A Liar-Like Paradox for Rational Reflection Principles

JOSHUA SCHECHTER[*]

Abstract

This article shows that there is a liar-like paradox that arises for rational credence that relies only on very weak logical and credal principles. The paradox depends on a weak rational reflection principle, logical principles governing conjunction, and principles governing the relationship between rational credence and proof. To respond to this paradox, we must either reject even very weak rational reflection principles or reject some highly plausible logical or credal principle.

1. Introduction

What is the rational credence in the claim A upon being given the information that the rational credence in A is some number r? A natural answer to this question is: r. Similarly, what is the rational credence in the claim A conditional on the information that the rational credence in A is r? A natural answer to this question is again: r. These answers can be used to motivate a principle that Christensen (2010) calls 'Rational Reflection'. In symbols:

$$\text{(RR) } Cr(A|Cr(A)=r) = r$$

where Cr(A) is the rational credence[1] in the claim A and Cr(A|B) is the rational credence in the claim A conditional on the claim B.[2] RR should be interpreted as saying that the identity holds

[*] This paper has been accepted for publication in *Analysis*, published by Oxford University Press. https://doi.org/10.1093/analys/anad070

[1] By 'rational credence', I mean 'ideally rational credence'. For simplicity, I will assume that there is exactly one ideally rational credence function, though my discussion would remain largely the same if this assumption were rejected.

[2] Christensen states the principle slightly differently: Cr(A|Pr(A)=r) = r, where Cr is the thinker's credence function and Pr is the credence function that would be ideally rational in the thinker's situation. Christensen's principle is

whenever the conditional credence Cr(A|Cr(A)=r) has a value. This conditional credence has a value if Cr(Cr(A)=r) ≠ 0, and perhaps in other cases, too.

RR is a kind of expert deference principle. It says, in effect, that it is rational to treat the rational credence function Cr as an expert and defer to it. Conditional on the expert having a certain credence in A, it is rational to have the very same credence in A.

It turns out that RR can be used to generate a liar-like paradox for rational credence, a paradox that will be presented below. This is a somewhat interesting result but is perhaps not entirely unexpected. There are many paradoxes concerning epistemic notions, including ones closely resembling liar and Curry paradoxes (for example, Kaplan and Montague's (1960) Knower paradox, Thomason's (1980) paradox for idealized belief, and Caie's (2013: section 2) paradoxes for rational credence). Moreover, there are two concerns one might have about the significance of such a result. First, RR is an extremely strong rational reflection principle, one that has been rejected by many – including Christensen (2010), Elga (2013), and Williamson (2011). And second, even putting aside that concern, one might worry that the existence of yet another liar-like paradox has little to teach us. If one thinks, for instance, that we should avoid the original liar paradox by moving to a non-classical logic, one might think that the diagnosis of a liar-like paradox for rational credence will be the same. Presumably, the paradox will rely on the same logical principles at issue in the original liar paradox – such as the Law of Excluded Middle or Reductio ad Absurdum – and so, presumably, we can respond to it in the same way we should respond to the original paradox, whatever exactly that is.

The purpose of this paper is to argue that these reactions are premature. There is a liar-like paradox for rational credence that depends on a surprisingly weak rational reflection

intended to be a constraint on ideally rational credal states. So it applies to the case where Cr is identical to Pr, which is the principle in the text.

principle and on surprisingly weak logical and credal principles. While, as we will see, several of the standard strategies for handling the original liar paradox can be extended to the paradox presented here, many of these extensions involve the rejection of some plausible claim about rational credences. So the paradox presented here does have something important to teach us.

2. Weak Rational Reflection

Consider the following consequence of RR, which I will call 'Weak Rational Reflection':

$$\text{(WRR) From } Cr(Cr(A)=0) \neq 0 \text{ it follows that } Cr(A|Cr(A)=0) = 0$$

WRR is the restriction of RR to the case where $r=0$. (It also requires that $Cr(Cr(A)=0) \neq 0$.) WRR is a very weak rational reflection principle.

