriotic priority, as I have argued they should.

**Conclusion**

In this paper I have shown that Richard Miller's argument for a patriotic bias rests on an uninformative definition of the principle of equal respect. Due to the indeterminate nature of this principle, it is unclear what sort of patriotic bias can be justified. Whether different allocations of tax-financed aid show equal respect for all becomes a matter of speculation and personal interpretations of human psychology.

I have argued that the philosophical human rights tradition can provide us with a more substantial account of what respect for persons entails. Recent influential works by James Nickel and James Griffin suggest human rights as a protection of abilities and interests necessary for living a minimally good life. Given the important nature of human rights to individuals persons, I suggest that equal respect entails the non-violation of these rights as well as a duty to protect and uphold them when one can do so at little cost to oneself. I note that this is merely the first step in the creation of such an account and that more work is needed to establish clearly the demands 'human rights as a standard for equal respect' can and ought to give rise to. I do contend that any account based on this new definition of human rights will fail to establish a patriotic bias as strong as the one argued for by Richard Miller.

A last remark with regard to the question of tax-financed aid is in order. As Charles Beitz has noted, discussions in the field of global economic justice often make too much of the importance

of transfer payments from tax dollars.[12] More effective, efficient and lasting solutions to problems of economic inequality and global poverty can likely be found through the structural rearrangement of institutions such that they favor - or at the very least cease to actively disadvantage - the global poor. For the purpose of this paper, which was a response to Miller's patriotic bias in tax-financed aid, such questions of institutional reform were unfortunately not within our scope. Discussions in the field of global justice and cosmopolitanism can perhaps shine a light on fruitful solutions in that direction.

[12] Charles Beitz, "Cosmopolitanism and Global Justice," *The Journal of Ethics* 9, no. 2 (2005)

## Bibliography

Blake, Michael. "Distributive justice, state coercion, and autonomy." *Philosophy and Public Affairs* 30, no. 3 (2001): 257-296.

Beitz, Charles R. "Cosmopolitanism and Global Justice." *The Journal of Ethics* 9, no. 2 (2005): 11-27.

Griffin, James, 2008. *On Human Rights*. Oxford: Oxford University Press.

Nagel, Thomas. "The Problem of Global Justice." *Philosophy and Public Affairs* 33, no. 2 (2005): 113-147.

Nickel, James W. 2007. *Making Sense of Human Rights*. Malden: Blackwell Publishing.

Miller, Richard W. "From *Cosmopolitan Respect and Patriotic Concern*." In *The Political Philosophy of Cosmopolitanism*, edited by Gillian Brock and Harry Brighouse, 127-147. Cambridge: Cambridge University Press, 2005

Pogge, Thomas W. M. "From *A cosmopolitan perspective on the global economic order*." In *The Political Philosophy of Cosmopolitanism*, edited by Gillian Brock and Harry Brighouse, 127-147. Cambridge: Cambridge University Press, 2005

Rawls, John, 2001. *Justice as Fairness: a Restatement*. Cambridge: Harvard University Press.

# A Defense of Public Justification

*Simon Pickus*

**Abstract** Public justification is a concept presented by John Rawls as a way to legitimize political authority and to make fundamental political arguments. In essence, the principle holds that one should only present arguments that the opposition can reasonably accept, as opposed to appealing to a religious or political conception of the good. This paper seeks to present a cogent conception of the principle of public justification. The strengths of the principle will be explained, and the main critiques of the position will be examined and defended against. By this method, Rawls' conception of public justification can be shown to be a compelling and robust position.

Among the more pressing issues that have persisted throughout Western political and philosophical thought have been how political power can be rightly exercised, and how can political disputes between passionate parties be fairly resolved. Under what circumstances can the coercive power of the state be implemented in a way that is just and right? Bloodlines, military might, and religious mandates have all been appealed to as justification for political authority, but these are all answers monarchs and emperors have given to their already cowed populaces. Compelling answers to these questions presented by thinkers such as Hobbes, Locke, and Rousseau emerged in the form of reasonable consent of the governed as a legitimizing factor for

political authorities. In the 20th century, the widely-read political philosopher John Rawls best articulated the concept of public justification, a principle in which political authority can be considered legitimate only insofar as the reasons given for political action could be reasonably accepted by those who are governed. For this project, I will begin by giving a general overview of the position as conceived and presented by Rawls in his more recent works. I will follow this outline of public justification by explaining why this view is appealing and what problems within political thought it solves, or at least purports to solve. I will then present brief explanations of some of the more pressing objections to the theory, and will conclude with a refutation of these critiques.

**The Idea of Public Justification**

For Rawls, the principle of public justification is one that exists within what Rawls refers to as a well-ordered society. This means that, for him, any discussion of public justification presupposes a democratic society with a political culture that is pluralistic and has a commonly accepted conception of justice. In addition, Rawls notes that, "Accepting this conception does not presuppose accepting any particular comprehensive doctrine."[1] To clarify, "comprehensive doctrine" is a Rawlsian term for a complete conception of the moral good and a thorough set of values. Although these are not by necessity comprehensive, what is important about them is that they comprise a set of values and a conception of the moral good. Some examples of comprehensive doctrines are religious beliefs and moral philosophical codes such

[1] John Rawls, *Justice as Fairness: A Restatement* (Cambridge: Harvard University Press, 2001), 26.

as utilitarianism. Here Rawls is emphasizing that the principle of public justification is distinct from any one conception of the good or set of moral values. It does not presuppose a religion or ethical code, and does not need to. As it is meant to function within a society that has a plurality of comprehensive doctrines that its citizens accept, public justification is compatible with all reasonable conceptions of the good.

It is important here to note the particular meaning of "reasonable" in this context, as it is a conceptually significant term. For Rawls, "…reasonable persons are ready to propose, or to acknowledge when proposed by others, the principles needed to specify what can be seen by all as fair terms of cooperation."[2] By this Rawls means that to be reasonable is to act fairly and to seek cooperation and the resolution of disputes. A reasonable person will not enter into an agreement knowing that they will later violate that agreement, not will they staunchly refuse any attempt at resolving a disagreement. Additionally, reasonable people will seek to end conflicts and live peaceably, even if doing so is not always in complete accord with their rational self-interests. Acting reasonably is, as Rawls sees it, distinct from acting rationally, although in no way does reasonableness preclude rationality. It is very possible, however, to act rationally and unreasonably at the same time. An example of this would be a person who enters a long-term agreement and immediately forsakes that agreement when they see a way to derive some advantage from it. Another way to conceptualize this distinction is in the context of rational self-interest. To act in accord with rational self-interest is always

[2] Rawls, *Justice as Fairness*, 7.

rational but not always reasonable. The example of the tragedy of commons demonstrates that rational self interest leads to what Rawls would call unreasonable behavior, because it does not indicate a desire for fair cooperation. Rawls' conception of the reasonable, I find, agrees in large part with commonly held intuitions of what it is to act reasonably.

The principle of public justification, once established in the Rawlsian political context, is the vehicle for those with political disagreements to discuss and resolve their disputes in ways that are reasonable and acceptable to all involved. As Rawls explains, this principle allows people and groups to "…justify to one another their political judgments: each cooperates, politically and socially, with the rest on terms all can endorse as just. This is the meaning of public justification."[3] Here Rawls explains the very basic idea of the public justification principle.

People within a well-ordered society, or any developed democratic society as we would recognize today, will inevitably disagree with each other and their leaders on their political and social policy judgments. This alone is difficult to dispute. There are many reasons, even within a well-ordered society with a shared conception of justice, for these disagreements, such as what Rawls refers to as the plurality of comprehensive doctrines. He claims that, "…a diversity of conflicting and irreconcilable yet reasonable comprehensive doctrines will come about and persist…This fact about free societies is what I call the fact of reasonable pluralism."[4] Once the aforementioned disputes arise, public justification acts as

[3] Ibid, 27.
[4] Ibid, 34.

a mechanism for their resolution. People and groups justify their political judgments by presenting arguments that their opponents can reasonably endorse as a means of making their views plausible within the worldviews of the other. Using public justification, they appeal not to their conception of the good, such as, for example, the principle of utility or the intrinsic value and dignity of a human being, but rather they appeal to political values and reasons they both share so as to cooperatively come to a conclusion. In this way political disputes can, ideally, be solved in such that all can reasonably accept the conclusion without having to violate their closely held values and beliefs. Rawls goes on to note that, "Public justification proceeds from some consensus: from premises all parties in disagreement, assumed to be free and equal and fully capable of reason, may reasonably be expected to endorse."[5]

The general aim of this principle, then, is to provide a way for political judgments to be justified without appeal to reasons that the disagreeing party would never accept. A utilitarian could never convince a Kantian that a political moral dilemma can be solved using the principle of utility, no more than an Orthodox Jew could appeal to his or her religious tenets to convince a political opponent who is an adherent of Islam. No matter how dearly someone holds their conception of the moral good, they will not be able to offer compelling arguments to me if I do not agree with that idea of the good. They would need to find a set of criteria we both accept. By avoiding argument entrenched in the values of a comprehensive doctrine, public justification aims to avoid some of the persistent and pressing disagreements that have plagued

[I] Ibid, 27.

political discourse. Additionally, it reinforces political cooperation and reasonable discourse in a way that is consistent with a functioning democracy.

One important distinction that Rawls emphasizes is that public justification does not have a basis in simple agreement. What sets public justification aside as unique is its appeal to a common ground of reasonable arguments based, in part, on a shared conception of justice that allows for important political disputes to be fairly solved. Rawls himself states that, "It is this last condition of reasoned reflection that, among other things, distinguishes public justification from mere agreement."[6] Here Rawls shows the true importance of justifying political positions by presenting reasons anyone could reasonably accept. It is this aspect of public justification that sets it apart and, as I will now explain, it is this aspect that makes the principle of public justification appealing.

**Why Public Justification is Compelling**

The theory of public justification has a variety of strengths that make it a very compelling way to approach political discourse and legitimacy. The first largely intuitive main strength of public justification is that it serves as an alternative to tyranny and oppression, and as construed here does not allow for tyranny or oppression of any sort. The very nature of public justification does not allow for any sort totalitarian coercive rule that is imposed on the populace of a nation unwillingly. This aspect of public justification, though simple and straightforward, is a significant

[6] Ibid, 29.

point in its favor.

A second way in which the principle of public justification is strongly compelling is that it provides a way to solve political disputes that otherwise seem too divisive or too deeply entrenched in moral values for either party to possibly accept the other's position. This is particularly relevant to American politics, and similar systems, in which there is a political culture of such profoundly divided adversarial fervor that a resolution between the adversaries, in this case the two political parties, seems completely unfeasible. Joshua Cohen, a prominent contemporary political philosopher, echoes this sentiment when he notes, "The more immediate concerns come from the pathologically polarized state of political discourse in the United States."[7] He goes on to state that the intention of politics is to confront and overcome important and pressing issues relating to people and what they value in their lives, which is significant because "…public reason arguably provides a more promising basis than polarized disagreement for doing the works of politics, and…decent and inclusive political life is not only a profoundly important good, but a painfully fragile one."[8] In essence, the principle of public justification allows us to do the important work of politics without being hobbled by the vehement political culture that currently exists in the U.S. All that is required for this to work is that those engaged in political arguments accept that giving conceptions of the good as criteria for political decisions is not only unreasonable but disrespectful, as it is essentially a demand that political opponents defer to one's

[7] Joshua Cohen, "Politics, Power, and Public Reason" (paper presented at the UCLA Legal Theory Workshop, Los Angeles, California, April 17, 2008): 2.
[8] Cohen, "Politics, Power, and Public Reason." 3

comprehensive doctrine. Were politicians and pundits to accept this burden of respect and consider the practical advantages of public justification, we would not be stuck in such a partisan rut. In this case, public justification is compelling in that it avoids this issue by leading the disputing parties to converse using reasons that the other side might reasonably accept. At the very least, this principle presents the possibility of progress beyond the partisan impasse that some see the United States to be stuck in, and in this way public justification is compelling to American political thinkers.

A third reason that public justification is appealing lies in the distinction between rational and reasonable. As a method for justifying political positions and authority, public justification as presented by Rawls prioritizes the reasonable over the rational. To some, the appeal to discussion between reasonable people without emphasizing acting in strictly rational ways lies in the avoidance of prisoner's dilemma-type situations. By this I mean that for some, political discourse is problematic because it can be said to be populated by those who act strictly in their own interests and the interests of their associates; people who act in ruthlessly calculating ways. Public justification, on the other hand, ensures political discourse in which nobody is trying to trick their opponents, but rather encourages reasonable people to make genuinely persuasive arguments so as arrive at resolutions of political disputes. This emphasis on reasonableness is appealing to some because it presents a less adversarial, more cooperative method of dealing with political disagreements. As an environment focused on the genuine resolution of the issues in reasonable and productive ways, public justification is an

appealing principle.

A fourth strength of public justification is the way in which it provides a means for the maintenance of both legitimacy and stability in contractarian societies, those societies with a basis in some sort of founding agreement or governing document. A very real issue for these societies is that in several generations that society will be comprised of citizens who were not party to the original contractarian agreement. In a Hobbesian society, for example, once this point is reached, and there is no effective institutional way to change the society, then it is only a matter of time until circumstances change such that sufficient people reject the original contractarian agreement and the foundations of the society collapse. For this reason, there will come a point at which the members of the society no longer see a compelling reason to continue to submit to the coercive powers of the state granted by their ancestors. Public justification becomes appealing in this circumstance because it provides a plausible means for the contractarian society to change according to the wills of its citizens without a fundamental threat to its stability. Since the society's basic institutions are now mutable according to the will of the current populace, general discontent with the contractarian nature of the state is no longer an issue. In this way public justification is a compelling principle to those who adhere to contractarian conceptions of statehood.

**Objection 1: Begs the Question**

In addition to its many compelling features, the principle of public justification has significant objections to contend with. To begin, it is necessary to clarify the concept of public reason and its

interaction with public justification. As Rawls puts it, public reasons are, "…the political values covered by the political conception of justice (or one of a suitable family of such)."[9] In essence, public reasons are those we can give to justify our actions and positions to others within our society who therefore share our basic political conceptions. As Cohen puts it, "…the ideal of public reason says that in our political affairs…justification ought to be conducted on common ground…common ground provided by considerations that participants in the political relations can all acknowledge as reasons."[10] Simply put, public reason is the vehicle of public justification; reasons that are publicly justifiable are discussed using public reason. It is the form of reason we use to justify our political judgments to others. In this sense a discussion of public reason goes hand-in-hand with one of public justification, and a rejection of public reason is a rejection of public justification.

The first objection I will address comes from a writer named Bruce Brower of Tulane University. In his article *The Limits of Public Reason,* Brower analyzes several ways in which Rawls can make public reason, and by extension public justification, compelling to those who do not accept the priority of the reasonable. If I can refute any one of these, it would show that Rawls' project does not succumb to the limitations Brower presents. I am choosing to address one of these lines of argument, in which Brower claims that the demands of public reason violate equal respect and can be shown to be compelling only to those that

[9] Rawls, *Justice as Fairness,* 90.
[10] Cohen, "Politics, Power, and Public Reason." 1

already accept the ideal of the reasonable. In other words, Brower argues that the case for public reason begs the question. Brower argues that the demands of public justification violate equal respect because they require people to abandon values and beliefs that are deeply important to them. As he writes, "Treating others equally and acting autonomously…requires us to ignore an important part of our character…"[11] Here Brower is arguing that in requiring that people not use their conception of the good to make fundamental political decisions Rawls is asking them to forsake something too important to simply discard. Brower goes on to argue that proponents of public justification, "…do ignore something 'constitutive' of our persons: that we care deeply about our conceptions of the good and associated justifications…The problem is…Rawlsian arguments will be acceptable only to those who have already approved the…ideal of the reasonable person."[12] This is the meat of the objection that Brower presents. Rawls fails to show equal respect because he devalues peoples' constitutive values on the grounds that they are not publicly acceptable reasons to give in a political sense. Because of this, Brower feels that Rawls is saying that people should not give morally-grounded justifications, and should rather give public justifications, which are more reasonable. But, Brower claims, this requires that someone has already accepted the priority of the reasonable. Because of this, public justification is only compelling to those who already accept it.

This objection is not as strong as it first appears, and it in

[11] Bruce W. Brower, "The Limits of Public Reason," *The Journal of Philosophy* 91 (1994): 14.
[12] Brower, "The Limits of Public Reason," 15.

fact undermines itself. There are two primary claims to deal with: the demands of public justification violate equal respect, and public justification is not compelling to those who have not already accepted it. A fair way to analyze this objection is to consider a political dialogue between two people and look to see if the issues Brower presented indeed occur. Abe is someone who wants to make political arguments based on his conception of the moral good. Zeke is a proponent of public justification. He adheres to a comprehensive doctrine but does not seek to ground political arguments in the values of that doctrine. Abe claims that society should implement policy A because it is consistent with his comprehensive doctrine's view of the moral good. Zeke says that that conception of the good conflicts with his own, and as such he cannot reasonably accept Abe's justification. Zeke suggests that Abe appeal to shared aspects of their society's political culture. Abe responds by saying that it is disrespectful that he be asked to discard his comprehensive doctrine, which is very important to him, when making this important political argument. This is the point Brower gets at. My response is to ask what, then, is the alternative? It seems as though the only way out of this impasse for Abe is that Zeke accept his conception of the good and therefore his political argument. But this undermines any attempt at equal respect that Brower wants to make. If this is what comprises equal respect, then Abe's demand of Zeke is no less disrespectful than Zeke's demand. For people who value conceptions of the good and their importance as much as Brower does, it follows that they would then find it unfair to ask someone else to defer to their conception of the good, as that would be demanding that they disregard a personally constitutive value.

I respond that Brower's standard for equal respect is too high to be feasible, and that it will inevitably lead to the impasse mentioned above. Given the aforementioned fact of reasonable pluralism, which I take to be uncontroversial in a free society, this impasse will occur constantly. Public justification is compelling precisely because it is a mechanism for this plurality of doctrines to exist without anyone having to defer to another's conception of the moral good. A more proper standard for equal respect is to consider each comprehensive doctrine to be as valuable as any other. This standard of respect, together with the fact of reasonable pluralism, leads us to conclude that those who hold conceptions of the good to be vitally important to people would in fact find a reason to endorse public justification. It provides a mechanism for political cooperation while maintaining everyone's deeply important values and ensuring that the standard of equal respect is not violated. This conclusion in addition to the strengths of public justification I mentioned earlier provides a very strong basis for the acceptance of public justification by those who do not necessarily endorse Rawls' ideal of the reasonable.

**Objection 2: Self-Defeating**

The second objection to the theory of public justification I will address is presented by Steven Wall in his article, *Is Public Justification Self-Defeating?*. Wall argues that public justification is in need of justification, and is unable to satisfactorily meet its own demands to justify itself to those who it would apply to. In other words, Wall is arguing that public justification is not in itself sufficiently publicly justifiable to justify itself as a principle for determining the legitimacy of political authority. Wall begins his

argument by defining public justification in a way that is coherent and continuous with the way it has been defined here. He claims that among the relevant requirements for political authority to be publicly justifiable is what he calls the "acceptability requirement." Wall defines this by saying, "…the justification must be one that can be reasonably accepted (or not reasonably rejected) by those to whom it is addressed."[13] There is nothing problematic here. He goes on to discuss how we must make a distinction between a public justification and a correctness-based justification. For Wall, a correctness-based justification is one that demonstrates that a conclusion is correct, whereas public justification, something that has already been made clear, is distinct from this. This is significant for Wall because if proponents of public justification do not hold that political authority must be legitimized by both of the aforementioned justifications, then they are left to answer why public justification is even worth discussing. Wall continues by explaining that this can be resolved by claiming that public justification serves to mark the outer limits of our freedom[14], and as such leads to what Wall calls the "reconciling function" of public justification, which serves to show that each person has a good reason, by appealing to public justification, to accept political authority. This function demonstrates why a correctness-based justification is not inherently sufficient for legitimizing political authority.

Wall argues that despite the appeal of the reconciling feature of public justification, it is still not an inherently correct

[13] Steven Wall, "Is Public Justification Self-Defeating?" *American Philosophical Quarterly* 39 (2002): 385.
[14] Ibid., 387

theory of political legitimacy. This, Wall explains, "…is why it is reasonable to say that [public justification] stands in need of justification."[15] In other words, since public justification does not claim to be correct on moral grounds, it needs to be justified by other means. So, Wall asks, what sort of justification is required? The answer is that public justification must satisfy its own requirements, and for this reason the theory might be self-defeating. As Wall puts it, "If [public justification] were indeed a self-defeating principle, then it would fail on its own terms. This would give us a reason to reject it."[16] Wall proceeds by claiming that supporters of public justification must now either demonstrate that public justification does not apply to itself, or that it does in fact meet its own demands. Wall addresses the first claim and argues that it is untenable because it contradicts the very purpose of public justification. To claim that public justification does not need to meet its own demands would be to say that any given authority is publicly justifiable but then not offer a reason to accept the constraints of public justifiability. This does not get us anywhere.

Wall addresses the second claim against the self-defeat of public justification in two ways. In the first, Wall argues that any attempt to argue that public justification applies to itself because of values that permeate contemporary democratic societies would have to contend with the objection that the principle of equal respect is in fact not embedded in modern democratic societies. This results in there being at least some people in contemporary

[15] Ibid., 388
[16] Ibid., 387

society who would not reasonably accept the theory of public justification. Secondly, Wall discusses how even if there did exist some sort of background political value that all members of a society shared, people would disagree as to the particular nature of that value. In this case there would be so much disagreement about the shared value that the value would be too thin a concept to appeal to when giving public reasons.

Wall concludes his discussion of public justification by expressing doubt that there is any recourse for those who support public justification to prove that it in fact is not self-defeating. Additionally, he notes, political legitimacy might be a matter of degree, and that public justification still serves to legitimize political authority better than any alternatives. He concludes by claiming that given that even if these might be valid options for the proponent of public justification, they do not refute the overall claim that public justification is self-defeating.

To begin my response to this objection, I note that Wall seems to give a charitable presentation of the general principle of public justification. I will also concede here that since public justification is not a correctness-based justification, it does need to be justified further. I will here accept the claim that in order to avoid being self-defeating, public justification must either be said to not apply to itself, or must itself be publicly justifiable. I will refute this objection by showing that public justification is itself publicly justifiable. This is because, despite Wall's insistence to the contrary, there is indeed a commonly held political conception of justice in contemporary democratic societies, and it is that conception of justice that can be appealed to in order to justify the theory of public justification, as well as other political claims.

Although there are considerable disagreements when it comes to moral conceptions of justice, such as how to punish criminals and what moral codes people ought to abide by, when it comes to political discourse Americans still share fundamental intuition about what political justice is. By this I mean our political culture holds that taxation without representation, for example, is unfair and unjust in a political sense. Americans expect the will of the people and the spirit of the constitution to be enforced as matters of justice and would as a group reject a leader or proposal that violates the basic tenets of democracy and representation. We have an understanding of society as what Rawls calls "…a fair system of social cooperation over time from one generation to the next."[17] We have a sense of basic liberties as defined by our constitution. This commonly-held conception of justice, broadly defined, functions as a baseline that publicly justifiable arguments can appeal to. In other words, this shared conception of political justice in American political culture is a common ground that demonstrates that the principle of public justification can be applied to the United States. I am confident that such shared conceptions of justice exist in similarly democratic nations.

Here it is important again to note the distinction between agreement and a shared political conception of justice. People agree when for whatever reason they both find an argument or idea appealing. A common conception of political justice, however, goes beyond agreement because it is a fundamental aspect of the democratic political culture that members of a free society share. They share it not because it is in accord with their conceptions of

[17] Rawls, *Justice as Fairness*, 5.

the moral good, but because it is part of the political culture they belong to. People who disagree on political and moral matters may still appeal to this shared political value and offer compelling arguments (i.e. public reasons) to each other. It is from these public reasons that people may come to an agreement about political decisions or policies. Because of this common ground I, or anyone else, can offer arguments in political disputes that are reasonable for my opponent to accept on the basis of political justice.

Wall also argues that even were a shared political value to exist within a society, "…it does not follow that everyone has reason to accept the particular interpretation of this principle that is needed to ground [public justification]."[18] I contend that even given differing interpretations of this value, the fundamental core of the value, such as justice, would suffice for the purposes of public justification. Additionally, Rawls himself addresses this concern in his presentation of the idea of an overlapping consensus, wherein he echoes my claim. As he writes, "While…all citizens affirm the same political conception of justice, we do not assume they do so for all the same reasons…but this does not prevent the political conception from being a shared point of view from which they can resolve questions concerning the constitutional essentials."[19] As a result, public justification is in fact not self defeating because it can meet its own demands, and it can be shown that modern democratic societies do have sufficient shared political values for public reasons to be feasibly presented.

[18] Wall, "Is Public Justification Self-Defeating?" 390.
[19] Rawls, *Justice as Fairness*, 32.

**Conclusion**

The principle of public justification, that political authority is legitimized and political disputes resolved by both parties appealing to arguments that the other side can reasonably accept, is to me a powerful principle. Because it is not limited by conceptions of the moral good and because it can help us to escape the partisan rut we as Americans seem to be stuck in, public justification can act as a means to end long standing and seemingly irresolvable political disputes. In addition, the emphasis of the reasonable over the rational ensures we avoid the pitfalls of unrelenting rational self-interest, such as those presented in the prisoner's dilemma and the tragedy of the commons. Although objections are leveled against the theory, they are not sufficiently strong to dissuade us from accepting public justification and its advantages in terms of fairness, respect, and pragmatism. In the end, public justification remains the most reasonable and compelling method for adequately resolving political disputes and legitimizing political authority. I genuinely believe that this principle is the best way to overcome the obstacles of political oppression and divisiveness, in spite of people's desires to adhere only to their conceptions of the good. Were just Americans to accept this principle, the contemporary political climate would improve tenfold, and much more genuine progress could be made.

## Bibliography

Brower, Bruce W. "The Limits of Public Reason." *The Journal of Philosophy* 91 (1994): 5-26.

Cohen, Joshua. "Politics, Power, and Public Reason" Paper presented at the UCLA Legal Theory Workshop, Los Angeles, California, April 17, 2008. 1-70.

Rawls, John. Justice as Fairness: A Restatement. Cambridge: Harvard University Press, 2001.

Wall, Steven. "Is Public Justification Self-Defeating?" *American Philosophical Quarterly* 39 (2002): 385-394.

# Having Children: Reproductive Ethics in the Face of Overpopulation

*Kianna Goodwin*

**Abstract** Overpopulation is a serious threat to future persons' quality of life. One that I believe can only be addressed by adopting reproductive values that inspire justice for future generations. In this paper I discuss theorists whose views I argue support limiting the right to procreate. I believe enforcing reproductive responsibility is necessary to curb the problem of overpopulation and therefore maintain a standard quality of life for future generations.

It's common to think of having kids as a personal opportunity to experience a unique happiness and our ability to make choices about procreating as a key expression of our identity and personal autonomy. These factors make us feel that the decision to have kids is a deeply individual choice and more importantly that there exists no ethical justification which could diminish this fundamental right.

