Extended Abstract for *Self-Referential Probability*

Catrin Campbell-Moore

June 26, 2017

This thesis focuses on expressively rich languages that can formalise talk about probability. These languages have sentences that say something about probabilities of probabilities, but also sentences that say something about the probability of themselves. For example:

(π) The probability of the sentence labelled π is not greater than $1/2$.

Such sentences lead to philosophical and technical challenges. For example seemingly harmless principles, such as an introspection principle:

> If the probability of φ is x, then the probability of 'the probability of φ is x' is 1.

lead to inconsistencies with the axioms of probability in this framework.

This thesis aims to answer two questions relevant to such frameworks, which correspond to the two parts of the thesis: "How can one develop a formal semantics for this framework?" and "What rational constraints are there on an agent once such expressive frameworks are considered?". In this second part we are considering probability as measuring an agent's degrees of belief. In fact that concept of probability will be the motivating one throughout the thesis.

Chapter 1 — Introduction

The first chapter of the thesis provides an introduction to the framework, including motivation for studying frameworks where self-referential probabilities are expressible. One of the key arguments considered is that such self-referential probabilities are unavoidable once one wants to adopt a framework that can express higher-order probabilities such as:

> Georgie believes to degree 0.99 that Dan believes to degree $1/2$ that the coin will land heads.

and quantification:

> Chris has non-negative degree of belief in every sentence.

Another reason for studying such languages that is presented is that the considerations required for studying such languages are required if one wants to study situations where what beliefs one has can affect what happens. For example

> James will be able to successfully leap across a chasm if and only if he is confident that he'll be able to do so (lets say degree of belief $> 1/2$).

which can be formalised using the sentence

(η) James's degree of belief in the sentence labelled η is $> 1/2$.

Both sentences ('James will successfully jump' and η) are true just if James has degree of belief $> 1/2$ in them.

This chapter also has some initial considerations and challenges that such languages face, for example presenting the conflict between probabilism and introspection that arises once such self-referential probabilities are expressible. The problems that these self-referential sentences lead to are very closely connected to those arising from the liar paradox, generated by a sentence:

(λ) The sentence labelled λ is not true.

And our strategy for understanding the expressively rich probability languages throughout the thesis is informed by work on the liar paradox and theories of truth.

Chapter 1 finishes with some more technical preliminaries, for example introducing the formal languages that will be used throughout the thesis. The languages we focus on are expressively rich ones that can express such self-referential probabilities. The distinction between the expressively rich and the expressively limited languages is roughly given by the difference between a predicate and an operator formulation of probability. An operator modifies a sentence φ to form a new sentence, $\mathbb{P}_{>1/2}\varphi$, whereas a predicate applies to a term, or name of a sentence to form a new sentence, e.g. $\mathsf{P}_{>1/2}\ulcorner\varphi\urcorner$. These names are given by a coding of sentences into the natural numbers; and then, as usual, one can prove a diagonal lemma showing that self-referential sentences can be expressed by the predicate language. The expressively limited operator languages are well-studied but the predicate languages have not received much attention and are problematic because of their ability to express self-reference.

Part I — Developing a Semantics

Chapter 2 — Preliminaries and Challenges

Part I is the more substantial half of the thesis and focuses on the question of how to provide a semantics for this expressively rich framework. Chapter 2 provides an introduction to this question and the method that will be pursued throughout the thesis. In this we first introduce the possible world structures that we will generally base our semantics on, called probabilistic modal structures. These have a collection of 'worlds' and a each world has an 'accessibility measure' over the other worlds, which should be probabilistic. For example:

In the expressively restrictive languages which formalise probability using an operator, and thus cannot express quantification or self-referential probabilities, one can then easily recursively define the semantics which says in which worlds the sentences are true or not. However trying to do this for the expressively rich language won't always work. For example consider a probabilistic modal structure such as:

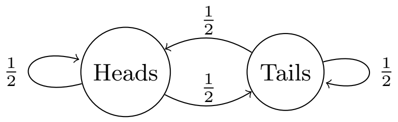

Figure 1: Example of a probabilistic modal structure.

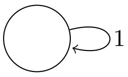

Figure 2: Omniscient probabilistic modal structure.

In this structure there is no interpretation of the probability notion which satisfies the intuitive criterion that corresponds to the recursive procedure used in the operator case:

> $\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner$ is true in w iff the collection of worlds where φ is true has measure (from w's perspective) of $\geqslant r$. I.e.:
>
> $$w \models \mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \iff m_w\{v \mid v \models \varphi\} \geqslant r$$

The reason is directly analogous to the liar paradox for truth: when trying to satisfy the T-biconditional $\mathsf{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$ the liar sentence λ leads to contradictions. For the same reason, a probabilistic liar sentence, π:

(π) The probability of the sentence labelled π is not $\geqslant 1/2$.

will cause the proposed criterion for probability in this probabilistic modal structure to lead to contradiction.