Notably, WRR avoids a central problem for RR. RR was motivated above by considering a case in which one learns that the rational credence in A is r and asking what the resulting rational credence in A should be. The apparent answer is r. (And similarly for the rational credence in A conditional on the information that the rational credence in A is r.) However, as Elga (2013) in effect points out, this line of thought is mistaken. The problem is that there is an ambiguity in the presentation of the case. Is one learning that one's rational credence should be r *upon learning that very fact*? Or is one learning that one's rational credence should have been r *prior to learning that very fact*? In the first case, it is intuitively plausible that one's credence in A should now be r. But in the latter case it is not. To motivate RR, it is the latter case that is relevant, since RR concerns the rational credence in A conditional on the rational credence in A having been r prior to learning the new information.

The intuitively right thing to do upon learning that one's rational credence in A should have been r (prior to learning that very fact) is something like this: First, one moves one's credence in A to r (if necessary). Second, one conditionalizes on the information that one's rational credence in A should have been r by moving one's credence in the claim that one's rational credence in A should have been r to 1 and making any corresponding changes to one's other credences. This second step may move one's credence in A away from r.

WRR avoids this problem with the motivation for RR. The reason is that if one learns that one's rational credence in A should have been 0 (prior to learning that very fact), what one intuitively should do is (something like): Move one's credence in A to 0 and then conditionalize on the new information. But conditionalizing on information in which one initially has a non-zero credence can never move one's credence in A away from 0. So WRR is not subject to Elga's problem.

WRR is also a consequence of several rational reflection principles that appear in the literature. Elga (2013) makes use of his criticism of RR to motivate a principle he calls 'New Rational Reflection':

> (NRR) $Cr(A|\text{the rational credence function is } Pr) = Pr(A|\text{the rational credence function is } Pr)$.

WRR can be proved from NRR, at least for the case where there are only a countable number of candidate rational credence functions.[3] WRR is also an immediate consequence of Dorst's (2020) principle 'Simple Trust'. Simple Trust is equivalent in orthodox probability theory to

> (ST) $Cr(A|Cr(A)\leq r) \leq r$

which quickly entails WRR (by setting $r=0$). So WRR should be acceptable even to many who reject RR.

---

[3] The core idea of the proof is that $Cr(A)=0$ is equivalent to the disjunction of claims of the form "the rational credence function is P", where P is a candidate rational credence function that assigns 0 to A. The argument for the countably infinite case depends on countable additivity.

3. *A Liar-Like Paradox*

Given WRR, we can straightforwardly generate a liar-like paradox. We first construct a sentence L provably equivalent (over some background theory we are rationally certain of) to the claim that Cr(L)=0. This can be done in several ways. For instance, L can be constructed using a device of direct self-reference (e.g., 'The rational credence of this very sentence is 0') or stipulation (e.g., 'Let L be the sentence "The rational credence of L is 0"'). Alternatively, L can be constructed using diagonalization, given a background theory of arithmetic (or syntax). The idea for arithmetic is that we start with a system of Gödel numbering for the sentences in our language and a functional expression '$Cr_f(x)$' that stands for the function that maps the Gödel number of a sentence to its rational credence. (This contrasts with 'Cr', which is an operator that attaches to a sentence to yield a term for a real number in [0,1].) We can then show that there is a sentence L provably equivalent (over the background theory of arithmetic) to the claim '$Cr_f(<L>)=0$', where '<L>' is the numeral for the Gödel number of L. So L will be provably equivalent to $Cr_f(<L>)=0$, which presumably will be provably equivalent to Cr(L)=0.

Given such a sentence L, the liar-like paradox is as follows:[4]

> Suppose (for Reductio) that Cr(L) ≠ 0. Under this supposition, Cr(L|L) ≠ 0. Since Cr(L) ≠ 0 and since L is provably equivalent to Cr(L) = 0, Cr(Cr(L)=0) ≠ 0. So by WRR, Cr(L|Cr(L)=0) = 0. So Cr(L|L) = 0. Contradiction!
>
> By Classical Reductio ad Absurdum, Cr(L) = 0. This is provably equivalent to L. So L. We have now proved L. So Cr(L) ≠ 0. Contradiction!