Our world population has doubled in the last 40 years, which means by 2050 we could potentially have 12 billion people in the world. Overpopulation occurs when the rate of birth exceeds the rate of death. People today have the capacity to live longer lives than ever before, yet lack of access to clean water alone prematurely kills millions across the globe every year. Despite the countless global struggles that lead to premature death we are still

reproducing at a rate that surpasses our rate of death. If we were to fix all the world's problems that lead to unnecessary death we must still contend with the fact that we are subsisting on a planet with a limited ability to provide space, supply food and produce energy. Even if it were possible to overcome the injustices of inequality by radically altering the distribution of resources or achieve technological advancements that are more sustainable there will still come a point at which none of these achievements will be enough to support the sheer number of people that will populate the earth. Overpopulation is a subject we do not breach publicly for fear of appearing absurd or anti-freedom; however I feel it is an issue of major ethical concern and one that needs to be addressed in order to negate this impending situation.

Discussing overpopulation is taboo because it threatens to breach the fortified value we have placed on reproductive autonomy. But I feel that the possibility of bringing people into a world headed for self-destruction is a greater ethical concern than avoiding taboo. Overpopulation is something that threatens the wellbeing of future generations and taking steps to alter this trajectory necessarily demands sacrifices from present generations, namely sacrificing complete reproductive freedom. I believe present people remain unconvinced of this necessity because their current reproductive values do not foster/support concern for future generations. So in order to properly address this issue of overpopulation, which greatly threatens future generations we need both a change in reproductive policy as well as a change in social values. Success is dependent on the implementation of both to make a difference because it would be impossible to enforce such infringing policies if they didn't reflect actual social values. In this

paper I will discuss some philosophical reasons as to how we might justifiably limit the right to procreate in the face of overpopulation. I am concerned specifically with the ethics involved and how we are able to reconcile concern for future generations vs. our own desires for procreative liberty. First, I will establish that a state of overpopulation is in fact undesirable and a situation to be avoided because it has negative consequences for the societies where it occurs. Secondly, the defining characteristic of overpopulation is that it's a problem which worsens over time, so next I will argue for why present generations should feel a connection to future generations who will inherent a worse problem than the generation before. Namely, I argue that the connection between generations is representative of how we understand our procreative duties and this in turn plays out in our reproductive ethic and how we relate to future generations. I will devote a section of the paper to deconstructing some of the reproductive ethics and customs we have now and examining how these views impact where our values lie regarding future generations. In the next section I will look at alternative ethics which carry different perspectives on procreation, therefore creating a different value system that I believe naturally prioritizes future generations. Finally I hope to make an appealing case for limiting procreative freedom in a way that reflects our values regarding having children, both present and future and provides them with a better quality future.

**How Having Too Many People Negatively Affects Everyone's Quality Of Life**

In his work "Tragedy of The Commons" Garrett Hardin

argues that there must be a restriction placed on limitless population growth because of existing persons' inability "to bear the full burden of the children they have." He insists that overpopulation is inherently a no win situation and the biggest mistake we make when thinking about overpopulation is our inability to factor in institutional sacrifice as a reputable solution. Population grows geometrically, i.e. exponentially and this means that eventually the world's resources are guaranteed to diminish because it is not possible to support an infinite population on a terrestrial landscape that is finite. Hardin uses the example of a "herder", who sees the common pasture as a limitless means to expand his herd of cows because they can graze freely and in as many numbers as he is capable of procuring. The herder does not consider this use of the pasture to negatively affect him on the individual-level, especially since he stands to gain so much personally from having a large and ever expanding herd. The "tragedy" is that everybody else has come to the same conclusion and so the pasture is not able to maintain itself under the strain of so many cows, let alone actually nourish them all. This is a simple analogy for the effect of large populations of self-interested people living in a limited world. Pollution also originates from the same thinking, except that instead of taking something indiscriminately from the commons something is indiscriminately put into the commons, which leads to the destruction of the original fruitfulness, so that we are effectively "fouling our own nest." [1] Having a limitless population, (again, actually impossible) or at

[1] Hardin, Garret, "The Tragedy of the Commons," in *Ethics and Population*, ed. Michael D. Bayles, (Cambridge, Mass: Schenkman Pub. Co., 1976), 9.

least a population double the current size would require that we learn to limit consumption of resources so as not to exceed the bare minimum needed to survive. That means if a man must eat a minimum of 1600 calories a day to survive/manage all his daily obligations then all calories consumed beyond that amount would be considered possibly beneficial but not necessary and therefore no longer part of his diet. Consuming more than this would be taking something beyond his share and therefore impending on someone else's ability to live. I do not think we can conceive of living on a planet with 20 billion people where our lives are so dependent on just servings for total survival. Hardin uses this example to emphasize that the more people we have on the planet the more we will be forced to downgrade from our expected quality of life, if we expect to continue without destroying our own living environment.

But this brings up questions like: Why care what happens to the planet beyond my lifespan? Or about the lives of people who don't already exist now? If having 15 babies and spoiling them to their heart's content suits me and is within my power to bring about then why not do it? I believe these ultimately disastrous sentiments reflect the current vision of reproductive liberty and can only be addressed by first understanding and then assuming other interpretations of reproductive rights.

**Reconsidering Commonly Accepted Values Regarding Procreation**

Procreation is normally understood as an autonomous decision in two fundamental and problematic ways: as an autonomous bodily decision and as something related to an

individual's self-conception. Understanding procreation as simply an expression of a one's bodily autonomy and an extension of one's ownership over their physical self is inherently problematic. This view focuses on the right to experience one's body in anyway one pleases, including pregnancy; and furthermore that being pregnant is a phenomenon like any other biological process. This makes it seem as if the birth of a child is an extension of one's physicality in the same way that growing out one's hair is, i.e. as if the unborn child were simply a by-product of one's sole individual organs. But becoming pregnant and maintaining the intention to carry the child to term so that it can eventually flourish as its own independent organism is something that's different in kind, not degree, from any other bodily function. Yes, any child who is born was at some point part of its mother's body. But after its birth it no longer functions as an extension of her body and instead lives as its own being; again, showing that the mother's body does not continue to wholly account for this new being's continued existence. In this case pregnancy acts as the original link in the causal chain that will become someone's entire life. While the pregnancy should necessarily be identified as this causal link it also means that the biological mother cannot claim her decisions affect only her and her own body when pregnancy leading to birth necessarily means that her decisions will come to affect at least two persons.

Here I think it is important to clarify a distinction made by Ruth F. Chadwick between begetting, bearing and rearing children because all of these are separate concepts silently at play when we talk about "having children". The fact that we indiscriminately employ the vague term "having children" inevitably leads to

misunderstandings. For example begetting is often a major part of how men conceptualize their procreative role; but if a man over emphasizes his role as begetter over and above his other duties because he has not internalized the two other roles associated with fatherhood then he might behave indifferently and spawn many illegitimate children. The greater outcome of this self-ascribed definition of father is that it can leave many children without the proper care they deserve.

What is important to grasp here is that each step in the procreative process is meaningful and necessary for creating new life but also potentially isolated from the other aspects involved. Secondly, a procreator may feel an emotional connection with any of the steps including: conception, gestation and labor, and the care/ raising of the child. It is also possible to connect with none of them, which is problematic for cultivating a society which demands accountability for their children's quality of life. In the same vein I realize not everyone is capable of every aspect of the procreative process; while some cannot conceive or carry a child others may not be able to rear one because of some critical personal deficiency/hardship. The problem remains that "having children" is an ambiguous undertaking at best. It might seem like this lack of clarity "issue" can be solved simply by separating out the rights that should pertain to each role (begetting, bearing or rearing) but on the whole this isn't too far from the system we have now. Currently, everyone has a right to procreate and to bear children at their own convenience. The same goes for rearing their children until reasons surface that expose them as unfit to care for a child and their right to raise their children can be taken away. But someone's right to conceive and bear children cannot be

terminated. We do not feel it is within anyone's moral capacity to force sterilization on someone who has demonstrated a severe inability to raise their own children in a loving, stable home. Similarly, but less problematic is that no one can be forced to raise a child they have conceived. These rights are all negative rights that allow us to relinquish our responsibility in some regard to our offspring and while we do have laws in place that require us not to brutalize, starve or sell our children I can't say that we have any that prioritize our children's right to a quality life over our own individual freedom.

Hopefully one can see that current procreative liberty operates as a very complex and far-reaching right. This is because the societal attitude implies that it involves anything one finds meaningful and fulfilling for his or her own private life. The problem is that what's considered meaningful and or personally beneficial to someone about reproducing is subjective and might include: experiencing the miracle process of labor, passing on one's genes by donating sperm or the choice to give up custody and terminate all parental rights. All of these examples involve extremely different intentions but nonetheless result in the creation of a new life. I think it's contradictory to be concerned with the wellbeing of existing children yet sanction all of the varied intentions that create new children who may end up suffering from difficult situations caused by those intentions. There are some possible intentions held by the begetters of children that directly lead to a lower quality of life for their child as they are assisted by attitudes of indifference, self-centeredness, or shortsightedness. A set of values that demands total procreative freedom as well as welfare for children is creating a hierarchy of values, which places

the interests of parents first and then scrambles to address the problems directly resulting from that hierarchy. I believe it's sound to question the intentions behind someone's involvement in any aspect of the reproductive process and more importantly to accept that some intentions are not justified when the impact or result of that decision carries such huge implications for persons other/beyond oneself. My point is that just because it is possible to separate the roles involved in procreating doesn't mean we should limit the responsibility regarding the care of children by believing that some roles bear no weight in the welfare of children.

**Why Care About People Who Do Not Exist?**

Philosopher Derek Parfit is also very concerned with doing the best for our children yet runs into a wall he calls " the non-identity problem" when considering choices that may affect their future. In a classic thought experiment we consider a woman who contracts an illness while pregnant, one that would cause a considerable deformity in the child resulting from the pregnancy. However, if the woman waits just three months to have a child the illness will be gone completely and her child will be perfectly normal. According to Parfit one's identity is necessarily rooted in the unique circumstances of their birth, three months later the circumstances would be entirely different the resulting person would be a product of these different circumstances and therefore a different person. Although at first it seems like the woman should wait to have the baby because it would be better for her child on closer inspection we realize that she is actually choosing between two different people and on this view we can't say that it would be better for the first child if the non-afflicted second child were born

instead. This realization leaves us in a bind where it would be better for no one either way as potential persons i.e. people who are not born have no concrete identity. However, Parfit does not want his view of identity to create an apathetic view of the future, and I feel that as long as we know that future people will exist, and they will, then we have a responsibility to them not to cause any harm, "Remoteness in time has, in itself, no more significance than remoteness in space. Suppose I shoot an arrow into a distant wood, where it wounds some person. If I should have known that there might be someone in this wood, I am guilty of gross negligence. Because this person is far away, I cannot identify the person who I harm but his is no excuse. Nor is it any excuse that this person is far away. We should make the same claims about the effects of people who are temporally remote." [2]

Unfortunately Parfit runs into more trouble when he tries to reconcile the non-identity problem with utilitarian values regarding future persons. He calls this new problem the "repugnant conclusion" and it stems from the idea that if we want to maximize happiness then if we have a population whose happiness is on average what we consider optimal then by adding a few extra people whose happiness is slightly below this the total amount of happiness increases from result from this addition. This ends up being a slippery slope where by adding more and more people we end up with an overlarge population whose lives are barely worth living. I believe these dilemmas to be counterintuitive in that they both assume what is important is that "happy people" be born, and

[2] Parfit, Derek. 1984. *Reasons and Persons*. (Oxford [Oxfordshire]: Clarendon Press.), 375.

seems to construct people as merely "happiness machines". "Just as a boiler is required to utilize the potential energy of coal in the production of steam, so sentient beings are required to convert the potentiality of happiness, resident in a given land area, into actual happiness. And just as the engineer will choose boilers with the maximum efficiency at converting steam into energy, Justice (utilitarianism) will choose sentient beings who have the maximum efficiency at converting resources into happiness." [3] It's not good that people exist because they're happy but that happiness is good for people who exist. What the repugnant conclusion assumes and the theorists that I reference deny is that we have an absolute duty to bring happy people into existence.

**Alternative Viewpoints That Better Support Future Generations**

When it comes to procreating *it is* possible to have a kid whom you love dearly, that you can provide for, who never experiences random terrible tragedy, who you have a great relationship with, who's healthy, that loves their life and is a good person. It might be the case that all of this characterizes your parenting experience, or it might not be… but there is no guarantee either way. David Benatar[4] is keenly aware of this and says that life inherently holds suffering as it necessarily involves enduring bodily decay and confronting mortality; there is however, no one who is possibly harmed by non-existence. He also believes that

[3] Bayles, Michael D. 1980. *Morality and Population Policy*. (University: University of Alabama Press.), 390.
[4] Benatar, David. 2000. "The Wrong of Wrongful Life". *American Philosophical Quarterly.* 37 (2): 175-183.

there are lives so miserable that by our own standards we could consider them not worth living. We therefore have a responsibility to avoid this cruelty and to not bring about such lives. So even on utilitarian grounds, more is not always better. But because the nature of existence is at best neutral (containing both happiness and suffering) we have no duty to bring "happy" people into existence either. The "neutrality" of life does not imply that great happiness and minimal suffering and great suffering and minimal happiness are ultimately equal in their value but that the potential for either scenario to occur or the scales to tip in either direction remains equally possible. Even if all precautions are taken to ensure a happy life for someone their life will necessarily contend with the presence of unhappy scenarios, which means there is no such thing as a non-tempered, unaffected and therefore totally happy life. We cannot say that existence holds the potential for total happiness and is therefore preferable to non-existence because we cannot possibly produce a sliding scale that shows the point where life is total happiness. Therefore you cannot bring into existence nor account for totally happy people in the world. However, you may be able to discern circumstances where someone's life is total suffering and therefore not worth living. The best that we could hope for is that they are contentment with the proportions of suffering and happiness in their life. Not bringing such people into existence causes them zero harm, not a proportional amount of harm, and so this option is always justified. The obvious consequence of adopting this view is that procreation is rendered seemingly… unnecessary.

Yet Benatar's view is that we may still choose to procreate if we wish so long as we're bringing into existence people whose

lives would be worth living. But how do we define a life worth living? This is where Benatar gets a lot of flack since it's unclear what decides whose life is worth living and whose is not. I think this is a misinterpretation of Benatar's intention in that it fails to differentiate ceasing to exist from having never existed. Benatar recognizes that people may have lives that started out as barely worth living but became lives of high quality and conversely that there are lives which started out worth living but are now barely worth continuing. Whatever the circumstance people's lives are necessarily linked to the individual suffering they've experienced, and asking whether they wish they'd never been born is completely futile. Despite whatever handicap they are faced with Benatar says people often view their lives through a distorted lens of attachment regardless of what they would say of their own circumstances objectively. What we are really talking about is not terminating existing beings but refraining from causing lives to begin that are not worth living; it's preventative. In effect, by limiting the amount of actual people who are harmed.

Shiffrin further uses the concept of harm to help us see how exactly the role of parent is to be understood. Like myself, she makes it clear that what she is *not* trying to do is belittle the difficulty involved in properly carrying out parental duties, but to draw attention to the moral implications involved in *creating* a life. She is therefore talking about a situation involving strict liability because of the inherent one-sidedness of this relationship where *the parent and only the parent* chose the life of the child. Furthermore this child will inevitably come to suffer harm in their life, the existence of which is a product of the parent's desire to have a child. She calls this "wrongful life". Shiffrin defines harm as it "primarily involves the imposition of

conditions from which the person undergoing them is reasonably alienated or which are strongly at odds with the conditions she would rationally will;" furthermore " harmed states may be ones that preclude her from removing herself from or averting such conditions." [5] What is important to note is that harm is firstly something that the person being harmed *did not will*. Harm is not just loss or pain but anything which, "exerts an insistent intrusive and unpleasant presence on one's consciousness that one must just undergo and endure." [6] This to me is a perfect description of the anxiety that is an inherent part of survival

The analogy often used involves a rescue scenario in which it is necessary to break the arm of an individual in order to get them free of a car wreck (where the danger could potentially escalate) and save their life. By choosing to harm this person in the act of breaking their arm you have also carried out the action necessary to save them from harms greater than a broken arm. The relevance is that it's necessary for people to suffer some harm in existence in order to enjoy the great benefit of life. Shiffrin openly denies that this is an accurate parallel. She says a "pure" benefit is not solely the removal of harm but the ability of the benefit to improve the overall quality of life for the recipient. The rescue case is not an example of a pure benefit because it addresses only the removal of a single greater harm, (greater injury or death for the victim in the accident), but does not necessarily disallow the existence of yet another harm to this person later in life. In real life procreation does act as a benefit which avoids obstructs any greater harm. The rescue scenario exemplifies Shiffrin's insistence that this analogy "illegitimately trades upon a common equivocation of

[5] Shiffrin, Seana Valentine. 1999. "Wrongful Life, Procreative Responsibility, and the Significance of Harm". *Legal Theory.* 5 (2):750
[6] Ibid., 750

"benefit." [7] In other words that we speak as though removing someone from harm is what benefits that person. In reality it does not follow that it is the act *of doing the saving* which is the moral justification for inflicting harm but the greater positive (beneficial) outcome that is the result of the saving. Conversely the beneficial act of creation doesn't allow justification for harm because the greater outcome of procreation is not that a greater harm is averted. It is not appropriate for us to think it acceptable to harm someone just to gain a benefit. Such an action only becomes morally innocent when we do it to remove some greater harm. We are certainly not justified in inflicting a minor harm for the prospect of a greater benefit.

There is another often-cited example used in attempts to emphasize the inherent good of life by drawing a connection between life and benefits which I believe is relevent. In this scenario the hypothetical character called "Wealthy" injures another character, "Unlucky" in an attempt to bestow benefits which would improve the overall quality of Unlucky's circumstances. Wealthy is a philanthropist of sorts who decides to charter a plane so that he may distribute his solid gold bricks indiscriminately by randomly throwing them overboard. One of these bricks falls on Unlucky and the impact injures him as one expects a hit from a gold brick would. Though Unlucky is caused significant pain from his injuries he will definitely live and the gold brick is his to keep. Once again the given example presupposes many things, including as already stated, the fact that it is morally justified to harm someone simply for the sake of what is *assumed* as a benefit at the time without the "beneficiary's"

[7] Ibid., 751

consent. Now, what if Wealthy included an additional 1.5 million dollars meant as anticipatory compensation for the injury caused by dropping the brick? Shiffrin and myself believe this is a false solution; if the compensation is "built in" to the harm then it seems as if Wealthy is preemptively pardoning himself from any culpability as well as disregarding his subsequent duty to seriously address any and all harm done. In order to legitimately act in compensation for a harm then one must seriously address the harm itself as it stands alone. This means as separate from the delivery or execution of the harm i.e. certainly not exploiting any potential for benefit in order to justify doing the harm itself. I think the concept of wrongful life is inherently different from the rescue or financial scenarios used in thought experiments for them to be compared. In the case of procreation not only are we committing the much more serious act of creating brand new life but in this case we neither save nor prevent anyone from a greater harm.

The key to understanding the wrongful life concept is being able to come to terms with naming all the things that are scary and difficult about having and raising children. No one wants his or her child to suffer, so then, why is it so difficult to understand that they *will* suffer? And how is it not in the nature of a parent to naturally assume responsibility for all that their child feels, endures, achieves, etc? This theory is really not much more than a reflection of these basic inclinations that are intrinsic to good parenting. I believe this appeals to the greatest of all parental instinct and that is to shield one's child from harm. Opponents to wrongful life might again say that any possible horror experienced by a child is not cause enough for a parent to call their child's life wrongful. I think Shiffrin would disagree and say that a parent's

instinct to protect is so severe that the failure to do so could potentially create such guilt that they'd prefer their child to never have been born. Not because they do not value their child's life but because they acknowledge the unfairness of a child suffering who did not ask to be brought into this world.

Another critique of wrongful life questions the point where a parent should cease to be liable for all harm experienced by their child. The concern for how far into lives of future people we are responsible for is something that concerns Parfit as well. Personally, I think that the point at which a parent ceases to be liable is relative to the initial harm incurred by the child in their youth. Again following Shiffrin and as well as intuition I think the concern is really whether the parent took proper steps prior to conception as well as during the child's early years that showed consideration for their future. Ideally, the child will become completely responsible for itself so far as they were provided the tools to do so by their parents. If the point at which their life becomes unmanageable can be traced back to an original and significant harm done by the parent then that parent should be held responsible contributing to the current situation. But again appealing to intuition it should follow that the older the child gets the murkier that trace line should be due to the growing agency (autonomy!) of the child. And this is true for Parfit as well; it would be wrong to deny the initial connection we do have to our children's future because we are not able to see forever into the results. The better it is seen to that children are given what they need to make their own decisions and inform their own actions the less it can be said that their lives are limited by the decisions of their parents. Similarly we must leave behind a quality of life that

reflects our own standards for our children or be responsible for negative quality of life they experience. Giving life is currently seen as a gift, something for which we should be never-endingly grateful for, something that is beyond reproach, we should not demand more of the giver. But giving life is not something that pardons you from your responsibilities, in fact quite the opposite, having children only extends your responsibility indefinitely.

**What Different Values Means Practically**

When we begin to grasp the kinds of values regarding parenting and procreation perpetuated by Benatar and Shiffrin I believe we are better able to accept a difficult course of action like limiting population. We see limiting our procreative liberty as less about our own limited freedom and more about doing what's right by future persons by providing them a certain quality of life. It's easy for us to accept that we have a moral duty not to force undesirable situations on others. We now have the ability to include future persons based on a strong understanding that we actually dictate who these people will be and therefore have just as much of a relationship with them.

According to population scholar Michael Bayles, the greater the need for population control the more likely there will be a greater need for limits on freedom as well. This is referring to problems which are dire (immediate) and require solutions beyond volunteerism or family planning. For Bayle guilt plays a major role in our society; it influences how we feel about our own actions; however it does not necessarily change them. The desire not to harm future generations may be instilled in present generations but it does not curb the tragedy of the commons. That is why we will

eventually need policies that allow us to execute these views. He insists that because no specific values regarding quality of life have absolute priority (subjective) it's necessary to evaluate policies based on their ability to successfully accomplish objectives for present and future persons. This means that a policy is only justifiable if it actually realizes the desired effects. Bayles also emphasizes that some freedoms are greater or more important than others and that this should also dictate how we are to address certain population concerns. He advocates a pragmatic use of our perceived spectrum of freedoms. For example, it is less of an infringement on peoples' freedom to be able to have up to two children rather than no children. The main difficulty of implementing such policies, whether they be positive incentives, negative incentives or compulsory is to insure a level of equality regarding the actual effects. Neither Bayles nor myself thinks that it is ethical for people of lesser means to bear the greater burden of limiting population growth. Again what this means is a pragmatic approach and an emphasis on equality. I think that it's also important to emphasize that poverty does not necessarily make for life barely worth living. There are other values in regards to quality of life to be prioritized which are more universal like, mental stability, sobriety etc.

Hardin states that humans intuitively feel guilt for however they've failed ethically. But regardless of whether guilt is a naturally occurring response, it's also useless in bringing about an optimal desired result. Along this line I believe any person is capable of feeling a deep love for their child and still failing them. Hardin proposes what he calls "Mutual coercion mutually agreed

upon." [8] He feels that coercion regularly practiced simply means bringing about the desired result that everyone wants but doesn't want to contribute to themselves, like taxes, and that the same can be said of limiting the resources/rights to reproduce infinitely. Responsibility Hardin says, is a product of social arrangement and does not occur on its own. We cannot measure, control, or affect how much a procreator loves their progeny but what we can do is take steps to ensure a basic quality of life for them so that they are able to pursue lives worth living.

**Conclusion**

By adopting reproductive ethics that inspire justice for future generations I believe the limits on procreative freedom become less burdensome for present generations. Whether institutionally enforced social responsibility is successful relies on our own personal relationship with the values we are upholding. Overpopulation is a threat to future persons' quality of life, which means essentially that it's a threat to our children and our children's children as well as to ourselves.

[8] Hardin, *The Tragedy of the Commons*, 7.

## Bibliography

Bayles, Michael D. 1980. *Morality and Population Policy*. University: University of Alabama Press.

Benatar, David. 2000. "The Wrong of Wrongful Life". *American Philosophical Quarterly.* 37 (2): 175-183.

Hardin, Garret, “The Tragedy of the Commons,” in *Ethics and Population*, ed. Michael D. Bayles, Cambridge, Mass: Schenkman Pub. Co., 1976

Parfit, Derek. 1984. *Reasons and persons*. Oxford [Oxfordshire]: Clarendon Press.

Shiffrin, Seana Valentine. 1999. "Wrongful Life, Procreative Responsibility, and the Significance of Harm". *Legal Theory.* 5 (2): 117-148.

# The Irony of Ironism: A Critique of Rorty's Postmetaphysical Utopia

*Jeffrey Rivera*

**Abstract** In Richard Rorty's work *Contingency Irony and Solidarity*, Rorty attempts to elucidate a mechanism for dealing with the public dissent likely to arise from a group of individuals he terms "ironists". This mechanism, a strong public/private distinction, he hopes will allow for a self proliferating, ever progressing liberal utopia. This paper will reject this distinction as internally incoherent under its own terms, and will assert that even if Rorty's distinction is successful, it ultimately attempts to proliferate the type of individual we would like to avoid.

In his book, *Contingency, Irony and Solidarity*, Richard Rorty urges us to rethink our conception of what a liberal society should look like, and which values it should hold and promote. Rorty claims that our current vision of a liberal society is one that is governed by the idea that cruelty, the promotion of suffering, is the worst thing that we as liberals do. In addition to this, Rorty appeals to the idea that a special kind of suffering, humiliation, for a liberal, is an especially bad form of cruelty. Rorty is aware however, that the type of individual likely to cause civil unrest and humiliation, the unorthodox thinker, is also a potential catalyst for political, cultural, scientific and philosophical progress. He is at once the liberal hero, an enigmatic poet who makes the world his own, but he must also be the villain: the egoist par excellence.

Recognizing the danger and importance of such individuals, Rorty creates a description of liberalism which might give us the best of both worlds; private self creation, as well as public unity and social cohesion. In this paper, I will argue that the mechanism that Rorty asserts to bridge this gap, the affirmation of a strong public/private distinction, will not feasibly do the work which he requires. Furthermore, I seek to show that even if this distinction holds up, the ironic liberal should not be the type of individual we would like to promote in a utopian society.

What is an ironist, and why should a liberal society protect the autonomy of these individuals? In order to answer this question, Rorty appeals to historical change as being a product of the evolution of language. Rorty describes the ironist as being indebted to a specific historical view, one that will see the strong poet, the thinker who re-describes and creates something new, as instrumental to intellectual progress.

The ironist understands that he is born into a specific historical juncture. This notion can be equated more or less to the existential notion of facticity found in thinkers such as Heidegger and Kierkegaard. This is the idea that, with the inception of one's life, comes a set of specific conditions which relate to and characterize that being. For Rorty, the most relevant aspects of one's facticity seem to be the subject's relation to history, and specifically historical discourse, and the language games he is prone to play given his position within this canon. This is a specific contingency which all beings must depart from in order to become self-creators.