This argument relies on the following principles: (i) WRR, (ii) Classical Reductio ad Absurdum, (iii) from the claims that B is provably equivalent to C and that Cr(B)=0 it follows

[4] See Campbell-Moore 2016: 27 for a related argument against a general class of expert deference principles.

that Cr(C)=0, (iv) from the claims that B is provably equivalent to C, that Cr(B)≠0, that Cr(C)≠0, and that Cr(A|B)=0, it follows that Cr(A|C)=0, (v) from a proof of A (perhaps using WRR) it follows that Cr(A)≠0, and (vi) from the claim that Cr(A)≠0 it follows that Cr(A|A)≠0.

Putting aside WRR, there are two principles that one might worry about in this argument. The first is (v) since one might think that instances of WRR should not be permitted to appear in proofs. However, the appeal to (v) can be avoided if we instead accept that it can be rational to have non-zero credence in the relevant instance of WRR and we accept that from a proof of B from A and from the claim that Cr(A)≠0 it follows that Cr(B)≠0. Given this, we can replace the end of the argument with the following:

> … So L. We have now proved L from Cr(L|Cr(L)=0) = 0. Cr(Cr(L|Cr(L)=0)=0) ≠ 0. So Cr(L) ≠ 0. Contradiction!

This modification, however, seems unnecessary. Proponents of rational reflection principles take such principles to be something like conceptual truths. So if we accept WRR, it is appropriate to permit its use in proofs.

The second principle one might worry about is Reductio ad Absurdum. The reason is that Reductio reasoning plays a rule in many semantic paradoxes. So it is natural to view it as a likely culprit in the paradox. However, if we have the following principle, we can avoid the appeal to Reductio:

> (☆) From A it follows that Cr(A)≠0

The argument is this:

> By (☆), from L it follows that Cr(L) ≠ 0. L is provably equivalent to Cr(L) = 0. So from L a contradiction follows. Any claim from which a contradiction follows has rational credence 0. So Cr(L) = 0.

We can then proceed with the remainder of the argument just as before.[5]

It is not clear to me whether we should accept (☆), even restricted to sentences like L. As is well-known, in infinite cases there can be truths that have rational credence 0. (Typical examples involve dartboards with an uncountable number of possible locations.) Nothing of the sort seems to be going on with L. Nevertheless, this is a place where one could reasonably balk.

There is, however, a different way to avoid the use of Reductio. Indeed, there is a simpler variant of the liar-like paradox, one that is more difficult to avoid. That is the topic of the next section.

4. A Simpler Liar-Like Paradox

WRR can easily be seen to be equivalent (in orthodox probability theory) to the following principle:

(WRR*) $Cr(A \,\&\, Cr(A)=0) = 0$

That is, the rational credence of the conjunction of a claim A with the claim that the rational credence in A is 0 is itself 0.

This principle can be motivated using the intuitions underlying Moore's paradox. Moore's paradox suggests that one rationally shouldn't believe the conjunction of a claim A and the claim that one doesn't believe A. Similarly, one rationally shouldn't believe the conjunction of A and the claim that believing A is not rational. And, similarly yet again, one rationally shouldn't have a non-zero credence in the combination of A and the claim that the rational credence in A is zero. This is WRR*.

[5] This point is analogous to Illustration 1 in Field 2015: 39–40, which shows that we can dispense with the use of Conditional Proof in the Curry paradox if we accept that valid arguments preserve truth.

There are also more theoretical motivations for WRR*. One motivation is that one's credences should cohere with what one takes the rational credences to be. WRR* provides a weak constraint on how one's credence in A should cohere with what one takes the rational credence in A to be. In this way, WRR* is a kind of epistemic anti-akrasia principle. A second motivation is that one's credences should cohere with what one takes one's credences to in fact be. If one is a rational thinker, and has credence function Cr, then WRR* provides a weak constraint on how one's credence in A should cohere with what one takes one's credence in A to be.[6] Viewed in either of these ways, WRR* is a plausible principle.[7]

We can make use of WRR* to generate a remarkably simple liar-like paradox. As before, let L be a sentence provably equivalent to Cr(L)=0. Then the argument goes as follows:[8]