The key difference between the ironist and the liberal is brought to light with regards to this realization. Whereas the liberal

is content to play the current language games and realize his self-creation through these paradigms, the ironist views his being born into a specific paradigm as constraining. He feels that if he is to be the strong poet, one who fears his self-creation is merely a replica of a past self, he must create a very strong sense of his own identity. This cannot be done within the current paradigm because it places importance on specific modes of thinking. Rorty shows us this in his analysis of the character of specific time periods. For instance, if we look at thinkers whom we perceive as being particularly influential, we see that they do not merely find or relay information in light of the current views on an issue, they seek to re-describe the phenomenon under a new sort of view. For instance, Einstein's theory of relativity does not simply work out some inadequacies of Newton's theory, it fundamentally re-describes all relevant phenomena in a completely different light. It somehow makes us see things in a different way and therefore makes things new. This is the sort of re-description that the ironist sees as important to his self creation.

We might use this sort of example to point out another important feature of Rorty's theory, namely that there are specific historical conditions of possibility for the adoption of new language games. The first of these is that new descriptions of the world are brought about in light of past inconsistencies or uselessness of older language games. This might be understood in a similar fashion as scientific theory choice. As discourse progresses within a subject (slowly, as a product of small contingencies), problematics arise within it. For instance, Newton's theory cannot properly describe phenomena when approaching the speed of light. These inconsistencies are typically dealt with by the

introduction of ad hoc solutions. Novel re-descriptions remove these inconsistencies by creating a new view of the interactions of the phenomena at hand.

It is important to note here that novel descriptions initially have no place within ~~a~~ the extant language game, because they are not truth candidates within that language game. As they are posited, re-descriptions are metaphorical, but have the potential to become truth candidates as they are adopted by language users – as those language users begin to interpret the world in that particular fashion. For Rorty, it is crucial to realize that this process of language adoption is not one of the language user's rational choice, but a *process*. Since the individual statements of novel language games are not truth candidates, the old and new languages are adopted not in light of a comparison between the novel and the previous descriptions of phenomenon, but by the slow shifting of the way particular agents see themselves and describe their world. Rorty recognizes that for most (for the non-ironists), the creation of an idiosyncratic language is non-essential to their notion of self-creation and as such, they are not want to change their manner of speaking. To put it another way, non-ironists don't necessarily see themselves as, but inherently are, people who value a form of historical continuity.

This valuing of continuity is also implicit in the liberal's relationship to what he calls his "final vocabulary". A user's final vocabulary is constituted by those terms which he uses to relate himself, his desires, his goals, and values, to others. A user's final vocabulary is "final in the sense that if doubt is cast on the worth of these words, their user has no noncircular argumentative

recourse."[1] Again, the liberal has no problem using typical language games to elucidate his ultimate conception of self. He sees the evolution of his final vocabulary as linear. The ironist that Rorty describes, on the other hand, sees particular vocabularies as constraining to his notion of self creation, as the ironist is someone who cannot simply take the paradigm which was factically imposed upon him and proceed from there in self creation. He must appropriate and re-describe the past in order to make it his own and become a completely idiosyncratic self creation. The liberal is content to move forward, while the ironist wishes to create an entire new line. He must idiomatically create the taste by which he will be judged. If the ironist creates his own vocabulary, he has thereby created his own novel system for truth candidacy and therefore can see himself as authentically created.

We might remark that this description of the ironist sounds very much like the picture of the "authentic being" described by Heidegger or Satre, or Nietzche's "ubermensch". Presumably, many of us would find the promotion of *this* type of self creator as questionable, as they have been traditionally linked to anti-liberal, (and sometime fascist) ideology. However, Rorty offers a different take as to why we should wish to steer clear of the ironist type.

Rorty describes liberals as those who think that the promotion of suffering as the worst thing liberals due. Further, Rorty describes a special sort of suffering that should be avoided within liberal societies: humiliation. Ostensibly, it is a special type of suffering for liberals because we, as liberals, are concerned with

[1] Richard Rorty, *Contingency, Irony, and Solidarity*. (Cambridge: Cambridge UP, 1989), 73.

self-creation. Following his linguistic self-narrative view, Rorty's formulation of humiliation is done in terms of linguistic communication. Humiliation is a special sort of suffering in the sense that it is a forced shift in one's final vocabulary and therefore one's self creation becomes compromised. Because the ironist is ever anxious about the terms in which he describes himself as a result of his rejection of objective language choice, and therefore truth, Rorty asserts that the ironist is the sort of human being *by nature* who has no respect for the humiliation of others' vocabularies. He is therefore the villain of the liberal society, while at the same time being the catalyst for change and progress.

What does it mean to humiliate someone linguistically, and what are the conditions of possibility for this form of cruelty? As we have discussed, Rorty believes that shifts in language are products of many small re-descriptions which lead to a shift in one's final vocabulary. These shifts in vocabularies slowly lead to paradigmatic language changes. Slow language changes are normal and covetable, as they are based on the decisions of the agent (or groups of agents) and help to inform his self narrative. What occurs when we liberals are humiliated, Rorty asserts, is that our final vocabulary has been forced to shift, resulting in a major challenge to one's identity. It is important to note that the ironist is immune from this sort of humiliation because they are aware that "the terms in which they describe themselves are subject to change" and they are "always aware of the contingency and fragility of their final vocabularies, and thus of their selves."[2]

[2] Ibid., 74.

Fortunately, Rorty seeks not to promote the ironist, but the *liberal* ironist. The liberal ironist, is an individual who holds fast to his ironic values privately, but shows no sign of ironic ideology publically. Rorty asserts that the liberal ironist can maintain this view because he understands that his, and everyone's, language is ultimately nothing more than a view informed by contingencies which an ironist must continually overcome. To put this point in another fashion, the ironist understands that truth is merely a property of a specific language game. He sees that the language games we choose to play are based upon contingencies about the way the world is, and thus how we see the world. He also recognizes that these games shift over time; they are savored or spit out by different cultures, political factions and intellectual movements. In short they see language and therefore truth evolving, and therefore reject the ability to make objective decisions about the value of playing any one language game over another. Rorty feels that this relativistic position allows a strong enough reason for the ironist to affirm a public liberal standpoint, while also embracing a commitment to hiding his ironism in the shadows of his or her private life.

It is this mechanism that will allow for his ever-evolving flourishing post-metaphysical utopia. Rorty claims that through the linguistic evolution the ironist offers, paired with a sense of solidarity afforded by his liberal values, we can create a stable liberal society full of ironists. Since each of these ironists seeks to break with the status quo, Rorty claims we will have more and more re-descriptions, and therefore more fuel for future ironists' self creation.

I would like to offer three criticisms here. First, such a liberal society has a low potential to be endorsed by the ironist, even if he firmly holds that the public/private distinction should be enforced. Secondly, the distinction causes the ironist to be an almost pitiable character and therefore we should not promote a society where "irony is universal." Lastly, the ironist by becoming a liberal destructs too much of what it means to be an ironist, to make the label "ironic liberal" plausible.

First, it seems to me that to force the ironist into affirming a particular political conception is antithetical to the idea of the ironist. Just because the ironist is ostensibly immune to humiliation by means of language (as his final vocabulary is ever in flux), does not mean that his language and desires are not limited by the holding particular political ideologies. In holding particular political doctrines, we are acting antithetical to the idea that the ironist is a human being who is ever in flux about his self description. By positing a reason to hold liberalist ideals, he is further constraining himself. He is more likely to be a mere replica, and therefore he might get the sense that his public affirmation of liberalism constrains his self narrative. Indeed the ironist does not merely mentally gratify his own idiosyncratic language, but seeks to use it to describe himself and his desires. If he finds his desires are contrary to the desires of liberalism, then he is at a loss to express his desires.

In examining the justification of his liberal ideals, it seems to me that these ideals do not stem from the ironist's ironic values. If the justification for the agent's irony comes from the realization of the contingency of his final vocabulary, then it seems to me that this view cannot inform a liberal viewpoint. Since the ironist thinks

it is fatuous to regard his or her final vocabulary as being stable, why should the ironist respect the contingent values of others? It almost seems like the liberal ironist takes more seriously the ideals and values of others, while rejecting and suppressing his own. If this is the case, he has no reason to be a liberal, since his self-creation is merely a secondary concern.

Furthermore, it seems like the ironist is precisely the type of person who would reject the type of justification which Rorty believes might inform the ironist's decision. The justification given for an ironic affirmation of liberal ideals is necessarily an inter-subjective one. If I, as an ironist, understand that each individual's views are unimportant, then I may see others as like myself and may seek to promote the welfare of others self creation. This to me seems to be antithetical to the sort of view which the ironist wishes to pursue in the sense that it seems close to positing an objective truth about the intrinsic nature of the self. It is the truth that each person's self narrative is important to his or her self, but recognizes that it is the product of a plethora of contingencies. As such we get a tacit appeal to inter-subjective truth when we posit the ironist's defense of liberalism. This sort of truth positing cannot be affirmed by the ironist.

Another seeming inconsistency within the ironist position is that he sees himself as somehow historically privileged. Although his view of history has led him to the Nietzschean conclusion that any truth about man is necessarily a truth about man for a small period of time (perhaps within a given language), he still has based this view on a particular conception of history. He sees himself as having found some sort of objective truth about the ebb and flow of historical paradigms. Not only has he

discovered the truth of this assertion, but he lives his life in subservience to this fact. His ever changing self narrative, his attitudes toward others, are dominated by this realization. This seems like the ironist ironically takes his beliefs a bit too seriously, and therefore must reject a major tenet of his ironism.

If the ironist is such that he sees his final vocabulary as utterly contingent, what does it matter if he or she has put their stamp upon history? It is just something that will be seen as fodder for re-description by a future agent. Since the ironist sees his facticity governing himself as a bad thing, as something in the way of self narrative, why would he want to join in creating a potentially entangling factical paradigm for future agents to live within? Of course, this fact is unavoidable. If the changing of language is a result of many small contingencies, then of course every agent who uses a language could possibly (and unbeknownst to that agent) contribute to a change in the predominant paradigmatic language. Therefore, the intention of the ironist must be misplaced. If there is no way of knowing what particular states of affairs our thoughts might manifest as a result of discourse, it should not be desirable that one language be put in place of another.

This criticism lends itself to the idea of the private containment of ironism. Since language for Rorty is a causal mechanism, it seems unlikely that private irony can be contained. We have no certainty over which statements might or might not influence other agents' self descriptions. Therefore the affirmation of a distinction between public and private asserted is obtuse. For instance, as a philosopher, I continually read other philosophers, and in doing such an action, simply reading a book, my final

vocabulary is under the threat of being affected. Any proposition, unbeknownst to me, might be some sort of secret key for deconstructing my entire vocabulary.

This point is echoed in Charles Taylor's "Ethics of Authenticity". In his own attempt to bring authenticity into the liberal sphere, Taylor rejects the premise that authenticity is a purely self-created notion. Our familial, social and political relationships are instrumental in our personal pursuit of authenticity. Taylor recognizes that the culture of authenticity within liberal societies is one where the value of self-choice is paramount. However, the reality of the situation is that when we make choices, we don't simply value the choice, we value specifically what we choose to defend and its relationship to our daily lives. "On the intimate level, we can see how much an original identity needs and is vulnerable to the recognition given or withheld by significant others."[3] Basically, the self creating individual cannot atomistically enjoy self creation. He cannot keep it private. He must use the external world to validate his language. The liberal ironist of course, in his anxiety over the potential contamination of the public via his ideals, does not have this option open to him. We see that atomism necessarily undermines ironism, as that ironism has no means of expressing itself and therefore the ironist has no way of seeing his language as useful.

To say that we must in fact revere the public/private distinction in order to protect ironists seems obtuse. Past political and intellectual cultures have been far more repressive with regards to autonomy. This did not stop any of the past ironists from

[3] Charles Taylor, *The Ethics of Authenticity* (Harvard University Press, 1992) 49.

having their work influence future historical paradigms. If the ironist has such a strong grasp upon history, how does he need protection? This seems to me to raise the suspicion that Rorty is not actually concerned with a wariness of the public's self description being undermined by the ironist, but a protection of the ironist from external forces. For if the ironist is such that he is disposed to regard his self created vocabulary as ever in flux, what reason should the ironist have to be wary of political institutions possibly dominating his ends?

Let us look further at my claim that the ironist cannot possibly stop himself from the threat of contaminating public liberalism. The sources which cause a language user to adopt certain ways of looking at problems, creating his own meaning, cannot be intrinsic to the self. Self creation is entirely a process which is co-formed with and projected upon external forces. If one wants to reject that any force outside of the self should be used as a tool for self creation, then one rejects any possibility of self-creation. If we are unaware of how languages shape or might shape our future language, then how can the ironist save his own final vocabulary? Isn't his final vocabulary continually barraged by external language games? On top of this, the ironist is already skittish about his ever changing final vocabularies and self perceptions. Given this picture of an ironist, it seems unlikely that he might avert the possibility of (even unintentional) humiliation at the hands of other language users. What is left for Rorty's liberal ironist but an ever anxious, hermetic existence?

However, this is not the kind of life we live. Political and social concerns are implicit in the idea of self creation. We do not live in some kind of personal vacuum of our own intuitions. Our

relationship to others and the world (perhaps also history), is important to our definition of self and without these we cannot be the ever changing self determiners that Rorty wishes us to be. Without conditions of significance, reasons to care about something or other, we simply have no criterion with which to make choices valuable to us. If we are to accept Rorty's paradigm of self choice, for your life's story to remain untainted, we would of course (if possible) have infinite control over our final vocabularies, we would reduce the possibility of humiliation. But what kind of life would the ironist enjoy? His self narrative would consist of pure self created fantasy. It would be trivial without an external public to project his ideals upon. Ironically, by having infinite power for self assertion and value creation, the ironist would have removed his possibility of having such a power.

The idea that an ironist can live in this way, is of course ridiculous. It seems that what Rorty is concerned with is not in fact, the firm distinction between public and private spheres, but of the protection of the individual's self narrative against societal commandeering. The purpose of positing the public/private distinction in the first place, was an attempt at the reduction of humiliation and cruelty: the worst thing liberals do. But it was posited in order to protect the general public against the ironist. However, what it looks like is that the ironist himself is the one which is being protected by the distinction. Since self creation and therefore irony, cannot possibly be privatized, anyone and everyone is subject to the humiliation of the ironist, (including other ironists). In short, there is no guarantee that private irony will not "contaminate" the public notion of liberalism: the aversion to suffering. In affirming the public private distinction Rorty is not

saving the public from the ironist, but the ironist from public interference.

The ironist is also put in a peculiar psychological disposition with regard to his work. As we've noted, the ironist is such that he regards the relative unimportance of his self creation as the basis for his public liberalism. It is hard to see how the ironist can see his views as being important to the progression of history, but yet as unimportant to others. In fact, doesn't the ironist wish to influence other, futurally contingent ironists? Because of his break with his facticity, he is concerned with the progression of history: of specific futural agents' potential synthesis with his vocabulary.

These remarks show that the ironist is in fact concerned with something external: with his position and relationship to the evolution of language and therefore historical paradigms. He regards his existence as contingent upon his history, and also as his self-creation as relational to this history.

We might ask ourselves now, if an ironist is unconcerned with external forces when it comes to self creation, aren't we affirming a metaphysical transcendent? It seems like in affirming the individualization of the self, atomization, we are falling into a pitfall where self-hood is no longer questionable. The ironist is a deconstructionist on many fronts, he is able to laugh at his own final vocabulary and assert its meaningfulness, but at the same time he is on a particular side of the metaphysical pole, a side which his heroes like Nietzsche and Heidegger are antithetical: the subject-object distinction. In a post-metaphysical society, it is unclear how Rorty can possibly start with a metaphysical claim: the self exists. As this claim is part of a justification of the

public/private distinction, and because as an anti-metaphysician the ironist can reject this premise, it is hard to see why all ironists might adhere to it.

Another worry about the ironist position is that in adhering to a strict privatization of ironists, is one raised by Daniel Conway. If we are to privatize the ironists' pursuits, we necessarily force him into an anti-social hermetic existence. The liberal ironist is one whose liberalism comes before his ironism. As such, the ironist feels responsible not to influence the final vocabulary of others. But if the ironist is afraid of this notion, and he is unsure whether his language may or may not change the self describing actions of others, he might not have any reasons to perform acts of overt kindness. As Conway puts it, the "liberal ironists thus double conserve themselves, sequestering themselves in the private sphere and ingesting moral edification that may prevent future expenditures of cruelty."[4]

Rorty perhaps attempts to give us a way out of this. The liberal ironist, in his commitment to avert suffering, can attempt to understand the ways those who speak with different vocabularies might be humiliated. To do this he suggests the ironic liberal to study authors such as Nabokov and Orwell, authors who describe humiliation.

Again, we might look at this sort of provision and evaluate whether the ironist is the sort of being we wish to encourage. In addition to his private self creation, the ironist is also compelled to study artistic works. He is committed to not only knowledge of

[4] Daniel Conway "Taking Irony Seriously: Rorty's Postmetaphysical Liberalism," *American Literary History* 3, no. 1 (1991): 200.

historical paradigms, but of an understanding of different types of cruelty. What sort of moral imperative is Rorty giving to the ironist? This seems to me to be a direct violation of the ironist's metaphysical aversion. Even privately the ironist is seen to be dominated by his political affiliation with liberalism. The ironist is committed to a form of hyper liberal asceticism.

In his work, Rorty has attempted to give valid grounds for the promotion of ironists within our society. However, it seems that this characterization is good for neither liberals nor ironists. Though Rorty seeks to (furtively) increase the autonomy of the ironist, he implicates him in a life without a possibility for authentic self creation. The onus is placed upon the ironist himself to avert anti-liberal claims, whereas the liberal comes off scot free. As such, we would do good not to create a liberal society where a strong Rortian public/private distinction is honored.

## Bibliography

Conway, Daniel. "Taking Irony Seriously: Rorty's Postmetaphysical Liberalism." *American Literary History* 3, no.1 (1991): 198-208. Print.

Taylor, Charles. *The Ethics of Authenticity*. Harvard University Press, 1992.

Rorty, Richard. *Contingency, Irony, and Solidarity*. Cambridge: Cambridge UP, 1989.

# A Defense of a Wittgensteinian Outlook on Two Postmodern Theories

*Sarah Halvorson-Fried*

**Abstract** The way postmodern thinkers deal with issues of language and power has been highly influenced by Ludwig Wittgenstein's later philosophy of language. Wittgenstein's conception of language as a collection of "language-games" based on agreement in use rather than a direct reflection of objective reality is central to these issues. In this paper, I will show how this Wittgensteinian conception manifests itself in two important contemporary theories: the liberal ironism of Richard Rorty and the feminist philosophy of Luce Irigaray. I will show how Rorty's and Irigaray's Wittgenstein-influenced theories both bring Wittgenstein's philosophy of language into a more social context, and argue ultimately that through such theories we can better understand social issues in our modern world.

Much of postmodern theory deals with issues of language and power. According to many postmodern thinkers, most of the relationships between language and power go unnoticed, as the public usually sees language as a neutral medium within which we can communicate. But language has the power to oppress, the power to assign identities, the power to liberate. The way postmodern thinkers deal with these issues has been highly influenced by Ludwig Wittgenstein's later philosophy of language. In this paper, I will show how this influence manifests itself in two

important theories: the liberal ironism of Richard Rorty, a "distinctive and controversial [pragmatist]"[1] and the feminist philosophy of Luce Irigaray, a prominent name in the French school of feminism. I will respond to criticisms of Rorty that call his theory misrepresentative, and identify the disparity between Rorty's and Wittgenstein's goals as a vital reason to accept Rorty's invocation of Wittgenstein. I will identify Wittgensteinian concepts in Irigaray's feminism and establish a similar disparity in goals. I will then use a Wittgensteinian reading of Irigaray to illustrate the purpose and value of analyzing postmodern theory under a Wittgensteinian lens. Ultimately, I believe that it is through such a lens that we can better understand many postmodern approaches to the relationship between humans, language and the world. In particular, I will show in this paper that his conception of language as based on agreement in use is central to both Irigaray's feminism and Rorty's liberal ironism.

**I. Rorty's Use of Wittgenstein**

Rorty refers to Wittgenstein's later work in order to argue against the prevailing acceptance of universality and representation of truth in political and philosophical systems. In *Contingency, Irony, and Solidarity*, he criticizes the basing of political systems on sweeping political theories and ideologies and proposes a new "politics of redescription." In *Philosophy and the Mirror of Nature*, he criticizes the epistemological tradition of Western philosophy, disparaging its perception of the ability to discover truth, and

[1] Bjorn Ramberg, "Richard Rorty," (*Stanford Encyclopedia of Philosophy*, 2007), http://plato.stanford.edu/entries/rorty/.

proposes a turn in philosophy toward a more conversational, less argumentative and truth-value-based approach. In both works, Rorty uses Wittgensteinian philosophy as a defense for his rejection of universalizing systems.

In *Contingency, Irony, and Solidarity*, Rorty spells out implications of Wittgensteinian philosophy of language, identifying Wittgenstein as one important thinker who revealed the human-created, shifting nature of "vocabularies." Rorty's "vocabularies" can be thought of as analogous to Wittgensteinian "language-games" and refer to specific cultural collections of ways of thinking, communicating, and acting (ways of living). Rorty argues that if vocabularies are indeed created contingently and in constant shift, if they are "optional and mutable,"[2] then the values they express, too, are optional and mutable. He asserts that neither the vocabularies nor their values should be imposed on anyone, and that political systems should seek to include *multiple* vocabularies. Such systems he terms "liberal utopias," inhabited by "liberal ironists" who would recognize their own contingency, acknowledging the possibility of shifting truth and shifting morality, which continue to change as they are influenced by different (contingent) factors. Seeking to provide people with the most freedom of expression possible and alleviate the most suffering possible (this is the "liberal" part), they would promote their causes through redescriptions rather than arguments.[3]

Like Nietzsche, Freud, and Donald Davidson, Wittgenstein is a stepping-stone on the path to Rorty's land of liberal utopias,

[2] Ibid.
[3] Richard Rorty, *Contingency, Irony, and Solidarity* (Cambridge: Cambridge University Press, 1989), 9.

where we all recognize contingency. According to Rorty, Wittgenstein helped us along this path by revealing the contingency of *language*: In positing that language *forms* an objective framework based on agreement rather than *adhering* or *corresponding* to an (already-existing) objective framework, Wittgenstein makes us see language as a product of historical contingencies. Here it is useful to explore Rorty's use of Donald Davidson's philosophy of language, another stepping-stone. Davidson, like Wittgenstein, asserted that what makes language work is understanding between speakers, not expression of truth. Davidson's notion of "passing theories" from his 1986 paper "A Nice Derangment of Epitaphs" states that understanding between two linguistic beings occurs when their concepts of a word's meaning converge. Each person's concept of each word's meaning is in constant shift relative to context, so understanding – and meaning – are also in constant shift. This assertion helps us recognize the contingency of language by revealing its lack of necessity, like Darwin's theory of evolution revealed the contingency of the biology of species.

> Davidson lets us think of the history of language, and thus of culture, as Darwin taught us to think of the history of a coral reef. . . . Our language and our culture are as much a contingency, as much as a result of thousands of small mutations finding niches (and millions of others finding no niches), as are the orchids and the anthropoids.[4]

[4] Ibid., 16.

Just as the present state of species has depended on many contingent factors, so has our language. Rather than an expression of or correspondence to reality, it is somewhat a product of chance: Things could easily be otherwise. In addition, they are bound to continue to change. For this reason, according to Rorty, no singular ideology can be the right one: The circumstances under which ideologies and social theories come into being will never be static. As situations change, so should the vocabularies we use and the values on which our political systems are based.

Rorty does for philosophy in *Philosophy and the Mirror of Nature* what he does for politics in *Contingency, Irony, and Solidarity*, presenting this idea of redescription rather than appeal to universal truth within the discipline of philosophy. In this book, Rorty criticizes the epistemological tradition and details what he sees as a necessary shift in Western philosophy. He uses the arguments of several philosophers, including Wittgenstein, to critique the representational view of knowledge central to traditional epistemology. According to Rorty, Wittgenstein, along with Sellars, Quine, Kuhn, and Davidson, showed that neither the mind nor language is capable of mirroring reality. Subsequently, the discipline of philosophy had to change, because epistemology ceased to make sense.[5] As such, the traditional questions of philosophy are no longer relevant to our time. They are not, as many believe, timeless. The last sentence of his book reads,

> The only point on which I would insist is that the philosophers' moral concern should be with

[5] Richard Rorty, *Philosophy and the Mirror of Nature* (Princeton: Princeton University Press, 1979, 169.

> continuing the conversation of the West, rather than with insisting upon a place for the traditional problems of modern philosophy within that conversation.[6]

We should not "insist on a place" for these traditional problems precisely because they will not, as so many philosophers have believed, lead us to discovery of universal truths. When we do philosophy, according to Rorty, we should neither assume that we operate outside the boundaries of contingency nor that we have a privileged ability to discover "truth." Rather than some sort of elevated search for truth, he claims that our Western tradition of philosophy is just another vocabulary (or language-game).

Instead, as in *Contingency, Irony, and Solidarity*, Rorty would have us enter a more conversational approach. Once more, Wittgenstein's influence is clear. Under Rorty's "naturally holistic conversational justification," which he favors over the "reductive and atomistic" justification of the epistemological tradition, social justification of belief creates knowledge. Just as language finds objectivity of meaning in social agreement under Wittgenstein, so does knowledge find objectivity of truth in social agreement under Rorty. Under this view, philosophy as a search for truth is nonsensical: We "have no need to view [knowledge] as accuracy of representation" since "we understand knowledge when we understand the social justification of belief."[7] Rorty terms this view "epistemological behaviorism" and once again attributes his theory to Witgensteinian influence.

---

[6] Ibid., 394.
[7] Ibid., 170.

> Explaining rationality and epistemic authority by reference to what society lets us say, rather than the latter by the former, is the essence of what I shall call 'epistemological behaviorism,' an attitude common to Dewey and Wittgenstein.[8]

And for Rorty, if we recognize philosophy's inability to discover truth in any objective sense, then we should change the discipline. Just as in *Contingency, Irony, and Soliarity*, Rorty would have us reject a privileged, contingently created position of philosophy in favor of a conversational discipline inclusive of multiple language-games.

A legitimate worry for many critics is that Rorty simultaneously makes normative claims while rejecting normativity. This may indeed be a problem for Rorty, but for the purposes of this paper it is not relevant. My task here is to show the validity of Rorty's invocation of Wittgenstein. Another worry is that in expounding on the created nature of meaning, Rorty is rejecting objectivity of meaning in any form; in ordinary words, for instance, like "apple" or "table." Such a rejection would make Rorty an anti-realist. I do not think he aims to do this: Rorty's concern is primarily with the abandonment of essential identities in order to allow for shifting notions of selves, cultures, and truths. He makes this distinction himself in *Contingency, Irony, and Solidarity*.