> By WRR*, Cr(L & Cr(L)=0) = 0. Since L is provably equivalent to Cr(L) = 0, Cr(L & L) = 0. So Cr(L) = 0. This is provably equivalent to L. So L. We have now proved L. So Cr(L) ≠ 0. Contradiction![9]

---

[6] A wrinkle with the first motivation is that a rational thinker (perhaps) need not be certain that Cr is the rational credence function. A wrinkle with the second motivation is that a rational thinker (perhaps) need not be certain that Cr is their credence function. In response, we can restrict our attention to rational thinkers who are certain that Cr is the rational credence function or that Cr is their credence function, but who may be uncertain which value Cr assigns to each claim.

[7] To get a sense of just how weak WRR* is, it may be helpful to consider a simple model theory based on classical modal logic, such as the one in Williamson 2011. In this model theory, a model contains a non-empty set of worlds W, a two-place accessibility relation R, an assignment of a weight in (0,1] to each world, and a valuation. The idea is that worlds are epistemic possibilities and wRv obtains if what one is rationally certain of at w is compatible with v obtaining. We require that R be serial and that the sum of the weights for all worlds be 1. The value of Cr(A) at w is the sum of the weights of those worlds accessible from w at which A is true. In this framework, it is straightforward to show that WRR* behaves the same as the modal principle (□T) □(□A→A), which is valid exactly on the quasi-reflexive frames (∀w,v∈W (wRv→vRv)). Reflexivity (and hence quasi-reflexivity) is a plausible constraint on models, since presumably what one is rationally certain of at an epistemic possibility is never incompatible with that possibility obtaining.

[8] This argument is loosely inspired by the version of Curry's paradox in Meyer, Routley, and Dunn 1979.

[9] There is an analogous paradox for the notion of having epistemic justification to believe. Let J(A) be the claim that one has justification to believe A. A plausible principle governing J is J(¬(A & J(¬A))). Let L be equivalent to the claim that J(¬L). Then we can argue as follows: J(¬(L & J(¬L))). So J(¬(L&L)). So J(¬L). So L. We have now proved L. So ¬J(¬L). Contradiction!

There is a second paradox for justification. Another plausible principle governing J is ¬J(A & ¬J(A)). Let L be equivalent to the claim that ¬J(L). Then we can argue as follows: ¬J(L & ¬J(L)). So ¬J(L&L). So ¬J(L). So L. We have now proved L. So J(L). Contradiction!

This argument relies on the following principles: (i*) WRR*, (ii*) from the claims that A is provably equivalent to B and that $Cr(A)=0$ it follows that $Cr(B)=0$, (iii*) from a proof of A (perhaps using WRR*) it follows that $Cr(A)\neq 0$, (iv*) from the claim that B is provably equivalent to C it follows that A&B is provably equivalent to A&C, and (v*) A is provably equivalent to A&A.

These principles are difficult to reject. Notice that there is no appeal to any of the logical principles that have sometimes been viewed as the culprits in the liar and Curry paradox, such as Reductio ad Absurdum, the Law of Excluded Middle, Conditional Proof, and Reasoning by Cases. The only logical principles appealed to are simple principles governing conjunction. The credal principles appealed to are nearly as compelling. So this paradox is a challenging one.[10]

5. A Curry-Like Paradox

A natural question is whether there are Curry-like paradoxes for rational reflection principles, too. Here is one. It is straightforward to show that WRR* is equivalent (in orthodox probability theory) to the following plausible principle:

This second paradox is analogous to a second paradox for rational credence. Consider the following principle: (WRR′) $Cr(A \;\&\; Cr(A)\neq 1) \neq 1$. Let L be equivalent to the claim that $Cr(L) \neq 1$. Then we can argue as follows: $Cr(L \;\&\; Cr(L)\neq 1) \neq 1$. So $Cr(L\&L) \neq 1$. So $Cr(L) \neq 1$. So L. We have now proved L. So $Cr(L) = 1$. Contradiction! In the model theory from footnote 7, WRR′ behaves the same as the modal principle (5c) $\Box\neg\Box A\rightarrow\neg\Box A$, which is valid exactly on the frames satisfying the following condition: For every world w there exists a world v accessible from w such that each world accessible from v is accessible from w. This is also a weakening of reflexivity.