> We need to make a distinction between the

[8] Ibid., 174.

> claim that the world is out there and the claim that the truth is out there. To say that the world is out there, that it is not our creation, is to say, with common sense, that most things in space and time are the effects of causes which do not include human mental states. To say that the truth is not out there is simply to say that where there are no sentences there is no truth, that sentences are elements of human languages, and that human languages are human creations.[9]

Rorty is decidedly not an anti-realist, though he does have a pluralist notion of truth: Since truth is not "out there," since it is created by humans, it can be created in many ways. The last worry I will explore in the next section: that in fact Rorty may not be able to use philosophers like Wittgenstein as he does; that he may be misrepresenting them and that his use of Wittgenstein may be unfounded.

**II. Is Rorty's Use of Wittgenstein Valid?**

Rorty makes bold claims when he uses philosophers like Wittgenstein to support his politics and philosophy of redescription. Is this use valid? We might ask, as some have: How can Rorty make the jump from Wittgenstein's notion of language as use to "contingency of language" in *Contingency, Irony, and Solidarity*? Does Wittgenstein really exhibit language's contingency? Does Rorty accurately represent Wittgenstein in *Philosophy and the Mirror of Nature*, when he cites Wittgenstein as one of the philosophers who changed the nature of

[9] Rorty, *Contingency, Irony, and Solidarity*, 5.

epistemology? Does he interpret Wittgenstein's notions of language-games and of language as agreement correctly? I argue first that he does in fact represent Wittgensteinian concepts of language accurately, and second that these questions are somewhat inappropriate, because Rorty and Wittgenstein have very different goals. Wittgenstein is trying to determine the nature of communication. His task is quite an apolitical one: He simply wishes to discover the true nature of language, and he discovers it to be a practice based on custom. Rorty has a larger goal in mind: He wishes both to convince us that all of our practices based on custom are not necessarily right, that we cannot justify anything with an appeal to "truth" since everything we do and think is not necessary but contingent, and to propose new systems – of society and of philosophy – based on this recognition. It is because of this disparity of purpose that Rorty's use of Wittgenstein is not, as some critics have proposed, invalid. Rather, Wittgenstein's philosophy of language, like Darwin's theory of evolution, is useful to Rorty for purposes of illustration: Wittgenstein serves both as a useful comparison and as an important predecessor. In appealing to Wittgenstein, Rorty is simply laying out for the reader Wittgenstein's influence on his own theory.

Wolf Rehder is one of these critics. In "Hermeneutics versus Stupidities of All Sorts: A Review-Discussion of R. Rorty's 'Philosophy and the Mirror of Nature,'" Rehder disparages Rorty for his use of philosophers like Wittgenstein.

> As witnesses for his holistic, antifoundationalist, and pragmatist new view of philosophy as hermeneutics, Rorty calls, among others,

> Foucault, Dewey, Wittgenstein, Sartre, Kierkegaard, Quine, Gadamer, Feyerabend and Heidegger, a truly motley group of big names. However, he makes only makes a meager case against epistemology and traditional philosophy with this impressive phalanx of witnesses for the prosecution. It is not going too far to say that his backing up his case with this echelon of genuinely great men does not only not do justice to their philosophical work, but even tends to demean their work and their role in the history of philosophy. This is so, because Rorty's 'positive' case, his hermeneutic turn and proposed transcending of truth-oriented inquiry is, unfortunately, surprisingly naïve.[10]

It is naïve, according to Rehder, because there cannot be useful conversation without conflict, nor can it exist without a common language or discourse. In Rehder's view, Rorty is proposing the opposite: agreement between different languages and discourses. "Any fruitful discussion is based on some sort of disagreement."[11] This is a commonly held view: To engage in conversation, we must share a language-game; and to debate, we must disagree. It seems to me, though, that in criticizing Rorty on this point Rehder is simply not taking Rorty seriously: Rorty's point is that useful conversation is possible – better, even – if it considers perspectives of multiple vocabularies. To say that useful conversation must happen within the same vocabulary is to refuse Rorty's proposed

[10] Wulf Rehder, "Hermeneutics versus Stupidities of All Sorts: A Review-Discussion of R. Rorty's 'Philosophy and the Mirror of Nature," *Zeitschrift für allgemeine Wissenschaftstheorie / Journal for General Philosophy of Science* 14, no. 1 (1983): 95, http://www.jstor.org/stable/25170640.
[11] Ibid., 96.

shift, to disregard his entire point of making the discourse of philosophy more inclusive of multiple language-games. Rorty's usage of all of these philosophers to defend his "naïve" system obviously troubles Rehder. After all, he says, "[It] does not only not do justice to their philosophical work, but even tends to demean their work and their role in the history of philosophy." It is this criticism that I will now address.

First, Rorty does seem to accurately represent Wittgenstein. Wittgenstein created a new framework for objectivity based on social agreement rather than on truth. This agreement in no way determines truth or falsity, but instead forms a new standard of objectivity. In response to the invisible interlocutor in section 241, "So you are saying that human agreement decides what is true and what is false?" Wittgenstein offers an alternative: "It is what humans *say* that is true or false; and they agree in the *language* they use."[12] Agreement does not determine truth in the world, only truth in our agreed-upon shared account of the world – in our shared language. It is this agreement that allows us to communicate with one another. People are understandable when their definitions accord with socially accepted ones. When Rorty says that Wittgenstein "[explains] rationality and epistemic authority by reference to what society lets us say, rather than the latter by the former,"[13] he seems to be correct: Wittgenstein's account of a socially formed objective framework does conform to Rorty's "epistemological behaviorism," as it locates objectivity in social accordance.

[12] Ludwig Wittgenstein, *Philosophical Investigations*, trans. G.E.M. Anscombe (New York: Macmillan Publishing Co., Inc., 1953), 88.
[13] Rorty, *Philosophy and the Mirror of Nature*, 174.

Second, it is useful to ask *why* Rorty appeals to "this impressive phalanx of witnesses." Does he aim to represent them? Given the difference in Rorty's and Wittgenstein's goals, strict adherence is not necessarily essential. Any apparent disparity between Rorty's Wittgenstein's systems is unimportant, because Rorty and Wittgenstein are not making the same kind of claim. They are not talking about the same kind of thing. When Rorty says, "the truth is not out there,"[14] he does not mean that we create the objective world. Indeed, he explicitly distinguishes between "the claim that the truth is not out there and the claim that the world is not out there."[15] He means that our social and cultural institutions, our beliefs, our methods of inquiry (like philosophy) are created in the same way that language is, in the same way that evolution is. Rorty does not really claim to adhere to Wittgenstein, so he cannot be criticized for it. In both *Philosophy and the Mirror of Nature* and *Contingency, Irony, and Solidarity*, Rorty invokes Wittgenstein as an important influence, but not as his only influence. Where Wittgenstein's goal is to discover and describe, Rorty's is to reveal, convince, and change.

**III. Illumination Through Irigaray**

Irigaray is Wittgensteinian in many of the same ways as Rorty: She holds a pluralist view of truth, rejects normativity, and uses Wittgenstein's notions of language-games and forms of life. But because she does not invoke Wittgenstein's name to defend her views, as Rorty does, she is never criticized for misrepresentation,

[14] Rorty, *Contingency, Irony, and Solidarity*, 5.
[15] Ibid., 5.

as Rorty is. This fact reveals Rorty's immunity to such criticism. Her theory also illustrates the effectiveness of applying later Wittgensteinian philosophy to postmodern theories. Through an exploration of her work, I hope to show this usefulness.

In *To Speak Is Never Neutral*, Irigaray questions the assumed impartiality of language and calls on us to recognize both its sexed nature as "the language of man" (a title of one of her chapters) and its unfairly universalizing tendencies. She states in her introduction, "This book is a questioning of the language of science, and an investigation into the sexualization of language, and the relation between the two."[16] In "Linguistic Sexes and Genders," she identifies the sexism inherent in language, examining particular words in her native French. In "This Sex Which Is Not One," she states that "female sexuality has always been theorized within masculine parameters"[17] and attempts to conceptualize it differently, outside these parameters. One of Irigaray's main concerns throughout her various works is to show how the current linguistic system is oppressive to women while claiming to be universally neutral, an idea clearly influenced by Wittgenstein, as I will show. Another concern is to show how change is possible through new feminist language-games, the details of which can be confusing and have been debated, but which is clarified through a Wittgensteinian reading of her theory.

Irigaray uses the Wittgensteinian notion of language-games

[16] Luce Irigaray, *To Speak Is Never Neutral*, trans. Gail Schwab (New York: Routledge, 2002), 5.
[17] Luce Irigaray, "This Sex Which Is Not One," trans. Claudia Reeder, in *New French Feminisms*, ed. Elaine Marks and Isabelle de Courtivron (New York: Schoken Books, 1981), 99.

as well as his conception of objectivity as agreement to describe the problem of a universal language that is catered toward men but purported to apply to women as well. According to Irigaray, the language we accept as universal – the language of politics, of science, of philosophy – is actually an oppressive, particular language-game.

> A sexed subject imposes his imperatives as universally valid, and as the only ones capable of defining the forms of reason, of thought, of meaning, and of exchange. He still, and always, comes back to the same logic, the only logic: of the One, of the Same. Of the Same of the One.[18]

Just as, in Wittgenstein, we cannot form a private language because all words we use are defined by the linguistic community, so, in Irigaray, is it nearly impossible to escape from the purportedly universal dominating male language-game. In the same vein as Rorty, Irigaray questions the value of rationality and criticizes the language of traditional philosophy, which is decidedly male and which is imposed on women while masking itself as universal to all.

> From [Irigaray's] point of view, the philosophers, of whatever persuasion, are comfortably installed in the male imaginary, so comfortably that they are completely unaware of the sexuate character of 'universal' thought.[19]

[18] Irigaray, *To Speak Is Never Neutral*, 228.
[19] Margaret Whitford, *Luce Irigaray: Philosophy in the Feminine* (New York: Routledge, 1991), 103.

How, then, is feminist theory even possible? The problem is as follows: "Not using logic risks maintaining the other's status as *infans* . . . Using logic means abolishing difference and resubmitting to the same imperatives."[20] If we operate outside the dominating language-game, we will not be taken seriously, and if we operate within it, we are giving in, trying to fit ourselves into the oppressive system.

Irigaray's solution, possible under Wittgensteinian influence, is to form a new language-game that challenges this discourse. Irigaray appeals to the female body in the formation of a new language of feminism, under two assumptions: First, that the male body is already intrinsic to philosophy – in ethics, for instance, where the point is to enhance positive effects on the body (e.g., health) and circumvent negative effects (e.g., death). Second, that the female body is currently defined by male desire and male language.[21] The body is important both in the symbolic and in its realized form for Irigaray. Rather than being forced to conform either to the supposedly universal language of men, based on the male body, or to form a new language based on the male-created female body, "the female body has to be allowed its own imaginary existence in the form of symbolic difference."[22] This imaginary existence can only be realized by privileging female life, female sexuality, and the real female body, as they are "for themselves."[23] Irigaray's solution is Wittgensteinian because it relies on Wittgenstein's notions of language-games as flexible, changing

[20] Irigaray, *To Speak Is Never Neutral*, 228.
[21] Whitford, *Luce Irigaray*, 150.
[22] Ibid., 103.
[23] Irigaray, "This Sex Which Is Not One," 106.

and organic and of language as a form of life. Formation of a *new* language-game is possible because language-games are always coming into and out of being. The female body itself is an important part of the female form of life, and so can be appealed to in Irigaray's formation of a new feminine language-game.

Importantly, Irigaray does not declare herself Wittgensteinian; but a Wittgensteinian reading of Irigaray both makes sense, as I have shown, and clarifies some aspects of her solution. Joyce Davidson and Mick Smith show how such a Wittgensteinian reading clarifies and does justice to Irigaray in "Wittgenstein and Irigaray: Philosophy and Gender in a Language (Game) of Difference." Specifically, a Wittgensteinian reading solves an interpretative conflict among Irigaray scholars. Critics have typically either called Irigaray essentialist, which she explicitly claims not to be (her disparagement of universalizing language is clearly anti-essentialist) or as speaking in metaphor or symbolism when she speaks about the body (since they know she is anti-essentialist, they cannot imagine she would invoke the *real* body). Even Margaret Whitford, a prominent Irigaray scholar, acknowledges the difficulty of reading Irigaray, in that "we are not quite sure what status is given to Irigaray's statements."[24] She wonders whether they are "empirical descriptions . . . ideal descriptions . . . descriptions of the reigning imaginary . . . or perhaps simply metaphors again."[25] Reading Irigaray under a Wittgensteinian lens, say Davidson and Smith, "might provide a

[24] Whitford, *Luce Irigaray*, 102.
[25] Ibid.

third alternative"[26] and solve this conflict: Through Wittgenstein, we can come to terms with Irigaray's simultaneous rejection of essentialism and appeal to the body in formation of a new, subversive, feminine language-game. Wittgenstein's notion of "blurred concepts" or "family resemblances" lets us recognize the possibility of using something like the female body to create a new language-game without essentializing it.

> Women's anatomy might be understood as a real component of the patterns, context, and environment that might give rise to a feminine language-game. So, while anatomy is not an *essential* referent to which language must be fixed, it is a valid and pertinent feature of a feminine form of life.[27]

Wittgenstein told us that definitions need not always be fixed, that a "the indistinct [picture] is often exactly what we need."[28] In the same way, female anatomy need not be essentialized to serve as a reference point for the creation of a feminine language-game. We see, then, that Wittgensteinian philosophy does not only manifest itself in Irigaray's theory; it can also help clarify it.

## IV. A Difference of Goals: Language and Power

Like Rorty, Irigaray has a political goal, one that is vastly different from Wittgenstein's descriptive one. Rorty and Irigaray

[26] Joyce Davidson and Mick Smith, "Wittgenstein and Irigaray: Gender and Philosophy in a Language (Game) of Difference," *Hypatia* 14, no. 2 (1999): 83, http://www.jstor.org/stable/3810769.
[27] Ibid., 84.
[28] Wittgenstein, *Philosophical Investigations*, 34.

both assume that language has power: In both of their theories, it is language that oppresses and language that has the power to liberate. This relationship between language and power was termed "discourse" by Michel Foucault, and refers to language and other shared aspects of culture as a mechanism that perpetuates itself through use, never calling itself into question. Central to this idea is the Wittgensteinian one that language is based on agreement in use, that social agreement in use forms the objective frameworks within which we communicate. Wittgenstein was the philosopher to assert that there was no ideal language capable of representing reality. Maxine Greene says in "Postmodernism and the Crisis of Representation" that our postmodern task "may be a matter of recognizing that there is no single-dimensional medium reflective of the 'facts' of the world, but a multiplicity of language games, as Ludwig Wittgenstein made so clear."[29] Postmodern thinkers like Foucault, Rorty, and Irigaray, as well as Judith Butler, Monique Wittig, and Edward Said, among others, have accepted this task, drawing out the social and political implications of Wittgensteinian philosophy of language.

Wittgenstein thus proves to be invaluable to postmodern theories of language and power: Though Wittgenstein never approaches the social and political ideas that theorists like Rorty and Irigaray do, his work is ultimately their basis. For this reason, and as we have seen through these two case studies, a Wittgensteinian reading of postmodern theories helps us understand them.

[29] Maxine Greene, "Postmodernism and the Crisis of Representation," *English Education* 26, no. 4 (1994), 208.

## V. Eliminating False Clarity: The Value of Wittgenstein-Influenced Postmodern Theory

Both Rorty and Irigaray use Wittgensteinian notions of language and social agreement to call into question the universality we so often use to solve political, philosophical, and scientific problems. Irigaray questions the universality of political, philosophical, and scientific language, while Rorty questions the ability of universalizing, truth-seeking systems of politics and philosophy to provide us with acceptable solutions. I once heard in an ecology class that "our best chance of solving problems is to recognize the complexity of the situation rather than appeal to an ideology." The professor said such an appeal gives us "false clarity." It seems to me that this is true, that more realistic views do not think themselves universal, and that Rorty's and Irigaray's Wittgenstein-influenced theories that seek to reveal the complexity of the situation in lieu of the false clarity of universalizing political, philosophical, and linguistic systems are ones to consider with utmost seriousness and thoughtfulness.

## Bibliography

Davidson, Joyce and Mick Smith. "Wittgenstein and Irigaray: Gender and Philosophy in a Language (Game) of Difference." *Hypatia* 14, no. 2 (1999): 72-96. http://www.jstor.org/stable/3810769.

Greene, Maxine. "Postmodernism and the Crisis of Representation." *English Education* 26, no. 4 (1994). 206-219.

Irigaray, Luce. "Linguistic Sexes and Genders." Translated by Alison Martin. In *The Feminist Critique of Language*, edited by Deborah Cameron, 119-123. London: Routledge, 1993.

Irigaray, Luce. "This Sex Which Is Not One." Translated by Claudia Reeder. In *New French Feminisms*, edited by Elaine Marks and Isabelle De Courtivron, 99-106. New York: Schocken Books, 1981.

Irigaray, Luce. *To Speak Is Never Neutral*. Translated by Gail Schwab. New York: Routledge, 2002.

Ramberg, Bjorn. "Richard Rorty." *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/rorty/.

Rehder, Wulf. "Hermeneutics versus Stupidities of All Sorts: A Review-Discussion of R. Rorty's 'Philosophy and the Mirror of Nature." *Zeitschrift für allgemeine Wissenschaftstheorie / Journal for General Philosophy of Science* 14, no. 1 (1983): 81-102. http://www.jstor.org/stable/25170640.

Rorty, Richard. *Contingency, Irony and Solidarity*. Cambridge: Cambridge University Press, 1989.

Rorty, Richard. *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press, 1979.

Whitford, Margaret. *Luce Irigaray: Philosophy in the Feminine*. New York: Routledge, 1991.

Wittgenstein, Ludwig. *Philosophical Investigations*. Translated by G.E.M. Anscombe. New York: Macmillan Publishing Co., Inc., 1953.

# THE NARRATIVE SELF-CONSTITUTION VIEW: WHY MARYA SCHECHTMAN CANNOT REQUIRE IT FOR PERSONHOOD

*Andrew S. Lane*

**Abstract** In her book *The Constitution of Selves*, Marya Schechtman names four features essential for personal existence: survival, moral responsibility, self-interested concern, and compensation. She rejects reductionists theories of persons, specifically that of Derek Parfit, claiming that they cannot support the four features. Instead, she proposes a theory of persons which she calls the Narrative Self-Constitution View. Because she believes this is required to support the four features, she also argues that for an individual to be a person they must hold this view. Drawing from the work of Derek Parfit and Galen Strawson, I will argue that her arguments are inconsistent and do not show that reductionist theories cannot support the four features. As a result, I conclude that Schechtman is wrong to require the Narrative Self-Constitution View for personhood.

This paper will deal with the theory of personal identity proposed by Marya Schechtman in her book, *The Constitution of Selves*.[1] In this work, Schechtman claims that there are four basic features of personal existence: survival, moral responsibility, self-

[1] Marya Schechtman, *The Constitution of Selves* (Ithaca, NY: Cornell University Press, 1996).

interested concern, and compensation.[2] These she abbreviates as the "four features." Regardless of potential additions or emendations to the list, I will not question these features. As far back as John Locke, accounting for moral responsibility is a key motivation for personal identity theory and this continues with more contemporary philosophers like Derek Parfit. Moral accountability seems required for a functional society. If a person at time $T_1$ does not survive and there is a new person at time $T_2$, generally our intuition is that the person existing at time $T_2$ would not be responsible for the actions of person existing at time $T_1$. Thus, it seems necessary that a person must survive across time to some extent, otherwise nobody could be held accountable for past actions. The work of Galen Strawson will be useful in considering this question of moral accountability. Self-interested concern and compensation also seem necessary for "personal" existence, though not for "impersonal" existence. It is not of necessity that the former is better than the latter, but this essay will set such considerations aside. I will take personal survival to be a valid target, which is Schechtman's aim, without justifying whether or not it is any better than impersonal survival. Schechtman believes that Reductionist views, like that of Derek Parfit, cannot capture the four features, and thus fail as accounts of personal identity. Instead, she advocates what she calls the Narrative Self-Constitution View, which she feels is required in order to capture the four features.

The Narrative Self-Constitution View holds that a person creates his or her identity by forming an autobiographical

[2] Schechtman, *Constitution*, 2.

narrative. According to this view,

> the difference between persons and other individuals...lies in how they organize their experience, and hence their lives. At the core of this view is the assertion that individuals constitute themselves as persons by coming to think of themselves as persisting subjects who have had experience in the past and will continue to have experience in the future, taking certain experiences as theirs. Some, but not all, individuals weave stories of their lives, and it is their doing so which makes them persons.[3]

Those who do not adhere to this narrative view, those who do not think of themselves as persisting subjects and construct narratives, are not persons according to Schechtman. I claim, however, that the narrative self-constitution view is not the only way to capture the four features. As a result, Schechtman is wrong to deny personhood to individuals who do not view themselves narratively. The motivation for her requirement that an individual view themselves narratively is that to account for personal existence, we need to capture the four features; thus, if we can capture the four features another way, while this does not exclude her narrative view as one of the potential ways, which I believe it is, it is not required, and thus individuals who are non-narrative should not be excluded from personhood.

**The Reductionist View of Derek Parfit**

Before considering the views of Derek Parfit, it will be

[3] Schechtman, *Constitution*, 94.

useful to establish some distinctions concerning identity. First, there is the distinction between numerical and qualitative identity. For example, take two sheets of printer paper. The two sheets are qualitatively identical, for they share the same qualities, but are not numerically identical, because they are two different physical objects. While the two sheets are not numerically identical with each other, each is numerically identical with itself; each is one and the same sheet of paper. This is one of the basic principles of logic: self-identity.

Second, there is strict and non-strict identity. Strict identity requires that $X_1$ and $X_2$ be exactly the same in all ways; the smallest change of any kind destroys the strict identity of the objects. With non-strict identity, however, some change is permitted without destroying the non-strict identity of the objects. With the paper example, $X_1$ and $X_2$ are not strictly qualitatively identical, because if we compare closely enough, the fragments of pulp are not arranged in exactly the same configuration. Strict identity in this case would require that all the atoms making up the paper, and their arrangement, be exactly qualitatively identical. However, $X_1$ and $X_2$ may be considered non-strictly identical. For most purposes, it would be more useful to a person to consider $X_1$ and $X_2$ (non-strictly) qualitatively identical, because what matters to us about the sheets of paper is not on the level of atoms; for our purposes the sheets are qualitatively identical. The criteria for what qualifies for non-strict identity will vary depending on the objects in question, and this will be dependent on the perspective of who is considering the objects and their purposes. The strict/non-strict distinction applies to numerical identity as well. With the problem of personal identity, the two objects in question will be in different

temporal locations. To say that the paper is self-identical in a given moment considers space, while the time aspect is constant. When considering whether the paper is numerically identical at different moments also considers time. Strictly, the paper would not be identical at different moments because the atomic makeup will have slightly changed, for example from the effects of light. However, we may say that they are non-strictly identical if all that has changed are the atomic differences from light, because these differences are irrelevant to what matters to us about paper.

One of Parfit's central concerns is moral accountability. As mentioned in the introduction, if a person at time $T_1$ is not the same person at time $T_2$, then it seems that the person at time $T_2$ could not be held accountable for the actions of the person at time $T_1$, for they are not the same person. When we look at an individual across time, they are never strictly identical at two different times. Atoms have changed and psychological makeup is in constant flux. Thus, when speaking of an individual at two different times, they are never strictly-identical on a reductionist account. If one holds that there is, as Parfit would say, a further fact of identity, then one may argue that there can be strict identity across time. If, for example, there were an immaterial, eternal substance, perhaps a soul, and this substance provides identity, then it may be strict identity. None of the philosophers discussed in this essay argue for such a substance, and because it is not within the scope of this paper to properly argue against it, I will set this possibility aside. The person at two different times may, however, be non-strictly identical. The question then becomes, what criteria should we use to decide whether or not they are (non-strictly) identical? For Parfit, the mind is more important than the body and thus seems

the natural place to locate this identity. As a result, he articulates psychological criteria for identity.

For this, Parfit defines three terms: psychological connectedness, strong psychological connectedness, and psychological continuity. Psychological connectedness is "the holding of particular direct psychological connections."[4] Parfit cites memories, beliefs, desires and intentions as examples of individual psychological connections. For example, if a person at age 18 has the memory of running from a dog when they were younger, and this person still has this memory when they are 20, this would be an example of a direct psychological connection. Parfit claims, "since connectedness is a matter of degree, we cannot plausibly define precisely what counts as enough. But we can claim that there is enough connectedness if the number of direct connections, over any day, is at least half the number that hold, over every day, in the lives of nearly every actual person."[5] Strong connectedness means over half of the possible psychological connections hold. Strong connectedness is not transitive. A person at time $T_1$ may be strongly connected to the person at time $T_2$, and the person at time $T_2$ to the person at time $T_3$, but it does not follow that the person at time $T_3$ is strongly connected to the person at time $T_1$. Psychological continuity is "the holding of overlapping chains of *strong* connectedness." [6] While strong connectedness is not a transitive relation, psychological continuity is. Thus, the person at time $T_3$ would be psychologically

[4] Derek Parfit, *Reasons and Persons*, (Oxford, Oxfordshire: Clarendon Press, 1984), 206.
[5] Ibid.
[6] Ibid.

continuous with the person at time $T_1$, because they are linked through time $T_2$ to which they are both strongly connected. A person at two different times may be considered (non-strictly) identical if and only if they are psychologically continuous. Like the Buddhists and David Hume, Parfit claims that there is no Self, where the Self would be an unchanging entity or essence that can provide identity for an individual across time. That is, there is no "further fact" of identity; identity simply consists in holding psychological continuity.

**The Extreme Claim and the Moderate Claim**

In his book *Reasons and Persons,* Derek Parfit claims that we are Selfless persons, that there is no Self to provide the further fact of identity, instead claiming that our identity simply consists in overlapping chains of strong psychological connections, but thinks that this is not such a terrible thing. In fact, he feels that adopting this view was a positive change in his life. In response to his view, however, he sees two possible reactions; one he calls the Extreme Claim, the other the Moderate Claim.

The Extreme Claim says that "if the Reductionist view is true, we have *no* reason to be concerned about our own futures."[7] If in the future, my future self will not be the same person as my current self, then I have no reason to care for this person. It is not me. For example, why should I care if smoking damages my body, for it will not be me who dies of cancer. The Moderate Claim, however, says that psychological continuity with a high degree of connectedness gives us a reason to be concerned for our future

[7] Ibid., 307.

selves.[8] Parfit believes that even though it will not be the same person in the future by strict criteria, it could be the same person on a reductionist account, and the present person may still have concern for the future person. He likens this to how we may be concerned for our children, even if they are not us. The relations that justify this are not a deep separate fact. If these relations give us reason to care, then psychological continuity may give us reason.