[10] It may be illuminating to contrast this paradox with two other liar-like paradoxes for rational credence – appearing in Caie (2013: section 2.2) and Campbell-Moore (2016: 27), respectively. Caie's paradox relies upon an introspection principle: It is possible for there to be a rational thinker who satisfies both $Cr(A) \geq .5 \rightarrow Cr(Cr(A)\geq .5) > .5$ and $Cr(A) \ngeq .5 \rightarrow Cr(Cr(A)\ngeq .5) > .5$. This principle is weak in the sense that it only requires the possible existence of a rational thinker satisfying the introspection condition, but strong in the sense that satisfying the introspection condition requires a thinker to perfectly discriminate between credences above and below .5. Moreover, in the model theory from footnote 7, while it suffices for the introspection condition to be valid on a frame that R be an equivalence relation – reflexive, symmetric, and transitive – it does not suffice that R have any two of these three properties. Campbell-Moore's paradox relies on any expert deference principle that entails both $Cr(A|Cr'(A)\geq .5) \geq .5$ and $Cr(A|Cr'(A)\ngeq .5) \ngeq .5$ for some credence function Cr′. Such principles are also rather strong. Both Caie and Campbell-Moore's paradoxes also rely on stronger logical principles than does the paradox developed in this section – either Reductio or both the Law of Excluded Middle and Reasoning by Cases, as well as principles connecting credence and negation. So the paradox developed here relies on more austere materials.

$$(\text{WRR}^{**})\ Cr(A \ \&\ Cr(A)=Cr(A\&B)) = Cr(A \ \&\ Cr(A)=Cr(A\&B) \ \&\ B)$$

Using this principle, we can show that Cr(X)=1 for any X we choose.

Let C be a sentence provably equivalent to Cr(C)=Cr(C&X). Then the argument is as follows:

> By WRR**, Cr(C & Cr(C)=Cr(C&X)) = Cr(C & Cr(C)=Cr(C&X) & X). Since C is provably equivalent to Cr(C)=Cr(C&X), Cr(C&C) = Cr(C&C&X). So Cr(C)=Cr(C&X). This is provably equivalent to C. So C. We have now proved C. So Cr(C) = 1. So Cr(C&X) = 1. So Cr(X) = 1.[11]

This argument relies on the following principles: (i**) WRR**, (ii**) from the claim that A is provably equivalent to B it follows that Cr(A)=Cr(B), (iii**) from a proof of A (perhaps using WRR**) it follows that Cr(A)=1, (iv**) from the claim that B is provably equivalent to C it follows that A&B is provably equivalent to A&C, (v**) A is provably equivalent to A&A, (vi**) from the claim that Cr(A&B)=1 it follows that Cr(B)=1, and (vii**) from the claims that a=b and that b=c it follows that a=c.

6. Conclusion

How should we respond to the liar-like paradox for rational credence? There are at least five options: First, we could entirely reject WRR*. This may be the best option, but it has a cost. WRR* is an extremely weak epistemic anti-akrasia principle. It is equivalent in orthodox probability theory to WRR, an extremely weak rational reflection principle. Rejecting these

[11] This argument is Curry-like in two ways. First, the familiar Curry paradox provides a derivation of an arbitrary claim X. This paradox provides a derivation of Cr(X)=1. Second, the familiar Curry paradox involves reasoning with a conditional. This paradox in effect involves reasoning with conditional credences – WRR** is a close relative of the claim that Cr(B|A & Cr(B|A)=1) = 1.

principles seems tantamount to giving up the search for an acceptable epistemic anti-akrasia or rational reflection principle.[12]

Second, we could accept the conclusion of the paradox and claim that L both does and does not have rational credence 0. This would require moving to a paraconsistent logic (such as the kind defended by Priest (2006)). This response has the drawback that it doesn't address the Curry-like paradox, since it is not rational to have credence 1 in every claim.