However, one may still object that it will not be one in the future, so why should one be especially concerned today about what one shall care about in the future? Why should a person care about either their future selves or other people's future selves? To this, Parfit says that he does not have an argument to completely refute the extreme claim. Both claims, he thinks, are defensible. Though, he believes that we are not forced to accept the extreme claim. He wonders,

> It may be wrong to compare our concern about our own future with our concern for those we love. Suppose I learn that someone I love will soon suffer great pain. I shall be greatly distressed by this news. I might be *more* distressed than I would be if I learnt that *I* shall soon suffer such pain. But this concern has a different quality. I do not *anticipate* the pain that will be felt by someone I love.[9]

Thus, because he cannot refute the Extreme Claim, he accepts it as a defensible response to his position. However, he maintains that

[8] Ibid., 311.
[9] Ibid., 312.

the Moderate Claim is also defensible. Neither claim, he thinks, necessarily follows from his theory. Which claim a person holds will depend on the feeling of that person.

**Schechtman's Argument from the Extreme Claim**

Schechtman believes that reductionism cannot support the Moderate Claim and as a result we are forced to accept the Extreme Claim. Because the Extreme Claim cannot support the Four Features, Reductionism, she claims, cannot be true. She maintains that instead of accepting this as an interesting result of Parfit's theory of personal identity, it should be seen as a *reductio ad absurdum* of Parfit's reductionist account, because it cannot support the four features.[10] Her argument has two premises. Premise 1 is that "the four features require numerical identity–qualitative similarity will not do."[11] This is because "self-interested concern is an emotion that is appropriately felt only toward my own self and not toward someone like me. We all know the difference between fearing for our own pain and fearing for the pain of someone else."[12] As Parfit himself recognized, this is a difference of kind and not of degree. While we may potentially care about another person's pain more than our own, we do not "anticipate" the pain. Premise 2 claims that "the psychological continuity theory collapses the distinction between someone *being* me and someone being *like* me–that all identity amounts to on this view is psychological similarity between distinct individuals."[13]

[10] Schechtman, *Constitution*, 63.
[11] Ibid., 52.
[12] Ibid.
[13] Ibid., 53.

Schechtman believes that the extreme claim follows from these premises. If there is no difference between being the same person, and being like a different person, how can we decide if it is the same person, and thus how could we consider them to have self-interested concern? If qualitative similarity between distinct individuals is insufficient to underlie the four features, then the continuity theory fails to account for the importance of identity. She believes that to avoid the Extreme Claim, we need a theory where one and the same experiencing subject can exist at two different times; if person-stages are the only subjects that have experience in the theory, and person stages are not of the same subject, then this cannot happen.[14]

**The Tribal Example**

Regardless of the Extreme Claim, Parfit insists that, even though his rejection of the Non-Reductionist view led him to be less concerned about his future, he was still more concerned about his own future than that of a mere stranger.[15] To account for this concern, and to counter Schechtman's argument that we are forced to accept the Extreme Claim, we need to deal with the problem of anticipation. The Narrative Self-Constitution view, I argue, does no better than reductionism on this account. We also need to show that this concern is of a different character than the concern for others, because otherwise she can simply claim that it is not self-interested concern and thus does not capture the four features. To approach this, let us look to an example that Schechtman herself uses while

[14] Ibid., 57.
[15] Parfit, *Reasons*, 308.

defending her demand for a conventional linear narrative against the claim of chauvinism: the "Tribal" example.

> At some point, the deviation of an individual's self-conception from the range of narratives standard in our culture can be so great that comprehension of and interaction with such individuals becomes difficult. This is the sort of divergence that can often be found in cases of extreme cultural difference. In such a case the narrative self-constitution view might recognize that this culture has persons, but also note that their concept of persons-and so the persons themselves-are quite different from in our culture. For instance, a tribal culture might assign to an ancestral lineage much of the role that the individual person plays in our culture– responsibility, for instance, may be felt most directly for all of the actions of an ancestral line rather than for the actions of the individual alone, and self-interested and survival concerns may also be primarily attached the lineage. Presumably the members of this culture would also recognize what we call a single person as a natural unit, but this unit would play a different role in their interactions and practices.[16]

Schechtman would still consider these people, even though they have distinct selves spanning multiple bodies across multiple lifetimes. The person here, would thus involve the entire lineage, which she feels means that their concept of a person is different, but that they can still meet her criteria of supporting the four

[16] Schechtman, *Constitution*, 104.

features. Schechtman does not deny that Parfit is correct that we are distinct selves at different times; rather, she feels that we need narrativity to connect these selves as a single subject in order to capture the four features. Although Schechtman uses this example to defend her theory, it may also be used to illuminate why we are not forced to accept the Extreme Claim.

**Why We Are Not Forced to Accept the Extreme Claim**

We may now turn to Galen Strawson. He speaks of people as either episodic or diachronic. Someone who is diachronic sees themselves as existing across time and feels a deep connection to their past, whereas an episodic "has little or no sense that the self that one is was there in the (further) past and will be there in the future, although one is perfectly well aware that one has long-term continuity considered as a whole human being. Episodics are likely to have no particular tendency to see their life in Narrative terms."[17] Further, Galen Strawson thinks that "the heart of Moral responsibility, considered as a psychological phenomenon, is just a sort of instinctive *responsiveness* to things, a responsiveness in the present whose strength or weakness in particular individuals has nothing to do with how Episodic or Diachronic or Narrative or non-Narrative they are."[18] For Strawson, moral responsibility does not depend on whether or not it was the same (transient) self in the past. He claims that he, the present self, feels responsibility for past

[17] Galen Strawson, "Against Narrativity," in *Ratio*. 17.4 (2004): 428-452. Rpt. in *The Self?* Ed. Galen Strawson, (Malden, MA: Blackwell Pub, 2005), 65.
[18] Galen Strawson, "Episodic Ethics," *Philosophy.* 82.320 (2007). Cambridge University Press. Rpt. in *Real Materialism and Other Essays*, Galen Strawson, (Oxford: Clarendon Press, 2008), 220.

actions that he, the present self, did not perform. While Strawson most identifies with the present self, which he claims is very short lived, he also recognizes that as a whole human being he exists across time. People may feel a sense of responsibility for the actions of their family members, or community, etc, even though they did not perform them. This is especially easy to see in the case of children. Parents often feel responsibility for the actions of their child, even though they are fully aware that the child is a distinct person. Strawson claims that in the case of responsibility, there is a "phenomenon of natural transmission" that does not require diachronic self-experience.[19] For example, when a person dies their family members often handle any obligations of the deceased that remain open, including debt, regardless of the fact that they are distinct persons. A person holds himself responsible when he feels this sense of responsibility, even if the present self is not the same self that committed the original action.

Parfit's theory considers a situation that is similar with his Nobel Prize Winner example. He writes, "Suppose that a man aged ninety, one of the few rightful holders of the Nobel Peace Prize, confesses that it was he who, at the age of twenty, injured a policeman in a drunken brawl. Though this was a serious crime, this man may not now deserve to be punished."[20] When considering his accountability, we question his present state, whether and in what way he is similar to the person who did the action. In the case of the Nobel Prize winner, we look to see if the present self is similar in certain ways to the past self, and this is

[19] Ibid., 221.
[20] Parfit, *Reasons*, 326.

relevant to whether or not we hold him responsible. That they may be considered two different people does not preclude us from holding the present person responsible for the past person's actions. Does the present self, the Nobel Prize winner, still attack police officers? Or, does he still have psychological similarities that are relevant to this question? Is he peaceful, does he respect the police and other people in general, does he have a temper, are all relevant questions. Further, these questions affect whether or not he, the Nobel Prize winner, will feel responsible for this action.

Schechtman, however, maintains that qualitative similarity is not enough for responsibility, but this does not seem to be universally the case. We find examples where people feel a sense of responsibility even if they (the present self) did not perform the actions. While Schechtman accepts transference between living bodies in the Tribal example, within the life of a single human, this is not much different. There are multiple selves within the lifetime of one body instead of multiple lifetimes with multiple bodies; if anything, this should be easier for Schechtman to accept than the situation in the Tribal example. The difference is only one of distance and greater known qualitative similarity. In contrast to the above example, one may feel a much stronger sense of responsibility for an action they committed yesterday than for the actions of their ancestors. Here, they know a much greater amount of qualitative similarity holds, and feel themselves to be much more the same person. Even an episodic person may say this. In the case of the Nobel Prize winner, the qualitative similarity may be much weaker, and thus he may feel less responsible, for this is pushing closer to the situation of someone feeling responsible for an ancestor's actions as opposed to feeling responsible for the

actions of yesterday. While this is not the same as it would be if it were the same self, strictly speaking, feeling the responsibility as who did the action, the practical result is not different in a meaningful way; the responsibility, as a feeling, does not necessarily require that it be the same self as the self who did the action.

Schechtman allows that these tribal individuals are people, just different people. They feel responsibility for their ancestor's actions. Schechtman denies that we may feel responsibility for what we, the present self, did not do if we accept the reductionist view, but she will allow this if the conception of a person ranges across multiple bodies, presumably if they are conceptualized in the right way, with narrative. This allowance, however, can be turned around. If a tribal person is allowed to range across multiple bodies and lifetimes, even though Tribal body 1 will not feel the pain of their son, Tribal body 2, they may still have concern for it, and she must accept this in the tribal society for her theory to be coherent. In this case, one may speak of self-interested concern without anticipation, which is inconsistent with her argument for why we are forced to accept the Extreme Claim. She doesn't argue for why her theory allowing the tribal lineages to be people does not apply equally in the case of a single individual with multiple selves in our own society. She merely rejects this possible conception of a person out of hand. Moreover, she gives no argument to justify *her* particular choice for what it means to be a person in our own society. Having considered the views of Parfit and Strawson, it is clear that there are other options for what it means to be a person and these alternative conceptions cannot be ruled out just because they are different. It seems like she must

actually require that these tribal individuals are not people, for they cannot anticipate the actions as their own, or she must expand her concept of anticipation so that if the individual conceptualizes themselves in such a way that they have concern for future persons they take to be themselves (in the non-present self sense), this must be as acceptable as the anticipation she believes she establishes with her narrative self-constitution view. Otherwise, her example is meaningless, and she is open to the charge of chauvinism, for she has no good reason to exclude other possible self-conceptions. And thus, she is wrong to require her self-conception for personhood.

## Bibliography


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

Schechtman, Marya. *The Constitution of Selves*. Ithaca, NY: Cornell University Press, 1996.

Strawson, Galen. "Against Narrativity." *Ratio*. 17.4 (2004): 428-452. Rpt. in *The Self?* Ed. Galen Strawson. Malden, MA: Blackwell Pub, 2005. 63-86. Print.

Strawson, Galen. "Episodic Ethics." *Philosophy*. 82.320 (2007). Cambridge University Press. Rpt. in *Real Materialism and Other Essays.* Galen Strawson. Oxford: Clarendon Press, 2008. 209-231.

# Consciousness and AI: Reformulating the Issue

*Patrick Holzman*

**Abstract** In this paper, I explore the “issue” of consciousness in artificial intelligences, the problem of whether they can be conscious, specifically going for simply asking what consciousness involves, instead of more technical aspects of the field. I use Robert Kirk's concepts of the "Basic Package" as well as "Direct Activity" to outline what being conscious involves, and attempt to apply it to artificially designed and constructed beings. I assume that artificial conscious intelligences will be constructed, eventually; my goal is to suggest a specific and more useful way of thinking about consciousness, which will hopefully accelerate the inevitable.

The science of artificial intelligence deals with attempts to make programs or machines that can function in an intelligent way. What "intelligent" means is dependent on our own judgment and defined for the most part in terms of our own actions. Humans (and animals) act "intelligently," and so when we want to create an artificial intelligence, what we want is something that acts like us, that at least appears to make complex judgments and choices about its environment. Note that I have deliberately phrased this description of AI with phrases like "acts intelligently," or "appears to make judgments," or "functions in a certain way." That is, I've put these goals in terms of what the intelligences do, what their

behavior is, without any mention of their internal structure, and in doing so l leave open the question of consciousness and the "mind."

From my fairly limited understanding of the perspective of those working with AI, this is entirely reasonable. The goal of engineers working in AI is to create something that acts intelligently. The challenge is the execution, the structure of the program, but the goal itself is purely based on behavior. I might even be so bold as to say that many researchers in AI assume that "consciousness," and rational judgment, whatever these involve, will come out in the wash, around when we get things that can truly act like a person. However, I feel that consciousness should be a goal in itself,[1] and the path to consciousness will involve the amplification of some already existing "bare awareness" or "internal life" found in all systems. I will first talk a bit about the field of artificial intelligence, then consciousness in general (that is, attempt to define what I'm talking about in the first place), then introduce Robert Kirk's idea of the "Basic Package" or the "decider," and then finally his concept of "direct activity" and the "Basic Package Plus" to work out how one could judge whether a thing has consciousness or not. Using these, Kirk constructs a model of consciousness, or at least the salient aspects of consciousness in terms of testing for it. I agree with his views, and ultimately I will conclude that the phenomenal aspect of consciousness is not as important as the ability to make judgments and to actually understand the world.

[1] More precisely, I feel that creating an AI that would qualify as one of Kirk's "deciders" is worthwhile as a goal; I think it will be clear why after I describe Kirk's concepts.

The claim "AI researchers don't care about consciousness" is something of a straw man, but I want to dispel even the smallest hint of behaviorism. To be blunt, logical behaviorism, the opinion that all there is to having a mind is acting in a certain way, seems absolutely incomprehensible. In this sort of behaviorism, the statement "he believes that it will rain" is identical with the statement "he carries an umbrella and otherwise acts in certain ways." But this is quite false. Consider a hypothetical table-based system, wherein all conceivable inputs are associated with various outputs: input A at state J causes output X and state K, input B at state M causes output Y and state N, all down a table. Given a long enough table, and a fast enough method to access it, you could have an AI that perfectly replicated a human.[2] However, it seems rather obvious that this would not possess a rational mind, would not analyze the world or make judgments, but instead function purely through reaction.

A behaviorist would say that this is an unfair criticism; such a thing would be impossible to execute. If we were to create a being that acted like a human, and fully like a human, able to react to an indefinite variety of situations, and its hardware was limited to something the size of a human head, then it seems reasonable to assume that such a thing would likely be acting in a complex, rich way, actually having an internal functioning of similar convolution to ours, if likely with a different sort of structure. Phrased this way, behaviorism is much more about practical judgments about the nature of things we could encounter or build. This still misses the

[2] I do mean, though, an *extremely* long table, with a great many states and inputs. Essentially the false human's entire life story would count as a single state.

point behind asking whether something really possesses a mind or is conscious—unless there is something beyond the material, possession of a mind must line up with some physical state of things. Furthermore, advances in technology may allow us to make astonishingly complex AIs that, nonetheless, will have no true mind.

The goal at the moment is to make programs that solve problems humans are still not very good at, such as traffic control or chess. Generally this is done by formalizing the situation mathematically, then writing a program to manipulate this formalization and find what best fits a certain criteria. It is not that "is efficient" is the criterion, but rather that there is some variable in the formalization that the program attempts to minimize or maximize. Once the formalization is "translated" back into our own understanding of the problem, this variable is identified with efficiency, but a significant part of the work when making the artificial intelligence is this formalization, in determining how best to abstractly represent the problem. Even when researchers attempt to create AIs that learn through something like a "neural net," they must first create a domain within which the AI will function; the problem has changed from solving a certain problem in a certain language, to working out the language and what problem is involved while still using a certain other language.

Here I want to begin to use words like "syntax" and "semantics," but I think doing so would be dangerous—such words have been used many times before and have vague definitions.[3] It

[3] I also suspect the way I use these words, or at least what I consider important about them, is different from many others'.

might be best to carefully lay out what I'm talking about. When I refer to consciousness, I am to a certain extent going by Nagel's idea of having a "what it's like."[4] My eventual conclusion is, however, that the phenomenal aspect of consciousness, this feel, is not what is valuable. Rather, the value comes in the ability to make judgments about the world; whether our perception is immediate or has no phenomenological aspect is irrelevant. The distinction I'm trying to focus on is one between "consciousness" and "awareness," between "perception" and "sensation." Awareness, sensation, is simply having a first-person perspective from which certain things are experienced, while consciousness, perception, is to have some context, some interpretation of that sensation. Perception is sensation with internal context; consciousness is awareness with actual *meaningful* content. This is a very fuzzy distinction, one that will be better distinguished when I get to Kirk's deciders and the basic package. One possible way to think of it is by the concept of "raw feels," which here would just be sensation. The "raw feels," the sensations, are the raw bits of context-less information that comes into a system, which for some things is then interpreted and becomes perception, becomes conscious. For those things which are not conscious, sensation cause some reflexes to fire, and in this manner they are yet aware.

**Terms and Assumptions**

In earlier versions of this paper I freely used "consciousness" when what I meant was "a mind," in this sense of

[4] Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review.* 83 (4): 435-450.

"rational and analyzing" that I've been stressing. To jump ahead, the distinction between an unconscious and a conscious mind is, in Kirk's terms, the addition of "direct activity," the fact that for a conscious mind perception is irresistible and happens automatically. Perception of the world directly affects the mind, changing how that being will achieve its goals or what its goals are, with no need to reflect on its body of knowledge. This “direct activity” is what is entailed by consciousness, the “what it’s like.” The actual rationality should not be properly referred to as “consciousness,” except in that explicit sense of rationality. That is, we are not discussing “conscious vs. unconscious,” but rather “conscious vs. reflexive” or something along those lines.

By "system" I really do mean any sort of system. For the most part I'm talking about complex life-forms, but computer programs, robots, even things like toasters or thermostats count as a "system." The "basic package" and "deciders" will be detailed in more depth later, but essentially a "decider" is something that analyzes information about its environment, forms goals, and then executes those goals. I will say that systems that are deciders have "minds," and "mind" here means "rational, complex mind." "Consciousness" refers to minds or deciders that have direct activity, Kirk's "basic package plus." Unfortunately, I do not have a simple term for systems that are not deciders that still have direct activity. I think they could be called "non-rational sensing systems."

I will go ahead and assume there is nothing beyond our physical bodies at work when we speak of the mind. Our brains do things, and this activity, from a different perspective, is called “mind.” Brain activity does not “produce” minds over and above

the brain, they is not “caused” by that activity. Rather, the activity in the brain somehow is the activity in the mind. You could not have a human brain functioning the way it does without also having a mind. This is not a contingent fact, but rather a fact about the sort of activity that occurs in the brain, that it is also conscious mental activity, when viewed from the inside. I take this as a matter of faith, and feel no need to defend this. It seems for the most part obvious, and, honestly, not that interesting.

I will also assume that consciousness is an interesting and worthy topic of discussion. There is something that it is to be conscious; you and I can feel it just by thinking. This needs to be acknowledged, and explored. Finally, I will also assume that the mind could best be described as "activity within the brain when viewed from a different perspective." Searle gives this formulation: “Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain.”[5] I would agree, with a caveat about the exact use of language. I want to stress that the mind is not “caused” by the brain, but is the brain; I think I mean the same as Searle, but that I am insisting on a certain language. Searle talks about the mind being an emergent property of the brain, in the same way that wetness is an emergent property of H2O. Now, is wetness “caused” by the H2O? Not exactly, not in the same sense that a rock causes a window to shatter. H2O does not “produce” wetness, but rather it is wet, in sufficient quantities. “Produce” and “cause” evoke to me feelings of “extrude” and “impart,” not “possess.” However, if we are to say that the brain “causes” the mind in the same way that H2O “causes” wetness,

[5] John R. Searle, *The Rediscovery of the Mind* (Cambridge: MIT, 1992, 1.

then it's quite fine. Indeed, the other half of his formulation is that mental phenomena "are themselves features of the brain." Although I've focused on just this example, I think similar sorts of situations abound and are what actually make up many of the apparent differences among various theories.

What is tenuous, and interesting, is the use of "from a different perspective," when saying what the mind is. An objection can be raised that this physicalist explanation does not account for all aspects of the mental, that there is an "explanatory gap." Nagel's Bat[6] and Jackson's Mary[7] are paradigmatic thought experiments/arguments for this "internal perspective." The explanatory gap implies that knowledge of the physical world will not give you the knowledge of "what it's like." However, for now I am not focused so much on *what* it is like to be something, but rather *whether* there is a "what it's like" for any specific thing. I suspect there is a "what it's like" to be a cat, and that there is not a "what it's like" to be a rock, and furthermore that we can tell this empirically, and where the line is, just from physical facts. Even if physical facts can't tell *what* it is like, they can still tell *whether* there is a "what it's like." It is interesting to pursue whether we can tell "what it's like" to be something, and I will do so, somewhat, but the difficulties we have in doing so do not change our knowledge that there is a "what it's like" to be something.

[6] Thomas Nagel, "What is it like to be a bat?" (*Philosophical Review,* 1974).
[7] Frank Jackson, "What Mary didn't know" (*Journal of Philosophy,* 1986). I'm going to assume some familiarity with both of these.

**Brain/Mind Identity**

Before I get much farther, it may be valuable for me to clarify my position, especially towards mind/brain identity. The Stanford Encyclopedia of Philosophy states "The identity theory of mind holds that states and processes of the mind are identical to states and processes of the brain. [. . .] Consider an experience of pain, or of seeing something, or of having a mental image. The identity theory of mind is to the effect that these experiences just *are* brain processes, not merely *correlated with* brain processes."[8] On my view, there are two ways you can interpret this in terms of AIs possessing minds. One way is to say that, obviously, an AI cannot have a mind, as minds are identical to brains and thus non-brain possessing AIs will not possess minds. The other way, my way, is to say that AIs can quite easily possess minds, just minds that are very unlike our own, as their brains are unlike our own, *in structure*. What is the mind in the brain is the brain's structure—the mind is not a non-physical object whose parts can be identified with the parts of the physical object of the brain, but rather the organizational relations of the mind are identified with the relations in the brain. The mind is already nothing more than a set of relations; what the mind is identical with in the brain is those relations of the parts of the brain.

The sort of identity theory I agree with is an odd sort of token-token identity. A token of some activity in the brain is identical with a token of some activity of the mind. Types of tokens in the mind are defined in terms of behavior and similar

[8] J. J. C. Smart, "The Identity Theory of Mind" (*The Stanford Encyclopedia of Philosophy*, 2008).

phenomenological characteristics, and those tokens in the brain have similarities as well, but the link of types of mental tokens to types of physical tokens are on account of the linkage of the tokens themselves. "Pain is identical to c-fiber firing" does not mean that a being with no c-fibers cannot feel pain. Rather, individual tokens of pain in humans are found to link to tokens of c-fiber firing, but the link between pain and c-fiber firing *in general* is only on account of the commonality of the tokens of pain. In a being without c-fibers that still feels pain, such as a robot, we could say "pain is identical to a red wire firing," or whatever the case is.

Obviously this leaves the question of whether the pain in the robot is the same as the pain in us. I feel it is not, unless we've made an effort to make a robot with the same physical and mental structure as us, but I still feel that it is reasonable to say it has pain, as long as it has connections to its physical body that induce unpleasant sensations in it and that serve a similar role as pain in us.; "unpleasant" will be dependent on whatever reward mechanisms we design it to have. If a robot is has a mind, has a way of forming goals, some of which include the preservation of itself, has ways of gathering information about damage to its body, and has some sort of unavoidable phenomenological sensation that carries this information to its mind which encourages it to avoid that damage, then it has sensations that can be usefully called "pain." It may not be pain *like ours*, but it is no *less* pain that ours is.

**Kirk's Basic Package**[9]

In order to deal with things like consciousness, or the mind, or the idea of an "internal perspective," you need fairly strong definitions, or at least reasonably clear guidelines of what will constitute such things. Instead of trying to form yet another new framework, I've decided to use Robert Kirk's ideas of the "basic package," and "deciders," as well as his "direct activity," because the entire system seems to be the most reasonable and acceptable one I've read yet. A decider is something that makes judgments about the world, analyzes it, and forms goals and what it sees as the most appropriate paths to those goals. This is in contrast with systems that act purely on reflex, and Kirk uses this contrast extensively to lay out what he means by a "decider." An example of a general reflex system would be a clam, which shuts its shell when exposed to certain sensations. It is important to remember the distinction I tried to make between consciousness and a "mind"—a conscious mind is a mind with this direct activity, but a system does not need to have a full rational mind to have direct activity. A clam still has sensation, and an internal perspective, despite not having the full, rich consciousness it would possess with the basic package. What Kirk focuses on is "perception," which refers to sensation in a system that can learn, and which is an integrated part of a conscious mind. Fully conscious systems are partially defined by perceiving their environment and learning

[9] Kirk uses the concept of the "basic package" extensively. It is developed through chapter 6, and put forward on p. 89-96, and throughout the rest. The concepts of various sorts of reflex systems, and deciders, are developed first, through p. 77-89. The discussion of sensation and consciousness is from p. 58-61, as well as p. 92-94.

from their perceptions; similarly, perception is only perception in the sense that the system can do something consciously with the information, instead of having the sensation just cause a reflex.

By Kirk's view, there is a succession of increasingly rich reflex systems. Initially, there is the "pure reflex system," such as a clam. These are systems with hardwired responses to stimuli, which are genetic (for biological systems) and cannot be altered by the system itself, nor are they designed to be altered by the external world. The "road to the decider" is not simply a matter of increasing complexity—a complex organism like an oyster is just as much a pure reflex system as a protozoon, although biological organisms with greater complexity are generally partially that way to allow for more complex responses. There are then "pure reflex systems with acquired stimuli," where there is a slight amount of room for new responses to develop, and "built in triggered reflex systems," wherein certain stimuli open up subsections of the list of responses, which themselves otherwise stay inactive. Finally, just before we cross the threshold into the deciders, are "triggered reflex systems with acquired conditions." Kirk's example is the dragonfly, which learns to have a specific nest, but for whom that learning process is automatically set up to happen. That is, the dragonfly does not decide "this is where I'll set my perch," but rather certain conditions cause the variable "perch" to get permanently filled in, which then gets plugged into the triggered reflex system.

The threshold between this and the decider is the capacity of "monitoring and controlling the responses," and is the important part Kirk as emphasizes:

> We have reached a highly significant watershed. For a system to monitor and modify its own behaviour involves a major break with the reflex pattern. Monitoring and modifying must involve not only the organism's being able to perceive its own behaviour, or at least the effects of its behaviour on its environment, but also to adjust its behaviour in ways appropriate to its goals. That requires it to be able to control its own behaviour on the basis of its information, in a way that none of the types of systems so far considered is capable of. [. . .] It seems probable that what we can conveniently refer to as 'monitoring', modifying', and 'controlling' are highly complex processes, capable of being realized to a greater or lesser degree, at different levels of organization in the system as a whole, and in an indefinitely wide range of possible internal structural patterns.[10]

He further says that what is important is the *integration* of all these processes. There is no requirement of how these processes must be executed, just that there are capabilities. To be a decider, to have the "basic package," is for something to be able to—

(i) *Initiate and control* its own behavior on the basis of incoming and retained information: information that it can use;
(ii) *Acquire and retain information* about its environment;
(iii) *Interpret* information;
(iv) *Assess its situation*;
(v) *Choose* between alternative courses of action on the basis

[10] Robert Kirk, *Zombies and Consciousness* (New York: Oxford, 2005), 87.

of retained and incoming information (equivalently, it can *decide* on a particular course of action); and

(vi) *Have goals.*

Moreover, all of these must be unified and integrated.

It's possible that a thing could have faculties similar to some of these, but to have these fully they must be all present and interrelated. Put another way, it makes no sense to talk of "goals" without something being able to acquire and interpret information, or to choose between various actions, nor does it make sense to talk about controlling behavior unless a thing has goals, or interpreting or assessing information unless it's going to be put to a use, to a choice. A thing can "sort of" interpret information, a thermometer for example, but it will not be doing so for itself. This again has a great deal to do with perception, which is just sensation that conveys information, that a decider can then act upon. Sheer sensation, experience, can be found in the simple reflex systems, without there being any understanding or perception, despite there often being some apparently intelligent reaction. This relates back to Kirk's definition of perception—sensation that conveys information, that a decider can then act upon. Sheer sensation, experience, can be found in the simple reflex systems, without there being any understanding or perception, despite there being reaction, often seemingly intelligent reaction.

Bringing this back to the subject of artificial intelligence, what we deal with when we have seemingly intelligent systems is instead this very bare pure reflex system. Kirk will freely admit that he does not know enough of the subject of animal neurology to give clear examples of each sort of reflex system. Similarly, I will say that I am not sufficiently familiar with the programming of AI

to say what sort of system any one example is. However, by my earlier outline of an AI, what we currently have is often still a very simple sort of reflex system. Even the "learning systems" are likely only so-called "triggered reflex systems with acquired conditions," where certain approaches to learning are acquired, but are still within the reflexive framework set up beforehand by the programmer. It is entirely possible, though, that I am wrong here, and that what is causing me to hesitate is something else.

**Direct Activity**[11]

In Kirk's view, the basic package is not sufficient for phenomenal consciousness. What is also needed is "direct activity," or the direct action of sensation on the creature's decision-making process. We all experience direct activity, when any sort of sensation comes our way, because we cannot help but sense it. Initially it's difficult to even understand what Kirk means by direct activity, because it's unclear what the alternative would be. The simplest example of information gained indirectly is subliminal information—when we do sense something, and file it away somehow, but do not notice it and actually perceive it at the time. The information has been acquired, and can be used to alter our goals or our methods, but in order to do so we must indirectly access them after the fact. Kirk stresses instantaneity and priority in direct activity. The perceptual information is instantly available to an organism, and it also holds priority, immediately changing our goals and choices about the world.

[11] Another important concept, direct activity is detailed in Chapter 9 of *Zombies*, pp. 140-163.

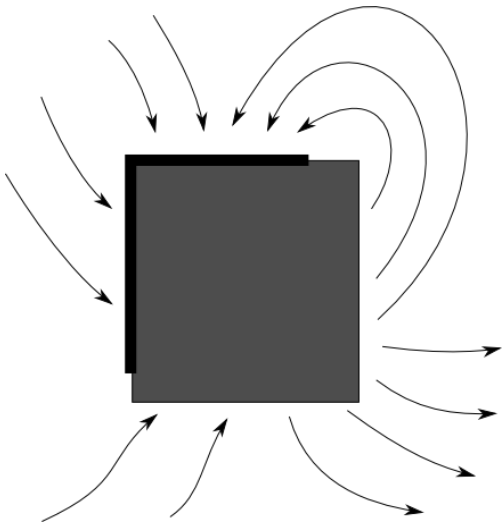
He uses what he calls a "rabbitoid" as an example,[12] stating that a "rabbitoid" is like a rabbit in all ways, except that sensory information does not act on it directly, but through some other method. It is difficult to imagine how this would work, but a possibility would be that the rabbitoid constantly queries its store of knowledge. When a fox comes up from behind a hill, the rabbitoid notices a second later during its regular "scan" of its knowledge base, and then bounds away, relying on its stored model of the environment to navigate. A conscious rabbit has an advantage over a non-conscious rabbitoid in that it will automatically notice changes in its environment, and will be able to alter its immediate goals accordingly, while the rabbitoid would always have some sort of delay in action. The very best that a rabbitoid could do, would be to constantly re-scan its knowledge multiple times per second. This distinction still holds if you assume that rabbits do not possess full rational minds; the reflex system possessed by a rabbitoid would still function better if information about the environment directly affected its system instead of it needing to constantly retrieve stored information about the world.

**The Red Herring of Thought**

I want to interject a bit about conscious *thought*, and then about bats, before returning to consciousness. Thought is often considered a very important aspect of being human, and seems conflated with consciousness itself. But what happens when we think? One might say, we become aware of what's going on in our

[12] Ibid. pp. 142.

mind, that we "look inside." I think there's a problem with this, that our everyday sensation of thought and introspection is too naïve, and problematic. Imagine this extremely simplified and kind of silly picture of the mind, gained (of course) from introspection: there is a sort of "black box," in which all mental activity occurs. This box takes certain inputs, many of which are "conscious," although some are not, and produces various outputs. These outputs include motion, activity, and speech, but (here is the point) also include "thought," which is nothing more than aborted speech and self-produced sensation, re-routed back into the box. On only part of the "edge of the box" is the "membrane of consciousness." In terms of this metaphor, things are conscious only as a result of passing through this membrane. Our knowledge of our mental activity is known only so far as we produce thoughts that are then reintroduced into consciousness. The activity within the black box is completely unknowable, and can only be inferred from the thoughts produced.



This is a very flawed picture. Consciousness is not a membrane, there is not a line when things "become conscious" in the brain. However, the salient point is that introspection is not directly accessing or monitoring our mental processes. Instead, thought is output that is reintroduced back into the system. This is

also flawed in that the reintroduction likely does not happen all at the same level—it would be better to imagine the circular arrows happening within the box, making loops of various sizes. But again, the point is that thought is activity re-routed, not the brain actively looking at itself.

This seems like it would be efficient, more than growing some specialized "introspection" capability. If the brain has taken the time to create systems dedicated to processing, say, language, it makes sense that when we think in terms of language we simply route the output of our thoughts, as if we were speaking, back into the language processing bits, using the same hardware we'd use if we were hearing, instead of developing a new system to "monitor" our thoughts. Similarly, at earlier stages, our ability to remember and imagine things significantly overlaps with our capacity to sense things, and so it seems reasonable that instead of developing a new "imagination" capacity, we rather develop the ability to stimulate those systems dedicated to dealing with perception. This also explains the sensation of thought, why it actually has a "sound," instead of just being abstract activity.

**So, What is it Like to be a Bat?**

I assume a bat has sensation, and also consciousness. What I mean is, there is something it is "like" to be a bat. Perhaps it doesn't have active thought, but it makes decisions, and its actions are complex and nuanced, reasoned. Nagel asked what it is like to be a bat; he, and others, concluded that we cannot know, that the life of a bat is fundamentally alien to us.

But at least attempt to imagine what being a bat is like. The problem, initially, and as Nagel stresses, seems to be echolocation,

something we have no real analogue for. But this doesn't seem entirely impossible, merely very difficult. Try this:

*Close your eyes (perhaps read the instructions first, or, imagine closing your eyes). You can still hear, can you not? With your eyes closed, drop something like a book to your desk, and notice how you intuit its position from your impression of the sound. If you were to reach out, you could grasp it with some difficulty. If it were to make noise constantly, you could grasp it with near ease. A sharp sound to your left will give you an impression of "something" there. With your eyes closed, a man walking around a room, or a floor above with a thin ceiling, will give you an impression of motion, of presence. Focus on that impression of presence, separating it from the sensation of the sound itself.*

*Now, with your eyes closed, feel out your surroundings. You can tell that this is a box, or that is a sphere. You can feel the dimensions of your desk, and you have an almost visual experience of this the size and shape of things. These sensations can be deceptive (how large are your teeth, when sensed with your tongue, and then when felt with your fingers?), but that is not surprising.*

*Imagine the sensation you experience when a noise is heard, the sense of location and position. Isolate the feeling of position, the feeling of "a presence," from the sensation of the noise itself, that it is a noise. Focus on the feeling of position and presence. Now, imagine the sensation of feeling the shape of an object, and isolate the impression of the form and size from the feeling of touch itself. Merge those feelings of position, as if you were*

*experiencing a thing's form through a constant torrent of sound, where the sensation of any individual sound was drowned out by the ubiquity of the torrent, leaving only the feeling of position, form, size, and distance. What is there is not the sound, but the almost eerie sense of "something there," the odd itching at your back, like the feeling of being watched without experiencing the watcher.*

*Pretend you were blind, and had to live off of touch and sound to navigate, but then were able to somehow merge the impression of presence you get from sound, having the sounds themselves fade into the background, and then were to combine this with the feeling of form and shape gained through touch, having the feeling of touch itself be replaced with that background noise, extended to the range of your hearing. You would reach out constantly, as if touching through sound.*

Nagel would say that this is what it would be like for a human to be a bat (and even then, only barely), and would press the point, asking what is it like for a *bat* to be a bat. A human has its own beliefs, desires, goals, and so on, and to imagine what a bat's internal life is like is impossible since these will always interfere with our attempts. However, I feel that you can run into the same sorts of problems with asking a question as apparently simple as "what is it like to be yourself?"

First ask, what *was* it like to be yourself? Imagine yourself ten years ago, or even a day ago. How do you do this? Well, you extrapolate. I myself at this moment a day ago was bumbling about, taking a shower, not really interested in anything, assuming I'd wake up a little in an hour or so and figure out what to do then.

Right now I'm coming off of a long, caffeine-fueled writing session. Ten years ago I would be napping on an hour-long bus ride; if I was awake at this moment ten years ago, I'd have been just woken up for some reason. Certainly, both of these involve being fairly groggy and sleepy, and to that extent I can imagine what it would have been like. However, the phenomenal quality of each experience is very different. The sleepiness I feel now is very different from ten years ago, and it is difficult to evoke that feeling in myself because to do so I would need to overwrite my current feeling. I cannot remember what it feels like, I can only extrapolate, evoke the feeling.

But is this only because sleepiness is a muddled, vague feeling? Consider pains. When I was younger, I stubbed my toe. I've done so many times over the years, in fact. And yet can I accurately remember what it felt like? No, only that there was an accompanying feeling of suffering. If anything, what I remember is the suffering, not the pain itself, and even that suffering is extrapolated. How I related to pain then is much different than how I relate to it now. What I feel when I imagine that pain is not what it was like for past me to feel pain, but what it would be like for present me to feel past me's pain, *and only poorly*. How different is this from trying to think what it's like to be a bat? Not impossibly so—and it is not a matter of kind, but of degree. It is much easier to imagine what it was like to be me feeling pain than what it's like to be a bat; but neither is perfect.

What if I asked, what is it like to be you, a second ago? No no no, that's silly, surely. But pinch yourself. Ow. What was it like? Well… it hurt, yes, but can you evoke that sensation again? Not really. You can recall the suffering, and what the pain was sort

of like, and how it still hurts a little now, but none of that is what it was like to be you a second ago, feeling that pain. So what is it like to be you, right now? Anytime you try to focus on that, you can only evoke the feelings a second later. What it is like to be you is constantly slipping away. You can only experience "what it's like" to be anything, namely you, as it is experienced. To actually feel what it's like, you need to have the feeling at the moment. This also somewhat makes sense evolutionarily—why would we go through incredible effort and cost to repeat pleasurable actions if we could merely evoke the pleasure in our minds on command?

That we cannot imagine what it is like to be a bat is not so surprising when we cannot even imagine what it is like to be ourselves. And yet this does not tell us that there was *nothing* that it was like to be ourselves, and it does not tell us that there is *nothing* that it is like to be a bat, and this says nothing about telling whether there is a "what it's like" for any entity through physical observation.[13] Our memories are not just "not as vivid," but entirely false, constructed. We cannot know the past "what it's like," or others' "what it's like"-s, just as we cannot know the "what it's like" for a bat—but we would not deny consciousness to our past selves, or to other people.

**What is it Like to be a Thermostat?**

To ascribe emotions, desires, or beliefs to a thermostat is silly. When I say that a thermostat has experience or sensation, I do not mean anything approaching our own experience. As Kirk

[13] A significant amount of this is paraphrased from Kirk (2005), ch. 5, especially p. 61-68.

would say, we are deciders, we interpret information and make decisions based on that information according to goals. A thermostat whose setting aligns with the ambient temperature does not "feel content." A thermostat set to a higher temperature does not "desire to make things hotter." A thermostat does not "believe maintaining the temperature is good." A thermostat does not *perceive*, because it does not interpret its sensations, does not work with information. Emotions, desires, and beliefs are fantastically complex and important aspects of our experience. Some day we will make a machine that does experience emotion and desire, and have beliefs, but it will be no time soon[14]. This is a reasonable and worthy goal, but it is important to realize how difficult it will be. So if we cannot say that a thermostat "desires to make things hotter," in what sense does it have an internal experience?

When I ask "what is it like to be a thermostat," I'm speaking of something that it is very, very difficult to imagine. It is hard enough to imagine what being a dog is like; harder still to imagine the life of a slug, and of a bacterium; so when we get down to something as bare and simple as a thermostat, we are truly a long ways away from our own experience. It is not even enough to try to sense things thoughtlessly, as the sensation of a thermostat is nothing like ours in any way. A thermostat is simpler even than an individual neuron.

All I mean is that the thermostat "senses." It senses the temperature the same way a protozoon senses light levels and moves accordingly, or the same way a bacterium senses a certain

[14] No, I don't have support for this, but I consider it the same sort of statement as "someday we will colonize other planets." Barring something horrible happening, or the discovery of some extreme limiting factor, it seems so.

chemical in its environment and stops dividing[15]—really it must be far more simple than that, but it is the same sort of "basic reflex" as a clam closing its shell in response to certain stimuli, or a slug retracting a feeler when it touches something rough. In the same way that animals have moved from those basic reactions to our own, we should attempt a similar project to move from the thermostat (well maybe something else) to a conscious being like us. What this requires is a move from the "reflex system with acquired conditions" to the actual "decider."

**Conclusion: Does "Consciousness" Matter When Thinking of Artificial Intelligence?**

It depends on what the connotations of "consciousness" are, which brings us back to Kirk's direct activity. If the difference is between having direct activity or not, between being a rabbit or a rabbitoid, it seems in fact that consciousness is of no importance, and the focus on "what it's like" is missing the point. If, instead, "consciousness" is taken to deal with the difference between sensation and perception, between acting on reflex, or making judgments, having goals, and so on, then it is obviously of high value. A system that can actually analyze the world and make judgments will have an advantage over something that acts on predefined rules, assuming it is meant to deal with the sorts of

[15] Certain protozoa sense light, and then move their flagella to move toward it, but only when it is fairly mild; bright, constant light has no effect. Colonies of certain bacteria maintain a size by having each bacterium secrete a chemical, and then stop division when the chemical reaches a certain concentration, which lines up with a certain population.

complex and variable situations that humans and other animals can handle.

In other words, we should not be asking whether computers will be conscious; that is a matter of how they relate to information. What matters is how they process it, and the incidental aspects of consciousness (the instantaneity, the priority) should not be taken as essential to having a mind.

## Bibliography

Churchland, Patricia S., Sejnowski, Terrence J. "Neural Representation and Neural Computation." *Mind and Cognition*, ed. W. G. Lycan (1990, 1994). Cambridge:Blackwell. pp. 224-252. (From *Neural Connections, Mental Computations*, ed L. Nadel et all, MIT Press (1989))

Jackson, Frank. "What Mary didn't know." *Journal of Philosophy* (83): 291–295. 1986.

Kirk, Robert. *Zombies and Consciousness.* New York: Oxford, 2005.

Lycan, William G. *Consciousness and Experience.* Cambridge: MIT, 1996.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review.* 83 (4): 435-450.

Searle, John R. "Can Computers Think." *Philosophy of Mind.* ed. D. Chalmers (2002) pp. 669-675. New York: Oxford, 1983.

Searle, John R. *The Rediscovery of the Mind.* Cambridge: MIT, 1992.

Smart, J. J. C. "The Identity Theory of Mind." *The Stanford Encyclopedia of Philosophy*, 2008.ed. Edward N. Zalta. http://plato.stanford.edu/archives/fall2008/entries/mind-identity/

# Could Consciousness Emerge from a Machine Language?

*Genevieve H. Kaess*

**Abstract** Behaviorists believe the following: if the output of artificial intelligence could pass for human behavior, AI must be treated as if it produces consciousness. I will argue that this is not necessarily so. Behaviorism might be useful in the short term, since we do not know what causes consciousness, but in the long term it embodies an unnecessary hopelessness. I will attempt to establish in this essay that certain empirical knowledge of consciousness is within the realm of possibility. I will then use my own definition of certain knowledge to shed light on ways in which computer programming falls short of producing human-like consciousness.

## I. Introduction

"The best reason for believing that robots might someday become conscious is that we human beings are conscious, and we are a *sort* of robot ourselves."[1] Daniel Dennett's offhand introduction to his essay "Consciousness in Human and Robot Minds" serves more generally as a summary of popular contemporary philosophical thought regarding artificial intelligence: it is possible, in theory, because human intelligence is

[1]Daniel C. Dennett, "Consciousness in human and robot minds," in *Cognition, Computation & Consciousness*, ed. Masao Io, Yasushi Miyashitatt and Edmund T. Rolls (Oxford: Oxford University Press, 1997), 17.

possible. Human life, and consciousness with it, is no more than the machinery of nature. What remains unclear is to what degree (if at all) and in what ways the mechanisms that produce human consciousness must be imitated in order to create artificial consciousness, and whether knowledge of the creation of artificial consciousness can ever be certain.

In this paper, I will argue that syntactical computer modeling is not sufficient for artificial consciousness. I will approach this point by first examining views of philosophers (specifically Alan Turing and Hilary Putnam) who have suggested behaviorism as the standard by which to judge consciousness in artificial life. I will suggest that although behaviorism provides an immediate solution to the problem of other minds, the adoption of behaviorism as a long-term solution embodies an unnecessary hopelessness regarding certain knowledge of consciousness. Applying the same standards for certain knowledge that we do for other phenomena, we can come to certain empirical knowledge of the causation of consciousness. The rejection of this claim, I will argue, is dualistic. Finally, using the standards that have traditionally been sufficient for certain knowledge, I will explain why one specific example (which I will discuss in section V) casts doubt on the claim that AI can be achieved through computer programming.

**II. Definitions**

For simplicity's sake, the term "artificial intelligence" (AI) will refer, in this paper, to artificial consciousness. Traditionally, consciousness has been deemed an unnecessary (or at least not necessarily necessary) condition for artificial intelligence. On the

contrary, I believe intelligence and consciousness to be inextricably linked. Intelligence, by definition, is the capacity to learn and understand[2]; understanding is a feature of consciousness. Information processing, then, can only be qualified as intelligence if it has conscious manifestation. Consciousness will be understood (in this paper) as thoughts and emotions such as humans experience them. I exclude non-human animal consciousness from my definition because the goal of AI scientists is to produce human-like intelligence (which, by my definition, entails human-like consciousness). By limiting the scope of the definition of AI in this way, a more comprehensible argument will emerge; current knowledge of the nature of consciousness in other organisms is imperfect, and any discussion of it would be based on conjecture.

The science of AI depends on the truth of one basic assumption: consciousness is a natural physical process. There is no spiritual realm of thought that exists separately from nature; therefore, provided limitless resources and a thorough understanding of the mind, we would be able to reproduce it artificially. *Computational* AI depends on the possibility that this can be realized using computer programming. In this paper, I will assume that AI is possible, but I will provide evidence that *computational* AI is not. Henceforth, "AI" will refer to computational artificial consciousness, and "computational

[2]This definitiveness of this definition is disputable. However, there is no doubt that this is one commonly used definition of "intelligence." Since I am merely using it to justify my choice to define AI in the way that I do, and not as a premise to any of my arguments, the definitiveness of my chosen definition is of little consequence.

functionalism" will be understood as the philosophical position that such AI is obtainable.

To say that the creation of artificial intelligence can be fully realized through computer programming is tantamount to saying one of two things: (1) the human mind is itself nothing more than a computer[3] – an information processing tool – or (2) computer programming and the mind can produce equivalent cognition without holding any additional features in common. I will discuss the second possibility in section VI. For the most part, computational functionalists hold that the first is true: information processing is the necessary feature of the mind. Certainly it is true that the human brain has a biological medium distinct from that of a computer, but that is all it is: a medium that realizes and supports the brain's intrinsic informational processing. Human consciousness, they believe, is a feature of the *processes*, not of the medium.

Computer programming, at its most basic level, is a series of 0s and 1s, which answer the question of whether or not various features exist. I will refer to these 0s and 1s as "computer syntax." Computer syntax is itself a mechanical feature of the computer, which is programmed in by humans. When prompted, it sets in motion a series of mechanical events within the computer that lead to the visible output on the screen or, in the case of AI, the observable actions of a robot. The 1s and 0s can be combined in very complex ways to produce impressive outcomes. In the 1950's, the research of Allen Newell and Herbert Simon suggested

[3]John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).

that "a computer's strings of bits could be made to stand for anything, including features of the real world, and that its programs could be used as rules for relating these features."[4]

The idea for AI was not born solely of the impressive capabilities of computers. It emerged also from the notion that computer programming is the best model for the workings of the brain. Most neurons give and receive signals in short blasts. They operate under an all or nothing principle - either they're firing or they're not. This is similar to the 1/0 duality of binary code. AI scientists posited that these neuronal impulses could be modeled by computer programming to the same effect: intelligence.

**III. The Problem of Other Minds**

But how would we know if that happened? Current scientific knowledge does not account for consciousness. This is called the "problem of other minds," and it is the foundation, as well as the limiting factor, for philosophical arguments regarding AI: we do not know what exactly consciousness is, and therefore we cannot test for it in others. One can only be certain of one's own consciousness. For some philosophers, this is grounds for suggesting the adoption of a behavioral standard by which we might judge what constitutes intelligence and what does not.

In his article "Computing Machinery and Intelligence," Alan Turing described his most lasting contribution to philosophy – the "Turing test." Turing devised a game in which two people (a man – "A" – and a woman – "B") sit in separate rooms as an

[4] Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1997), x.

interrogator questions them. All identifying features are hidden from the interrogator. His goal is to determine which is the man and which the woman; the goal of one of the two competitors is to confuse the interrogator and the goal of the other is to help him. Turing then posed the question: "What will happen when a machine takes the part of A in this game?"[5] The interrogator now must determine which of the two is the machine. Turing asserted that if a machine could win this game as frequently as the typical human, it would be unfair to deny that it had consciousness. After all, we do not require proof of consciousness in one another. Until consciousness is de-mystified, Turing believed, we must adopt this principle of equity.

Although the Turing test is not a definitive test for consciousness, many have accepted it as the standard. We do not have the knowledge to recognize consciousness in others; therefore we are engaging in cognitive chauvinism if we suggest that a machine with humanlike cognitive capabilities (insofar as they are measurable) lacks consciousness. Turing's solution is pragmatic: to avoid prejudice, we must judge consciousness in non-humans in the same way we do in humans – behaviorally.[6] The strength of his position is that it is safe; it makes no conclusive claim about what constitutes consciousness, but instead suggests the adoption

[5] A.M. Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 434.

[6] I would argue that we do not always usually use behavioral characteristics to determine whether other humans are conscious. Instead, we assume that they are conscious (because of their biological status as humans) regardless of whether or not they could pass the Turing test. However, I will grant Turing this point, since it is probably true that the reason we assume humans have consciousness, even if they cannot pass a Turing test, is because as a general rule, humans behave as if they are conscious.

of a standard.

Hilary Putnam expressed slightly stronger opinions in his essay, "Robots: Machines or Artificially Created Life": we cannot expect to gain complete understanding of psychological states by studying brain physiology. "Psychological laws are only statistical … to say that a man and a robot have the same 'psychology' … is to say that the behavior of the two species is most simply and revealingly analyzed at the psychological level (in abstraction from the details of the internal physical structure), in terms of the *same* 'psychological states' and the same hypothetical parameters."[7] For example, anger is defined by one's claims and actions, not physical brain states. It is identified by behavioral features, not biological ones. This being the case, Putnam contended that "it is … necessary … that one be prepared to accept first-person statements by other members of one's linguistic community involving these predicates, at least when there is no *special* reason to distrust them."[8]

Putnam constructed the following scenario to illustrate his point: suppose that sometime in the future the robots we have invented build robots of their own (Putnam calls these "ROBOTS"). The philosopher robots then sit around debating whether or not ROBOTS have consciousness. This is akin to our current actions. Since we do not understand consciousness, we have no less duty to ascribe consciousness to robots than we do to one another. The question of consciousness, Putnam concludes, cannot currently be solved. Whether robots should be treated as if

[7] Hilary Putnam, "Robots: Machines or Artificially Created Life," *The Journal of Philosophy* 61, no. 21 (1964): 677.
[8] Putnam, *Robots*, 684.

they have consciousness, then, "calls for a decision and not for a discovery. If we are to make a decision, it seems preferable … to extend our concept so that robots *are* conscious – for 'discrimination' based on the 'softness' or 'hardness' of the body parts of a synthetic 'organism' seems as silly as discriminatory treatment of humans on the basis of skin color."[9]

The acceptance of a behavioral standard may be the most appropriate immediate solution, but Turing and Putnam seem to have been content to let it go at that. Turing declared the concept of consciousness "too meaningless to deserve discussion."[10] They adopted a perplexing stance for philosophers – agnosticism – and many contemporary philosophers are happy to follow suit; debate over consciousness is not just meaningless, they believe, but impossible to resolve. The turn to behaviorism came not from conviction of its worth, but from the lack of a better option. I will argue that such a position of hopelessness is unnecessary; consciousness can be known empirically.

The problem of other minds rests on the assumption that consciousness is accessible only through first-hand experience. But this is dualistic. If each person's consciousness exists only in a special bubble that has no physical manifestation, then it is not physical. To say that consciousness is both material in nature and fundamentally undetectable is to make a claim that is dramatically inconsistent with contemporary scientific thought. Substance is thought to break down into particles that have both charge and extension; if consciousness is material (an assumption required for

[9] Putnam, *Robots*, 691.
[10]Turing, *Computing Machinery and Intelligence*, 442.

any form of artificial intelligence), it must be detectable at some level if the detector knows where to look for it. But that is the problem: how do we figure out what to look for when we don't know what to look for? How do we make the connection between objectively viewed matter and that which we experience as consciousness?