Third, we could somehow restrict WRR*. For example, we could claim that WRR* should be restricted to sentences that have rational credences and that L does not have a rational credence. Such a response is natural for proponents of paracomplete approaches to the semantic paradoxes (such as the strong Kleene logic-based approach in Kripke 1975 and its strengthening in Field 2008). Such a response is also natural for some proponents of classical logic-based approaches to the semantic paradoxes (such as the approach based on the 'closed off' construction in Kripke 1975). This response has the cost that it requires claiming that not all (apparently) meaningful sentences have rational credences.

Fourth, we could reject a credal principle used in the paradox. One way to do so would be to mimic Tarski's (1983) response to the original liar paradox. On one such view, there is a hierarchy of expressions for rational credence, 'Cr^1', 'Cr^2', …, analogous to the Tarskian hierarchy of truth predicates. $Cr^n(A)=0$ if A contains an occurrence of 'Cr^m' for any $m \geq n$. On this view, it will be false that from the claims that A is provably equivalent to B and that $Cr^n(A)=0$ it follows that $Cr^n(B)=0$, since A but not B may contain an occurrence of 'Cr^n'. Similarly, it will be false that from a proof of A it follows that $Cr^n(A) \neq 0$. This response has the same drawbacks as

[12] Lasonen-Aarnio (2015) suggests that we give up the search for an acceptable rational reflection principle.

the Tarskian view of truth (Kripke 1975). It also requires rejecting there being a rational credence function that behaves as expected on sentences of the entire language.

Finally, we could reject a logical principle used in the paradox. For example, we could reject the claim that A&A is provably equivalent to A. This is the most natural response for proponents of non-contractive substructural logics (such as Zardini (2011)). Given how simple and compelling this logical principle is, this response again has a serious cost.

On any of these five approaches, some highly plausible claim has to be rejected.[13]

References

Caie, M. 2013. Rational probabilistic incoherence. *Philosophical Review* 122: 527–75.

Campbell-Moore, C. 2016. *Self-Referential Probability*. Doctoral dissertation, Ludwig-Maximilians-Universität München.

Christensen, D. 2010. Rational reflection. *Philosophical Perspectives* 24: 121–40.

Dorst, K. 2020. Evidence: A guide for the uncertain. *Philosophy and Phenomenological Research* 100: 586–632.

Elga, A. 2013. The puzzle of the unmarked clock and the new rational reflection principle. *Philosophical Studies* 164: 127–39.

Field, H. 2015. What is logical validity? In *Foundations of Logical Consequence*, eds. C. Caret and O. Hjortland, 33–70. Oxford: Oxford University Press.

Field, H. 2008. *Saving truth from paradox*. Oxford: Oxford University Press.

Kaplan, D. and R. Montague, 1960. A paradox regained. *Notre Dame Journal of Formal Logic* 1: 79–90.

[13] Thanks to Stephan Krämer, Stephan Leuenberger, Bruno Whittle and three anonymous referees for helpful comments.

Kripke, S. 1975. Outline of a theory of truth. *Journal of Philosophy* 72: 690–716.

Lasonen-Aarnio, M. 2015. New rational reflection and internalism about rationality. *Oxford Studies in Epistemology* 5: 145–71.

Meyer, R., R. Routley and J. M. Dunn. 1979. Curry's paradox. *Analysis* 39: 124–28.

Priest, G. 2006. *In Contradiction*, 2nd ed. Oxford: Oxford University Press.

Tarski, A. 1983. The concept of truth in formalized languages. In his *Logic, Semantics, Metamathematics*, 2nd ed., 152–278. Indianapolis: Hackett Publishing Company. [English translation of the 1933 Polish original.]

Thomason, R. 1980. A note on syntactical treatments of modality. *Synthese* 44: 391–95.

Williamson, T. 2011. Improbable knowing. In *Evidentialism and its Discontents*, ed. T. Dougherty, 147–64. Oxford: Oxford University Press.

Zardini, E. 2011. Truth without contra(di)ction. *Review of Symbolic Logic* 4: 498–535.