Those who find the problem of other minds unsolvable might answer that we need proof, and that proof is impossible. First-hand experience cannot provide conclusive evidence regarding the nature of consciousness. Self-reporting is not sufficient for understanding of consciousness, because we are unaware of the causal mechanisms within our own brains. However, it seems to me that if we could thoroughly observe an individual's brain in conjunction with honest reporting of his mental states, we would discover much about the nature of consciousness, and perhaps even its causation. Honesty cannot be ensured for any given individual, but given numerous repetitions of the experiment and the assumption that most people are honest, useful data would emerge. For example, consider the following: the materialist understanding of consciousness requires that it must be possible, in theory, to replicate minds. This would be done, perhaps, by tweaking one person's neurons in various ways until the person had the personality, memories, etc. of another; the purpose of this exercise would be to learn which changes in the features of the brain are necessary for changes in consciousness.[11] Depending on how we tweaked the neurons and to what effect, we

[11] Obviously, there are ethical and practical barriers that would prevent the manifestation of this scenario, but I intend it only as a hypothetical situation to help illustrate my later point.

could draw links between brain states and conscious experience, from which we could conclusively accept or reject computational functionalism.

One objection might be that hypothetical scenarios like this one spawn sticky questions regarding personal identity. If my consciousness changes entirely to that of another person, or even if it just changes a little bit, do *I* really still exist or has my body just taken on a new identity? If I cease to exist, then clearly *I* cannot testify regarding certain knowledge of the change in my consciousness, in which case the success of the experiment (drawing links between consciousness and brain state) will depend on correct behavioral analysis. If I claim to have experienced a change from one personhood to another, in fact it suggests that I have *not* experienced such a change; upon becoming the second person, I would lose memory of the first. Even slight changes might be impervious to awareness. If I lose a memory, for example, and all memories of that memory, I cannot know that I have lost it. Self-reporting, even combined with brain observation, therefore becomes an inadequate method for the discovery of mental causation and third-person reporting of consciousness is not definitive. Furthermore, even if we *do* establish, using inductive reasoning, that a certain change in the brain produces a certain change in the *nature* of consciousness, it still does not speak to whether that feature of the brain caused that *moment* of consciousness itself. The brain might be an intermediate link in the consciousness-producing causal chain. For some philosophers, the lack of the plausibility of certain knowledge regarding the causation of consciousness is reason enough to dismiss the entire question.

Those who get caught up on the problem of other minds are forgetting one of life's early lessons: knowledge of causation in the physical world is never certain. Young children often are preoccupied by the question, "Why?" Adults who are grilled by these children are usually eventually reduced to the answer, "Because that's just the way it is." We can superficially understand causation, but when we examine our understanding, it becomes clear that all we actually do is recognize patterns. For instance, we think we understand why a ball rolls (it was pushed) and we think we understand why the push causes the ball to roll (the transference of energy). For many of us, the understanding ends there, but an expert in physics might be able to answer the question "why?" a few more times. Even our physics expert, however, is eventually forced to concede a lack of understanding. You do not wholly understand a cause if you do not understand the cause of the cause. Furthermore, all of these alleged causal understandings are actually theories based on induction. We believe that if the ball is pushed (under certain conditions), it will roll. But that belief is based on our repeated observation of this phenomenon. We have merely recognized a pattern, and concluded from it a causal relationship. Humans are only capable of identifying correlation. Causation is supposed, never known.[12]

Furthermore, we assume similarity in internal structure in entities that display similar characteristics. If a rat is born of a rat, looks like a rat and acts like a rat, we feel certain that it has internal organs much like those of other rats and we will come to

[12] David Hume, "An Enquiry Concerning Human Understanding," in *Modern Philosophy: An Anthology of Primary Sources, Second Edition*, ed. Roger Ariew and Eric Watkins (Indianapolis: Hackett Publishing Company, 2009).

conclusions based on this assumption. We believe in those conclusions with such absolute certainty that we bet our lives on them; rats are often used to test products to determine their safety for humans. If we truly believed extreme variation in the physical nature of rats possible, such tests would be worthless. Induction is by its nature uncertain, but humans trust it.

If we adopt a standard for consciousness in the name of objectivity, but refuse to accept that the causation of consciousness can be understood empirically, we have, in fact, failed to view the situation objectively. As the example of the rolling ball demonstrated, inductive correlative reasoning is good enough to use to identify other causal physical relationships. In the case of the rolling ball, we have come to the inductive conclusion that pushing the ball causes it to roll. If we repeatedly observe that a certain brain state corresponds to a certain mental characteristic, it is fair to assume causation, just as we assume that it is the push that causes a ball to roll, not that the push was an intermediate link in the causal chain[13]. Correlative evidence can demonstrate a link (or lack thereof) between brain physiology and consciousness. This evidence can be used to make conclusive claims about the nature and causation of consciousness.

Of course, the problem is that we have not yet accumulated enough correlative evidence to make conclusive claims about the causation of consciousness. But the situation is not hopeless. By adopting a position of behaviorism, one approaches this problem

[13] Additionally, if brain states are intermediate links in the causality of consciousness, then it is unlikely that syntactical modeling would produce consciousness, since it models a feature of brain states and would therefore be modeling an intermediate step.

from the wrong angle. If you turn to robots for the answer to the question of consciousness, you are looking in the wrong place. Clearly, one cannot look into a robot to determine whether or not it has consciousness. That would be like trying to determine whether something plays music without any knowledge or understanding of the nature of music. A more practical course of action is to look for the root of consciousness, and to do that, it is far wiser to look where we assume it does exist (in humans) than where we are trying to create it (robots).

**IV. On Correlation**

Correlation can be used in two ways. First, as I have suggested, positive correlation can lead to valid causal claims. If a light turns on every time I flip a functioning light switch, I might make the inductive claim that flipping a functional light switch causes a light to turn on. Induction is useful, but not a logically strong form of reasoning. It might be, for example, that one cause has two effects, and I correlate the two effects to each other rather than to their mutual cause. For example, a faulty light switch might produce a spark immediately after I flip it, just before the light turns on. I might induce that the spark causes the light to turn on. This would have the same inductive validity as the claim that flipping the switch turns on the light, but it would not be correct.

Negative correlation, however, is logically conclusive. Only one instance of the correlation of A and B is required to disprove the conditional statement, "If A, then not B." For example, the belief that no dogs bite humans can be disproved by the single instance of a dog biting a human. If use of computer programming to produce AI tends to have human-like results in

behavior, the behaviorist might inductively conclude that the two are equivalent and computational functionalism correct. However, it takes only one demonstration that the brain and the computer, given equivalent structural changes, produce different results to show that, at the very least, our current programming provides a flawed model of the brain.

**V. Implications for Artificial Intelligence**

In Section III, it was established that the search for the root of consciousness need not be futile so long as one is looking in the right place: the human brain. When we pose the question of whether AI might produce consciousness, it is important to recall that most of the initial hope for AI stemmed from its similarity to brain processes. Neurons send signals to one another with short blasts of energy, which is in some ways similar to how computers process binary code. However, it is important to note that this is not strictly true. Not all neurons fire in short bursts; some send longer signals not accounted for by computer syntax. Additionally, neurons exist in a net, whereas binary programming is linear. In his book What Computers Still Can't Do, Hubert Dreyfus described the problem of "know-how." When a person becomes an expert at a task, he no longer needs to think through all the steps of the task, but rather the proper course of action is immediately obvious. For example, a master chess player does not have to think through the rules of the game before making a move, but rather sees the position of the pieces on the board and knows instantly what to do. By contrast, the more data the computer chess player has about the game of chess, the more information it will have to analyze before making a move. Although, in general,

consciousness alone is a poor means for understanding underlying mental causation, in this case it was indicative of an underlying mechanism. Neuroscientists have explained the "know-how" phenomenon by the fact that when two neurons are simultaneously excited, the connection between them is strengthened.[14] Newer models of AI ("connectionist" models) have incorporated links like these into programming, but they are poor models for neural nets. Ultimately, even connectionist programming boils down to binary code.

For the sake of argument, however, let us grant that neuronal impulses are the source of consciousness and that binary code is a decent model for them. The question now is whether being a model is good enough to produce consciousness, or if there is some further biological feature necessary. For binary code to model neuronal information processing, one must be able to imagine that at any given moment, the neurons of the brain can be mapped syntactically. The alteration of patterns in binary code must produce output to the alteration of neuronal patterns. A recent study led by Mriganka Sur casts doubt on the causal nature of brain structure. Sur and his colleagues performed surgery on newborn ferrets,[15] so that each had one eye that sprouted connections into the part of the brain that is generally dedicated to hearing (rather than into the visual thalamus and visual cortex).

---

[14] Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press, 1997.

[15] Granted, I stated at the beginning of this paper that I was not going to tackle the notion of animal consciousness. However, the scientific community often extrapolates findings concerning animal physiology to humans, and I am assuming that this study is accurate in suggesting that there would be similar findings if we were to perform this study in humans.

There was no resulting change in the ferrets; they continued to see with the affected eyes, using the auditory portions of their brains.[16] An immediate change in neuronal patterns (and in our imaginary syntax which we have mapped onto the brain) produces no change in consciousness. This suggests plasticity of consciousness that is not observed in the output of AI. By comparison, it is difficult to believe that significant change in syntax would not produce observable change in computer function. In other words, in the case of computer syntax, there is a conditional relationship: if there is considerable change in syntax, there will be change in output.[17] For neurons, we have seen the equivalent conditional statement disproved. Here we have established lack of correlation between the result of neuronal behavior and that of syntactical programming; at the very least, we must conclude that current efforts to use computer syntax to model brain functions are fundamentally flawed. Just as a fundamental change in a recipe would not necessarily produce an observable change in outcome, but would very likely do so, this does not prove that syntax does not produce consciousness, but it suggests as much.

**VI. Discussion**

We have established that if neuronal impulses and syntactical programming each produce consciousness, they must

[16] Alva Noe, *Out of our Heads: Why You are Not Your Brain, and Other Lessons from the Biology of Consciousness* (New York: Hill and Wang, 2009), 53-54.
[17] One possible response to my argument would be a rejection of this claim. I am not a computer scientist, so I cannot say with absolute certainty that such a response would be unfounded. However, I think it is undisputable that if the syntax experienced the same degree of change as the neuronal impulses in this example, there would be noticeable change.

do it in different ways. Stalwart defenders of AI might claim that this is possible: that AI and the brain are fundamentally different from one another, yet produce equally valid consciousness. To defend themselves, they would likely revert to the problem of other minds. However, as I have already claimed, the problem of other minds should be dismissed as subjective. The claim that consciousness could be formed in two completely different ways is, first and foremost, unrealistic. It stems, I believe, from the belief that consciousness is spiritual – that it rises above and inhabits the physical world. If we instead accept consciousness for what it is – a biological phenomenon – it seems no more likely that computer programming (having proved dissimilar to the brain in every important way) could produce *it* than any other biological phenomenon (e.g. photosynthesis). Furthermore, if we reject the spiritual view of consciousness, yet accept that consciousness could be produced in a way that does not model the workings of the brain, we have no basis to judge what is conscious and what is not. The notion of consciousness becomes meaningless.

The conclusion to be drawn is that there is good reason to believe that syntax based AI does not produce consciousness.

## Bibliography

Dennett, Daniel C. "Consciousness in human and robot minds." In *Cognition, Computation & Consciousness*, edited by Masao Io, Yasushi Miyashitatt and Edmund T. Rolls (Oxford: Oxford University Press, 1997), 17-29.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press, 1997.

Hume, David. "An Enquiry Concerning Human Understanding." In *Modern Philosophy: An Anthology of Primary Sources, Second Edition*, edited by Roger Ariew and Eric Watkins (Indianapolis: Hackett Publishing Company, 2009), 533-600.

Noë, Alva. *Out of our Heads: Why You are Not Your Brain, and Other Lessons from the Biology of Consciousness*. New York: Hill and Wang, 2009.

Putnam, Hilary. "Robots: Machines or Artificially Created Life." *The Journal of Philosophy* 61, no. 21 (1964): 668-691. <http://www.jstor.org/stable/2023045>

Searle, John R. *The Rediscovery of the Mind.* Cambridge, MA: MIT Press. 1992.

Turing, A.M. "Computing Machinery and Intelligence." *Mind* 59, no.236 (1950): 433-460. <http://www.jstor.org/stable/2251299>

# Spatial Information and Diagrams

*Meghan Ertl-Bendickson*

*[This paper received the* 2011 Jakob Laub Prize *in Philosophy.]*

## Introduction

In recent times, it has become undesirable to use diagrams in logical proofs. Logical proofs, even in geometry, are ideally purely formal representations. Recent experiments by David Kirshner and David Landy, however, have shown that the way in which we physically arrange symbols on a page when we write a formula affects whether or not we compute it correctly. Specifically, we normally place multiplied (or divided) terms closer together than added (or subtracted) terms – following the order of operations. The operations which are supposed to be performed first are placed physically closer together than those which are done later (I shall refer to this as the "Rule of Spacing"). When formula are written inconsistent with this rule, people make more computational errors. Landy claims that this implies, through his "longer is larger" hypothesis and his "syntax" hypothesis, that there are diagrammatic elements to our formal representations. I argue that even if these spatial relations are diagrammatic, it is not a problem for logic the way using a conventional diagram would be. However, while I agree that these results are very important and need to be discussed, I argue that these spatial relationships are not actually diagrammatic.

**Why Diagrams are Problematic for Logic and Math**

Before we can examine Kirshner and Landy's results, we need to understand some background information about diagrams and why, exactly, it is no longer considered acceptable to use them in logical proofs. Diagrams were originally developed, in the times of Ancient Greece, for use in cartography and to find ways to accurately measure spaces and distances. This means that the first diagrams were meant to describe contingent, extensional properties of the real world. "Geometry as a discipline originated in the need to solve problems concerned with distances and areas in surveying and cartography. Its subject matter was therefore the physical features of the world, and the logical relationship its conclusions bore to these features was therefore contingent, akin to that of any physical theory." [1] They were used to deal with specific instances in space and time, for instance mapping a real landscape in a particular area. Geometry developed out of these issues.

However, it has since become something quite different. A critical change came when Descartes presented to us a way to describe geometric diagrams algebraically, allowing us to convert diagrams into formal representations.[2] This was beneficial to the study of geometry in a number of ways. It allowed geometry to directly profit from advances made in the rest of mathematics, so that if a new discovery were made elsewhere it could be applied to geometry, as well. It also solved the issue, which had been recognized for many, many years, that relying too completely on a diagram can cause error solely because the actual diagrams we

[1] Greaves, Mark. 2002. *The philosophical status of diagrams*. (Stanford, Calif: CSLI Publications), 77
[2] Ibid., 78

draw are fallible. No drawing of a triangle is ever going to be a perfect triangle, so basing your calculations on a specific drawing of a triangle can cause mistakes. Working instead with the algebra allows us to talk about "perfect" geometric shapes, without having to worry about whether our diagrams are accurate. Finally, though, Descartes allowed us to begin to discuss things that are not visualizable or intuitable. Geometry was no longer restricted to the domain of things that humans are capable of visualizing. We can talk, now, of 5-dimensional objects, or shapes with more sides than we can picture, etc. This final point makes it clear that geometry had begun to move away from its original purpose – the study of the real world and extensional, contingent spaces.[3]

Another shift came with the discovery of Non-Euclidean geometry. "After this discovery, it was unclear whether the theorems of geometry could even be considered to be *true* of objects of the world, let alone descriptive of their necessary properties, because of the uncertainty about the world's actual geometry." [4] Now there were actual aspects of geometry that specifically did *not* relate to our experience of the world. In fact, we were now left a little uneasy about the exact nature of our world – what kind of geometry do we actually have? We had assumed that there was only this one type of geometry based on rules which govern the real world. But now we could see that there were others, which follow different rules, leaving us unsure as to which one we actually live in. And for those types of geometry that do not represent our world, no diagram could now be of use to us.

[3] Ibid., 78
[4] Ibid., 79

Diagrams had at one point been essential to the study of geometry, but since then the development of geometry itself has tended in a direction in which diagrams can no longer be of substantive use.

Greaves discusses a number of the fundamental reasons diagrams cannot serve a real purpose in logical proofs. The first involves the "requirement of indeterminacy of interpretation."[5] Basically, diagrams inherently impose one interpretation on a problem, but there may be others. Using solely formal representations keeps us from becoming biased towards one particular interpretation. The second reason is slightly more subtle and more pertinent to our present discussion. Logicians, mathematicians, etc have wanted very much to keep psychological processes out of our rules of reasoning. "...the consensus among nineteenth-century mathematicians that proofs in any sort of mathematics be free of any dependency on facts unique to our particular psychology..."[6] Logic is meant to be objectively true, independent of particular human cognition. If the rules of logic are based on a particular human psychological process, then it functions only for human beings, not for the objective world. Further, if a rule of logic is based on a quirk of human cognition, we cannot be entirely sure it is true. We want to describe the world as it objectively is, not the world as we subjectively experience it.

The most fundamental problem for diagrams, however, has to do with a very basic assumption of logic. A logical proof is meant to be as broad as possible. A proof is not valid if it works only for one particular instance on one particular day, or if it

[5] Ibid., 80
[6] Ibid., 80

functions only for one discipline but not others. “A single fundamental principle has been at the center of the way that logicians from Aristotle to Frege have structured their accounts – namely, that the scope of a legitimate logical theory should be as broad and general as possible...logic should not be artificially limited in its domain of applicability, and thus it should attempt to model whatever is common about reasoning broadly conceived, however small that common fraction may be.” [7] We do not want one system of logic for biology, one for chemistry, and another for philosophy. Logic is meant to be a tool applied across all disciplines to make sure that all disciplines are consistent with the real world, not just with our own thoughts. Greaves calls this the *principle of maximal scope*. Diagrams, we have seen, were developed for a purpose in direct opposition to this. Diagrams were *meant* to describe specific, contingent instances, not broad axiomatic laws. This makes diagrams fundamentally at odds with the aim of logic.

**Visual Elements in Formal Representations**

So, we can see now why it has seemed so important to remove all aspects of diagram from our formal representations. Diagrams are contingent, so any diagrammatic element of a formal representation is a potential weakness to the proof. It is a point at which we cannot be sure the proof is following the principle of maximal scope or that it is detached from our psychological processes. Kirshner and Landy's experiments, however, highlight the possibility of just such an element. When we write a formula

[7] Ibid., 194

on the page, certainly that is a visual object. We may call it 'writing' instead of 'drawing', but we must admit that both are visually processed and involve spatial relationships on the page. So we need to clearly distinguish what makes something a formal representation on a page, and what makes it a diagram.

Landy describes two distinctions that have been made. The first is the concept of the difference between intrinsic and extrinsic representations. Diagrams are *intrinsic* representations, because the truth I am trying to show with my diagram is intrinsic to the diagram itself. I can draw a diagram illustrating that line A is longer than line B by drawing one line longer than the other – the difference in lengths of the lines is inherent to the drawing. In a formula, however, all of the symbols involved are arbitrary. The truth I am trying to show is extrinsic to the symbols I make – when I say 1+1=2, nothing about any of those squiggles on the page is inherently related to the numbers involved or the process of addition. The drawing of the lines, on the other hand, is *not* arbitrary.[8]

Another way of getting at this difference is to say that diagrams are direct representations, whereas formal representations are indirect. The formula 1+1=2 is indirect because I arrive at the truth of the statement only through knowledge of outside laws (what the symbol '1' means, what the rule of addition is, etc). But in the diagram of the lines, the truth directly shown to me through the symbols involved. I need no outside knowledge (besides knowing the definition of 'longer') to understand what is being

[8] Landy, David, and Robert L. Goldstone. 2007. "Formal notations are diagrams: Evidence from a production task". *Memory & Cognition.* 35 (8): 2033.

stated.[9] What both of these theories are getting at is the idea of arbitrariness. Formal representations are arbitrary, diagrams are not. So in order to decide whether something is diagrammatic or formal using these definitions, we have to ask whether it is arbitrary, direct, and intrinsic.

Landy aims to show that there are diagrammatic elements to formal representations by showing how the spatial relationships between our arbitrary symbols on the page reflect the processes going on in our calculations and also how making those relationships differ from our norm causes us to make errors. "…the rule system that governs the interpretation of formal systems carry functional spatial information – in other words, they are diagrammatic." [10] Before Landy published his papers, Kirshner[11] published a paper examining the curious fact that when people write out formulas, they place operands closer together or farther apart in reflection of the order of operations. So, 1+2x3=7 tends to be written 1 + 2x3 = 7, with the multiplied terms placed spatially closer together on the page than the added terms. He wished to see if this spatial grouping affected the way we compute, or in other words, if these spatial relationships inform the steps we take to solve an equation.[12]

To do this, Kirshner made a system called a Nonce Notation, which is a system of arithmetic completely divorced from any of the symbols we currently use. This Nonce Notation

[9] Ibid., 2033
[10] Ibid., 2038
[11] Kirshner, David. 1989. "The Visual Syntax of Algebra". *Journal for Research in Mathematics Education.* 20 (3): 274-287.
[12] Ibid., 287

had two difference versions. The first was “unspaced”, the second was “spaced”. The unspaced version had nothing in common with our current notation, the spaced version was exactly the same as unspaced, except following this Rule of Spacing we apparently use. So the two systems were thus:

| Current | Unspaced | Spaced |
|---|---|---|
| a+b | aAb | a A b |
| a-b | aSb | a S b |
| axb | aMb | a M b |
| a/b | aDb | a D b |
| a^b | aEb | aEb |
| b | aRb | aRb |

In the spaced version, the operations which are supposed to be performed first are placed closer together than those which should be performed last, just like what we tend to do when writing in our own notational system.[13]

Kirshner took a group of highschool students and first tested them on how well they understood math in our current notational system. Those who made minimal errors on the test then went on to take the same type of test, except using the Nonce Notation. The first test was unspaced, the second was spaced. These were students who understood the laws of math and the order of operations, so any mistakes they made would mostly be due to having trouble with the new notation. He compared the scores of the first, unspaced test to the scores of the spaced test and

[13] Ibid., 277

found that indeed, students did much better when the notation was spaced. Since the only difference between the two was the spacing, it had to be the spacing itself which made their scores go up.[14] This spacing, which is reflective of the order of operations, does seem to inform our calculations. It is not irrelevant.

David Landy did a series of experiments to follow through on these findings. In his first experiment, he tested how well people could judge the truth of a statement when the spacing of it was inconsistent (meaning, when the statement did not follow the Rule of Spacing). So he asked people (in his case, college students), whether a series of statements were true or false. Some were consistent (i.e., does "axb + cxd" necessarily equal "cxd + axb"? For which the answer is yes), and some were inconsistent (i.e., does "a+b x c+d" necessarily equal "c+d x a+b"? For which the answer is no). He found that people made six times as many errors when the spacing was inconsistent.[15] Inconsistent spacing apparently interferes with people's ability to judge the truth of a statement.

Next, Landy tested whether people really do consistently add these spacings to statements when they write or type them out. First he wrote out formulas in words (so, "one plus one equals two") and asked his participants to write the same formula out in symbols ("1+1=2"). He found that people did indeed place multiplied items closer together than added items.[16] Thinking perhaps this was a quirk of handwriting having something to do with the length of time it takes a person to think about the formula

[14] Ibid., 282
[15] Landy, *Formal Notations as Diagrams*, 2034
[16] Ibid., 2034

(meaning perhaps the gaps were due to a pause in thought), he tested whether the same would happen when typing on a computer. This time, participants were asked to convert English sentences into logical symbols ("if Jack is happy, then Jill is happy" would then become "A → B"). Again, however, people left spaces between groups reflective of the order of operations. So the spacing was present whether the formal sentences were handwritten or typed.[17]

Lastly, Landy tested how spacing affects people's ability to correctly solve formulae. First he had them solve simple expressions with just one operator – so, 1+1, or 2x3. Again, these were either consistently or inconsistently spaced. He found that the spacing mattered mainly for addition. For formulae where addition was the operator, when the spacing was wider than normal participants tended to overestimate, but when the spacing was narrow, they tended to underestimate (Proximity, 13). The last experiment involved compound computations, with more than one operator (i.e. 1+2x3=7). He found that inconsistent spacing led to errors in selecting the correct operation – operands placed closer together tended to be multiplied and operands placed farther apart tended to be added regardless of what the operator actually was.[18]

Landy proposed hypotheses to explain these phenomena beyond simply ascribing it to reflecting the order of operations. He wanted to say that this is not just a representation of the rule itself, but rather a spatial reflection of the cognitive processes that we use

[17] Ibid., 2036
[18] Landy D., and Goldstone R.L. 2010. "Proximity and Precedence in Arithmetic". *Quarterly Journal of Experimental Psychology.* 63 (10): 1953-1968, 18

to follow the rule. For the simple expressions, he proposed what he called the "longer is larger" hypothesis. He speculated that we all have a "mental number line" in our heads and when we do addition (but not multiplication), we start at the first number and "move" ourselves along the line the required number of steps and then see where we end up.[19] So for 1+1=2, I would start at one on my mental number line and then take one step forward. I see that I landed at two, and therefore know that the answer is two. But when spacing is abnormally wide or narrow, it influences my perception of the question so that I overestimate or underestimate the correct response, respectively. Thus the spacing of the formula on the page is a visual representation of the act of walking along my mental number line.

For the compound expressions, Landy offers a somewhat more subtle explanation. He claims that when terms are grouped closer together, it is a spatial representation of how *syntactically* bound together they are (I shall call this the "syntax" hypothesis). "...if, as we suggest, understanding formal symbol structures typically involves spatial resources, then symbolic productions might be expected to reflect syntactic structure: The less tightly two adjacent terms are bound syntactically, the farther apart they should be placed physically." [20] In the expression 1+2x3, 2 and 3 are more tightly syntactically bound than 2 and 1, so I place 2 closer to 3 than to 1 as a visual representation of that tightness.

[19] Ibid., 10
[20] Landy, *Formal Notations as Diagrams*, 2034

**The Rule of Spacing and the Principle of Maximal Scope**

If these are in fact diagrammatic elements in our formal representation, we have to ask what this shows. We have striven to remove diagrams from our computations and proofs because historically, diagrams were meant to represent contingent objects, relations, etc. Because they are contingent, they cannot follow the principle of maximal scope, which means whenever possible we should avoid them in order to keep our math and logic as broad as is possible. The other problem with the Rule of Spacing is that they seem to represent, according to Landy's hypotheses, our cognitive processes. We have tried hard to remove any psychological factors from math and logic, because, again, we do not want math or logic to be contingent on the human mind. Theoretically, another species ought to be able to use logic exactly the way we do. It ought not to work only for human beings.

However, we cannot just reject the Rule of Spacing *solely* because it is diagrammatic. We need to ask whether this is indeed a weakness, whether it does fall prey to the above problems. I argue that if these tendencies are diagrammatic, they do in fact still follow the principle of maximal scope exactly the same way that any arbitrary, formal representation would, and thus are not in fact a problem we ought to eliminate. These diagrams are of a different sort than, say, a drawing of a triangle. Yes, they are a reflection of the cognitive processes we use to solve equations, but so is the plus sign or the equals sign. These things are symbolic ways of communicating the steps we take to solve an equation, and if they are standardized, the way the equals sign is, we eliminate most the problems psychological interference might cause. They are not representations of contingent, extensional objects or relations in

the material world like our diagrams in cartography were. So while we have diagrammatic elements in our formal representations, it is not problematic in the same way.

**Why the Rule of Spacing Does Not Yield Diagrams**

I do not, however, fully support the idea that these *are* diagrammatic elements – specifically *because* of the differences between them and conventional diagrams mentioned above. Certainly they are visual and imagistic. But not all images are necessarily diagrams – all of our arbitrary symbols we use in formal notations are also imagistic in that they communicate their information visually. The distinction we have made is that diagrams are direct and intrinsic. For the "longer is larger" hypothesis, there could be ways to directly represent that. If we do perform addition by walking a "mental number line", a *direct* representation of this would involve making the spaces between symbols bigger for formulas in which the numbers involved are bigger; so we might have 1+1=2, and 3 + 5 = 8. This is a direct representation of our mental number line: we have to go further down it to get to 8 than we do to get to 2, so the formula directly represents this by spacing the numbers farther apart.

But this is not what Landy shown. In fact, what he has shown is the exact *opposite.* He proved that there is a common distance we put between the symbols, and that when that distance is inconsistent, it throws us off and we come up with the wrong answer. This may be proof that we are walking a mental number line and that that is how we do addition, but it is not proof that the Rule of Spacing is diagrammatic. We have a consistent distance, and any deviation from that distance is problematic. So while the

spatial relations of the symbols on the page are important, they are not any sort of direct representation, and thus are not diagrammatic.

For the "syntax" hypothesis, there is more of a sense in which the spatial relationships Landy found are direct. We are saying that two terms are closer together syntactically, and so we place them physically closer together on the page. This seems like a direct representation, or at least, it certainly does not seem arbitrary. However, the idea of two things being "more tightly syntactically bound"[21] is not a reference to a spatial relationship in the first place. The word "close" is misleading – we are referring here to a different kind of closeness. Saying two things are closer syntactically is different than saying Minneapolis is closer to Chicago than to Paris. There is no real physical distance involved in syntax, and there never could be, because syntax is not a physical object to begin with.

What do we mean by "syntactic closeness", then? We may say that being "more tightly bound" is referring to *temporal* distance, in that the terms are more tightly bound because they are dealt with first and are therefore temporally closer together ("tighter"), but then we are right back to referring directly to the order of operations. They are only temporally closer together because the rule of the order of operations says they should be, and if these spacings are only reflective of our rule, then they are most certainly not diagrammatic, unless we want to say that parentheses (which are also only reflective of the order of operations) are also diagrammatic. The spacing would then only be an arbitrary symbol

[21] Ibid., 2034

of the rule, the same way that addition and multiplication symbols are arbitrary symbols of their own respective rules. If we think about the order of operations and what it actually says, there is nothing about physical closeness that directly implies it the way saying a diagram represents that line A is shorter than line B because I have physically drawn line A shorter than line B. The Rule of Spacing, as a representation of the order of operations, is intuitively helpful, but not intrinsic. Again, for these spacings to be diagrammatic and not merely imagistic, they have to represent something in a direct way, and for these reasons if they are only representative of the order of operations, they do not.

Taking a step back, we have to further note that Landy has not in fact *proved* either the "longer is larger" hypothesis or the "syntax" hypothesis. He has shown that the spatial relationships between the symbols on the page affect the way we compute formulae. He has not shown *why* this is the case – that would require a whole different type of experiment. These two hypotheses might be plausible explanations, but they have not yet been proven or even strongly supported. Perhaps the Rule of Spacing *is* only a symbol of the order of operations, and thus arbitrary. Perhaps it is not indicative of some deeper cognitive process. Again, we use parentheses in algebra to help us follow the order of operations, and we do not consider those to be diagrammatic, even though they (like any other symbol, even the numbers) are visual.

The underlying point here is that just because something is visual does not mean it is diagrammatic. The requirement for something to be diagrammatic, by Landy's own standards, is that it is direct and intrinsic. In order for him to support his claim that

there are diagrammatic elements to formal representations, he needs not only to prove that the way symbols are arranged on the page affects the way we think, but also that the spatial relations involved are direct representations and not merely arbitrary symbols. Without this second step, all he has shown is that formulae are imagistic and that there is an aspect of that trait that affects the way we compute that we have not yet acknowledged.

**The Import of the Data**

I am not, however, dismissing the findings of Kirshner and Landy as insignificant. I believe it is still highly important to examine what their results mean. The issue I see for the disciplines of mathematics and logic is not that we have diagrams in our formal representations, but rather that we have implicit rules at play. It seems that an undiscussed rule has developed and been passed from teacher to student, and that it is powerful enough to cause people to make computational errors when it is disobeyed. Why the rule developed in the first place, which is what Landy is discussing with his two hypotheses, is an important and interesting question, but not necessarily relevant to mathematicians, logicians or philosophers. For those disciplines, the fact that the rule exists is the crux of the issue.

There are two ways we may address the Rule of Spacing: We may either actively suppress it, which requires explicit discussion of its existence and then for teachers to make certain they are not subconsciously passing it on to their students; or it needs to be defined and standardized, the same as the rule of addition or the order of operations. Without doing either of these, our psychological processes *are* interfering with our computations

in exactly the way we fear. The math or logic we do is being influenced by subconscious mental processes, and there may be differences in this from person to person. Perhaps what is "close together" or "far apart" for one person is different for another, and so when that first person writes out a formula in what they think is consistent with the Rule of Spacing, it is inconsistent for the second person, causing them to make a computational error. But if we make the rule explicit, perhaps standardize the distances between operands and particular operators, then this would hopefully minimize the interference of our own subjective psychologies.

There are, as Landy points out[22], a number of benefits to this rule, such that perhaps we ought not to bemoan its presence. The fact of the matter is that we are not purely linguistic beings. We also necessarily process information through our senses, since that is how we acquire it. This is unavoidable. For the purposes of written logic and mathematics, this means we process the information visually as well as linguistically. So incorporating visual elements into our rules might make it easier for us to process the information we are trying to convey. Particularly, when we first teach a student arithmetic, making the order of operations a spatial as well as a syntactic rule might make it easier to remember and follow. This would minimize the number of mistakes we make when computing formulae and help us learn faster.

In fact, the rule could be helpful for teachers as well as students.[23] If we had such a visual rule representing the steps we

[22] Ibid., 2038
[23] Ibid., 2038

took to compute a formula, a teacher could more easily see why a student got the wrong answer on a test or assignment. If a student writes 2 x 2+3=10, it is most likely that he or she did not follow the order of operations correctly and thus the teacher can much more easily correct and instruct him or her. On the other hand, if the student writes 2x2 + 3=10, it is of course still possible that he or she does not understand the rule of operations, but it is also more possible that there is some other error responsible. Basically, this visual rule is a way of representing the steps we took to solve an equation, the same way we use parentheses. So, it can communicate more efficiently to a teacher whether a student correctly understands the rule.

**Conclusion**

For many years, logicians and mathematicians have worked to remove diagrams from logical proofs and formulae for the reason that diagrams, due to the nature of their origins, do not follow the principle of maximal scope. We have drawn a strict distinction between diagrams, which are intrinsic and direct, and formal representations, which are extrinsic and indirect, or arbitrary. Kirshner and Landy, among others, have rather convincingly shown, however, that there are relevant spatial relationships to our formal representations – mainly, we tend to spatially represent the order of operations by placing physically closer together those operations which ought to be performed first. Landy explains these tendencies using what he calls the "longer is larger" hypothesis in simple expressions, and what I have called the "syntax" hypothesis in compound expressions. Because these spatial relationships so strongly affect the way we compute, Landy

claims they are diagrammatic.

I argue that if this were so, these diagrammatic elements would in fact follow the principle of maximal scope and therefore not be a problem the way a diagram of a triangle, for instance, is. I further argue, however, that even though the Rule of Spacing is visual and imagistic, it is not diagrammatic because the way it represents the information it is conveying is not direct or intrinsic. Regardless, the Rule of Spacing is currently an unacknowledged rule affecting the way we compute, which is problematic and needs to be addressed.

## Bibliography


Greaves, Mark. 2002. *The Philosophical Status of Diagrams*. Stanford, Calif: CSLI Publications.

Kirshner, David. 1989. "The Visual Syntax of Algebra". *Journal for Research in Mathematics Education.* 20 (3): 274-287.

Landy, David, and Robert L. Goldstone. 2007. "Formal notations are diagrams: Evidence from a production task". *Memory & Cognition.* 35 (8): 2033.

Landy D., and Goldstone R.L. 2010. "Proximity and precedence in arithmetic". *Quarterly Journal of Experimental Psychology.* 63 (10): 1953-1968.

# Epistemic Justification and the Possibility of Computer Proof

*Drew Van Denover*

**Abstract** Some mathematical theorems can be proven only with the help of computer programs. Does this reliance on computers introduce empirics into math, and thereby change the nature of proof? I argue *no*. We must distinguish between the warrant the proof gives for its conclusion, and our knowledge of that warrant. A proof is a priori if and only if the conclusion follows deductively from the premises without empirical justification. I start by defending this definition, and proceed to demonstrate that computer-generated proofs meet its criterion.

For more than one hundred years, mathematicians tried and failed to produce a valid mathematical proof of the "Four Color Theorem", or 4TC. First proposed in 1852, the 4TC conjecture remained unproven until Kenneth Appel and Wolfgang Haken published their solution in 1976. Debate immediately erupted about the legitimacy of their methods. Unlike every previous proof, Appel and Haken's work made ineliminable use of a computer program. Their knowledge of the 4TC depended on the operations of a physical machine—apparently introducing empirical elements into mathematics, the purest a priori science. Thomas Tymoczko soon emerged as a chief critic of the possibility of a "computer-assisted proof." These CAPs, he alleged, incorporate contingent facts about the world, whereas

mathematical proofs require a priori certainty. On his account, we should reject the 4TC as a true "theorem" lest we fundamentally alter the nature of mathematical truth. He writes:

> [The] use of computers, as in the 4CT, introduces empirical experiments into mathematics. Whether or not we choose to regard the 4CT as proved, we must admit that the current proof is no traditional proof, no a priori deduction of a statement from premises …. I will suggest that, if we accept the 4CT as a theorem, we are committed to changing the sense of "theorem", or, more to the point, to changing the sense of the underlying concept of "proof."[1]

I disagree with Tymoczko; CAPs *can* be a priori in the requisite sense. Something is a priori if it has a non-empirical justification—regardless of whether humans have a priori knowledge of that justification. We must distinguish between the warrant the proof gives for its conclusion and *our knowledge of that warrant*. I contend CAPs provide excellent, a posteriori reasons for thinking that Appel's proof has an a priori justification.

Most of the debate turns on what we mean by "a priori proof." I begin by discussing competing definitions, and then offer an account of how computer-generated proofs satisfy the best one. I conclude that we need not choose between CAPs' legitimacy and the aprioricity of mathematics.

[1] Tymoczko, Thomas. 1979. "The Four-Color Problem and Its Philosophical Significance". *The Journal of Philosophy.* 76 (2): 58

**Assumptions**

I want to make explicit some of the background assumptions underlying my thesis. First, I assume that normal mathematical reasoning, such as we find in ordinary human-produced proofs, counts as a priori. Following Frege, this is not to say that we discover arithmetic truths without reference to sense experience, but rather that their *ultimate justification* makes no use of it. Contemporary philosophers of mathematics seem largely to accept this thesis, and anyone denying it would see no epistemic difference between computer-derived proofs and the more natural kind. For the purposes of this paper, we shall therefore bracket objections to the aprioricity of mathematics in general.

Second, we need to outline our general conception of "proof." I agree with Rota that a mathematical proof is fundamentally an *argument*—a "sequence of steps which leads to the desired conclusion."[2] Like any other argument, proofs proceed from a set of premises to a conclusion, which we call a mathematical theorem. I see at least two necessary conditions for proof-hood (although more may exist). An argument is a mathematical proof only if (1) the argument is deductively valid and (2) it is in some sense a priori. These are distinct criteria. Heuristic arguments are increasingly common in the field, and indeed they can provide legitimate a priori mathematical knowledge—however, "The proposition was true for all of the $10^6$ cases we tested" does not amount to a *proof* of that proposition. Observe that Goldbach's Conjecture, for all its inductive support,

[2] Rota, Gian Carlo. 1997. "The Phenomenology of Mathematical Proof". *Synthese.* 111 (2): 183

has yet to achieve the status of "theorem." Similarly, many arguments deductively entail their conclusions, but because their premises are fundamentally empirical claims, they do not enjoy a priori status. Tymoczko's argument denies the second condition that CAPs are a priori, but we will seek to reaffirm it.

**Defining "A Priori Proof"**

We must clarify what we mean by "a priori." In this section I reject the definition Tymoczko uses, which requires proofs necessarily to generate a priori knowledge. Instead, I offer my own definition which does not refer to any particular individual's knowledge at all.

Recall that aprioricity is an epistemological concept. It primarily concerns *knowledge*—that is, justified true beliefs.[3] Specifically, it concerns the "justified" part of knowledge. A given belief is a priori when its justification does not depend on sense experience. I agree with Kripke that, strictly speaking, the predicate "… is a priori" applies to *knowledge* and *belief* exclusively, for they are the only bearers of justification.[4] We know something a priori when we know it on the basis of strictly non-empirical evidence.

As such, calling a proof "a priori" involves a little sleight of hand. Proofs are neither beliefs nor knowledge. They are arguments—abstract mathematical constructions consisting of a set of premises, a conclusion, and the inferential relations between them. An argument is a proof whether or not any particular person

[3] Where the justification and the belief are related in the right way, of course.
[4] Kripke, Saul A. 1980. *Naming and Necessity*. (Cambridge, Mass: Harvard University Press.), 35

*knows* it is a proof, and whether or not anyone *believes* it is a proof. We need to stipulate what "a priori" means when applied to mathematical arguments.

Before presenting my own definition, I want to discuss what I take to be the received definition of "a priori proof":

(1) An argument is an "a priori proof" if and only if it is capable of providing a priori knowledge of its conclusion to people with sufficient mathematical ability and knowledge of the involved concepts.

Intuitively, I find this view highly plausible. As mathematical apriorists by assumption, we think that all mathematical truth can be known without sense experience. Naturally, proofs should provide exactly that knowledge. This definition paints the following picture: When a mathematician reads the proof of a theorem, he mentally internalizes each proceeding step. He holds the entire proof in his mind, and can *see* why it is true. Because he knows the workings of the proof, he believes the theorem it underpins. If asked, he can rely on his understanding alone to justify that belief without recourse to experiential propositions. His knowledge of the theorem is completely a priori.

On definition (1), CAPs are not a priori because they are not surveyable. Since no one mathematician can read the proof in its entirety, no one person can truly *know* it. Appel presumably understands the concepts involved in his proof of 4CT, but when he justifies the results step by step, he must refer to empirical work done by computers. For this reason, Tymoczko denies that CAPs are truly "proofs"—they cannot actually provide a priori

knowledge:

> The mathematician surveys the proof in its entirety, and thereby comes to know the conclusion …. The proof relates the mathematical known to the mathematical knower, and the surveyability of the proof enables it to be comprehended by the pure power of the intellect—surveyed by the mind's eye, as it were. Because of surveyability, mathematical theorems are credited by some philosophers with a kind of certainty unobtainable in the other sciences. Mathematical theorems are known a priori.[5]

I agree with Tymoczko that CAPs are not surveyable in the sense he requires, and if we accept (1), CAPs are not truly proofs. However, I think we have good reason to reject (1) as the criterion for a priori proofs: requiring that proofs be capable of generating a priori knowledge indexes what counts as "proof" to particular, individual minds. On (1), whether a given argument is a proof depends on facts about the person attempting to understand it.

Because knowledge is a species of belief, it belongs to individuals. When Jones and Smith witness the same event, they form their own separate beliefs about it, which then count as knowledge if and only if they are true. So "Jones' knowledge" and "Smith's knowledge" are distinct entities. Further, what is sufficient to provide Jones with "knowledge of x" may not be sufficient to provide Smith with "knowledge of x." What actually *will* generate knowledge in a person depends on facts about that

[5] Tymoczko, *The Four-Color Problem*, 60.

person's perception and reasoning processes, and such contingencies are unacceptable for a good definition of proof.

Imagine an argument that requires hundreds of billions of pages to write down on paper (for example, suppose we somehow printed the results from every computation performed during Appel's the proof of the 4CT). That argument would be unsurveyable in a very real way. The time required to read and absorb it would exceed the human lifespan several times over. By (1), the argument is not a proof. But suppose now that modern technology increases human life expectancy tenfold, and cognitive enhancements permit us to read quickly enough to digest the argument and know its contents. The same definition dictates that now, the argument *is* a proof. Its proof-status changed because of strictly empirical facts which had nothing to do with the argument itself! Suppose further that an environmental disaster destroys the technology, but leaves record of the argument intact. Has it now ceased being a proof?

Mathematicians and philosophers often assert that "false proof" is a contradiction in terms.[6] Proofs are certain and timeless. If Euclid proved a proposition in 300 B.C., that same proof remains equally valid today. Definition (1) does not capture this character of mathematical proofs. We do not want our criteria for proof-hood to depend on any one person's a priori knowledge, because what is a priori knowable in practice will always be contingent. We need a different concept of "a priori proof."

A better definition of "a priori proof" will determine the argument's epistemic status using only features of the argument

[6] Rota, *The Phenomenology of Mathematical Proof*, 183.

itself—not features of the entities reading it. Remember, to call something a priori is to say that its *ultimate justification* does not depend on empirical propositions; whether any one person's knowledge of that justification is also a priori is irrelevant. Hence, I offer a counter-definition:

(2) An argument is an "a priori proof" if and only if:
  (a) none of its premises depend on empirical evidence for justification; and
  (b) the conclusion follows from the premises using only rules of inference with non-empirical justification.

Unlike (1), (2) does not depend upon contingent facts unrelated to the argument itself. The argument will be a priori or not regardless of whom or what is reading it. Moreover, (2) best captures the spirit of a priori as a feature of justifications, rather than genesis. (1) seems dependent on the "context of discovery"—it asks, "How, in practice, did some mathematician come to know the theorem in question?" (2) cares only about how we might, in principle, *justify* that theorem. If we can do so independently of sense experience, our theorem has achieved a priori status. On (2), "a priori proofs" are arguments guaranteed to generate a priori justifications, which is precisely what proofs ought to do.

Given our assumption that "normal" mathematical knowledge is a priori, we can derive the following:

(2*) An argument is an "a priori proof" if:

(a) all its premises are mathematical axioms or theorems; and

(b) the conclusion follows from the premises using only rules of logic.

Deciding whether computer-assisted proofs are legitimately a priori requires only determining whether they meet our two sufficient conditions. Do the computers assisting us employ only mathematically warranted inferences? We have excellent reason for believing they do.

**Do CAPs Meet Our Definition?**

Consider Appel and Haken's proof of the 4CT, for example. Exactly what role did computers play? We should remember that one hundred percent of the conceptual work for the proof was developed by *humans*. Stated roughly,[7] Appel and Hanken developed an *algorithm*—a mechanical procedure for applying a finite number of mathematical operations to some input, terminating in some output. The algorithm—like any valid algorithm—involves only mathematically warranted steps. The mathematicians proved, using tried-and-true human-generated methods, that when the algorithm takes a graph as input, a certain output results if and only if the graph has the property of being

[7] The description that follows oversimplifies a complicated and technical mathematical process, but I believe it accurately portrays the philosophical elements involved.

"reducible".[8] They further proved that if every one of a particular set of graphs *is* reducible, the 4CT must necessarily be correct. No suspect "computer-proof" has been invoked thus far.

Applying the algorithm by hand, however, is simply impracticable. The procedure requires "analysis of about ten thousand neighborhoods of vertices" for each of about fifteen hundred graphs.[9] Given the computational nature of an algorithm, the only reasonable way forward involves outsourcing these calculations to a machine. To do so, they wrote a machine-language program—another series of mechanical instructions that, in theory, cause the machine to run through the algorithm precisely as Appel and Hanken described it, storing its data in bits of RAM. On the hypothesis that the computer functions properly, it executes the algorithm using only inferences with a priori justification.

Three things in this process are of note. First, the work done by computer in CAPs remains purely combinatorial—different in scope, but not kind, from the role that calculators and even abaci serve in "normal" mathematics. That role comes nowhere near the creative artificial intelligence Tymoczko imagines:

> Suppose that advances in computer science lead to the following circumstances. We can program a computer to initiate a search through various proof procedures, with subprograms to modify

[8] I will not discuss here what "reducibility" means as a property of graphs. For details of the proof, see Appel and Hanken, 2002.
[9] Appel, Kenneth and Wolfgang Haken. "The Four Color Problem," in *Philosophy of Mathematics: An Anthology*, ed. Dale Jacquette (Oxford: Blackwell Publishing, 2002), 207

> and combine procedures in appropriate circumstances, until it finds a proof of statement A. After a long time, the computer reports a proof of A, although we can't reconstruct the general shape of the proof beyond the bare minimum…. [T]he question is whether mathematicians would have sufficient faith in the reliability of computers to accept this result.[10]

The kind of method Tymoczko describes goes far beyond a computer-assisted proof—it represents a computer-*generated* proof. Specifically, Tymoczko hypothesizes a scenario in which a computer creates a "proof" of Peano arithmetic's inconsistency. Surely, he says, logicians would find this result "hard to swallow." I agree; we should be very skeptical of such a hypothetical proof—but that hesitation does not indicate that mathematicians lack confidence in the *basic calculations* computers perform. Again, CAPs require only this latter kind of combinatorial computation.

Second, we see that computers might introduce error into proof results in two ways: through flaws in their programming (a software bug), or malfunctions in the physical processes underlying their data storage systems (a hardware bug). Both are real possibilities, but neither differs substantially from the errors commonly found in flawed attempts at proof by humans. We misuse notation and make similar syntactical mistakes with regularity, and our calculations are exponentially more error-prone than those of machines. If I ask a mathematician for even a (relatively) simple combinatorial result—say, the rational

[10] Tymoczko, *The Four-Color Problem*, 74.

representation of $\left(\frac{32497}{8237}\right)^{234} - \sum_{i=1}^{73^7} \frac{587i^{13}}{7}$, he will immediately reach for a calculator or (even more likely thirty years after Tymoczko published his paper) a computer. Why? Because empirically, computers are simply more reliable than humans. Appel, in his philosophical defense of his work, observes:

> When proofs are long and highly computational, it may be argued that even when hand checking is possible, the probability of human error is considerably higher than that of machine error; moreover, if the computations are sufficiently routine, the validity of programs themselves is easier to verify than the correctness of hand computations.[11]

His last comment raises the final, most important point of how computer-derivations function in practice: they are subject to easy and repeated *verification*. Certainly, it is possible for a single processor or a single program to malfunction in some way and thereby produce a false result. But CAPs like that of the 4TC have been reproduced on hundreds of individual computers, and their results agreed upon by numerous independently-coded programs. In fact, new implementations for deriving the 4CT proof continue to appear even in the 21st century. Granted, these results should not give us complete, absolute confidence in its validity (as philosophers, we regard very few things as *certain beyond a doubt*). But given the rigor and frequency of their verification, we can be just about as confident that Appel and Haken's algorithm

[11] Appel, *The Four Color Problem*, 207.

indeed generates the desired output as we can be about any empirical fact.

I say "empirical" without concern, though Tymoczko and his sympathizers would balk at such an admission. They grant that computers are almost always reliable, but argue that when assessing their capacity to prove theorems, we are exclusively concerned with a priori evidence. Tymoczko says as much:

> [T]here is a great deal of accumulated evidence for the reliability of computers in [CAP] operations, and the work of the original computers was checked by other computers....The reliability of the 4CT, however, is not of the same degree as that guaranteed by traditional proofs, for this reliability rests on the assessment of a complex set of empirical factors.[12]

In my estimation, this common argument misses the crucial distinction between the proof's a priori justification for its conclusion, and our knowledge of that justification. As per our definition, proof-hood requires that arguments begin from a priori premises, and proceed along a priori methods; our belief that it does so needn't be similarly a priori. We have overwhelming a posteriori evidence that the computer's methodology follows strict a priori guidelines, and therefore meets our criteria for an "a priori proof."

[12] Tymoczko, *The Four-Color Problem*, 74.

**Conclusion**

Tymoczko and I start from fundamentally different conceptions of what "a priori" means in the context of mathematical results. He roots his entire project in the idea that that "mathematical theorems are known a priori."[13] Are they always? Remember that knowledge is proprietary to individuals. One person can have a priori knowledge of a fact another person knows only empirically, and this principle does not change when applied to mathematical knowledge. Much (dare I say, *most*) mathematical knowledge exists on an a posteriori basis. For example, I have no graduate training in mathematics, but when a Fields medalist informs me she has proven an extremely high-level theorem, I believe her. Is my belief justified? I say *yes*. This woman is likely the most knowledgeable expert on the planet. She has nothing to gain from lying, but everything to lose if caught. If I cannot trust her opinion, I can trust no one's. Is my belief true? If she really has proven the theorem, it must be. In such a case, my belief constitutes a posteriori knowledge of a mathematical theorem. I expect that most undergraduates accept their professors' word about theorems prima facie, and thereby create knowledge of a similar kind. Asserting that theorems are necessarily known a priori seems simply unrealistic.

We better capture the aprioricity of theorems with reference not to how particular individuals *actually* know them, but how those theorems are *justified*. For this, we must look to the proofs' methods. As per (2*), mathematical arguments follow a priori methods when neither their premises nor inferences depend upon

[13] Ibid., 60

sense experience for justification. This certainly seems to be the case for Appel and Haken's proof of the 4TC, and for other CAPs like it.

Tymoczko rightly asserts that *mathematicians' knowledge* of CAPs is necessarily empirical. That fact is difficult to deny. However, it does not speak to the internal operations of the proof, which (in my estimation) are the sole determinants of the proof's a priori status. As long the proof offers an a priori justification for its conclusion, it does not matter whether humans know of that justification in an a priori way. In essence: *we need not know a priori that the proof's warrant is a priori.* Insofar as we trust our belief that hundreds of tests run on hundreds of thousands of combinations of software and hardware platforms cannot *all* be completely mistaken, we should trust our belief that CAPs justify their conclusion without reliance on empirics. Anyone suggesting that CAPs are not sufficient "proofs" for lack of a priori justification cannot ignore this result.

## Bibliography

Appel, Kenneth and Wolfgang Haken. “The Four Color Problem,” in *Philosophy of Mathematics: An Anthology*, ed. Dale Jacquette (Oxford: Blackwell Publishing, 2002)

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

Rota, Gian Carlo. 1997. "The Phenomenology of Mathematical Proof". *Synthese.* 111 (2): 183-196.

Tymoczko, Thomas. 1979. "The Four-Color Problem and Its Philosophical Significance". *The Journal of Philosophy.* 76 (2): 57-83.

# MACALESTER COLLEGE

# JOURNAL OF PHILOSOPHY

**VOLUME 20**
**SPRING 2011**

---

The Macalester College Journal of Philosophy is published every spring. Student submissions in any area of philosophy are considered. Submissions for future volumes should be submitted to a member of the Department of Philosophy at Macalester College.

---

**Citation of Articles**



---

**Contact Information**

Department of Philosophy
Macalester College
1600 Grand Avenue
Saint Paul, MN 55105

http://www.macalester.edu/philosophy