The strategy that will be used throughout the thesis is to generalise theories and semantics developed for the liar paradox, typically by the addition of the probabilistic modal structures. The following three chapters present a systematic study of the way the different semantics for truth can apply to probability.

Chapter 3 — A Kripkean Theory

In Chapter 3 we will present a semantics that generalises a very influential theory of truth: a Kripke-style theory (Kripke, 1975) using a strong Kleene evaluation scheme applied over the probabilistic modal structures. The general strategy follows insights from Halbach and Welch (2009) and is informed by developments by Stern (2015b). The idea of this is that one starts off with no facts about sentences involving probability and truth and iteratively adds more and more probability and truth facts given prior information. So for example at the first stage 0=0 is evaluated as true; then at the second stage one can also evaluate $\mathsf{P}_{=1}\ulcorner 0{=}0\urcorner$ as true; and at the third stage evaluate $\mathsf{P}_{=1}\ulcorner\mathsf{P}_{=1}\ulcorner 0{=}0\urcorner\urcorner$ as true. At no stage in this process does the probabilistic liar, π, become true or false.

This process is more generally and formally done using the probabilistic modal structure; for example the following information about the probabilistic

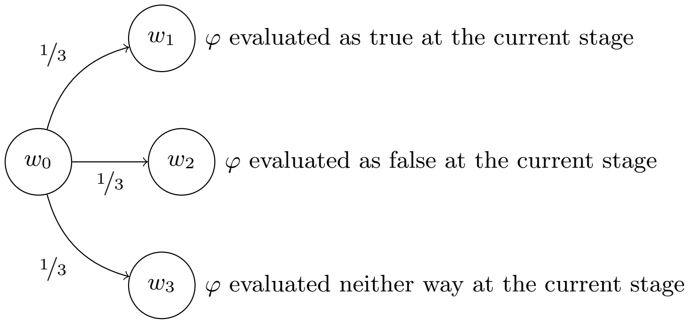
w_0
w_1
w_2
w_3
$^1/_3$
$^1/_3$
$^1/_3$
φ evaluated as true at the current stage
φ evaluated as false at the current stage
φ evaluated neither way at the current stage

Figure 3: A fragment of a probabilistic modal structure representing the information required to evaluate $\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner$ in w_0 at the next stage.

modal structure will allow us to work out whether at the next stage $\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner$ should be made true or not at w_0:
At the next stage $\mathsf{P}_{\geqslant 0.3}\ulcorner\varphi\urcorner$ will be true at w_0 because more than 0.3 weight goes to worlds where φ is currently true; $\mathsf{P}_{\geqslant 0.7}\ulcorner\varphi\urcorner$ is false because too high a proportion of the worlds currently make φ false; but $\mathsf{P}_{\geqslant 0.5}\ulcorner\varphi\urcorner$ is neither. Formally, we consider evaluation functions f, which assign to each world the collection of sentences evaluated as true in that world, and this is revised to obtain a new evaluation function $\Theta(f)$ by:

- $\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \in \Theta(f)(w) \iff m_w\{v \mid \varphi \in f(v)\} \geqslant r$
- $\neg\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \in \Theta(f)(w) \iff m_w\{v \mid \neg\varphi \in f(v)\} > 1 - r$

Since this is monotone (if one has evaluated something a specific way that evaluation will never change) there will be a fixed point: a stage where continuing to attempt to evaluate more sentences won't change anything. These fixed points are proposed as providing the semantics for these expressively rich probability languages.

This results in a final semantics that is not fully classical, which affects the probability notion by assigning sentences intervals as probability values instead of single numbers. For example there can be cases where neither $\mathsf{P}_{>0}\ulcorner\varphi\urcorner$ nor $\mathsf{P}_{<1}\ulcorner\varphi\urcorner$ is true, and we might then think of the probability of φ as being the interval $[0, 1]$. Certain axioms of probability have to be dropped, for example $\mathsf{P}_{=1}\ulcorner\lambda \vee \neg\lambda\urcorner$ is not satisfied in the construction, but the semantics can be seen as assigning non-classical probabilities (Section 3.2.4).

This semantics allows one to further understand the languages, for example the conflict with introspection, where one can see that the appropriate way to express the principle of introspection in this case is in fact to use a truth predicate in its formulation (as investigated in Section 3.4.1). So instead of

$$\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner \to \mathsf{P}_{=1}\ulcorner\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner\urcorner$$

one uses

$$\mathsf{T}\ulcorner\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner\urcorner \to \mathsf{P}_{=1}\ulcorner\mathsf{P}_{\geqslant r}\ulcorner\varphi\urcorner\urcorner.$$

This follows a strategy from Stern (2014a,b) where one should use the truth predicate to perform the function of quotation and disquotation.

In this chapter we also consider how this construction relates to an alternative construction done with a probability operator and truth predicate (whereas this one which was done using both probability and truth formulated as predicates) and note that the two constructions are equivalent. This is analogous to the result for the case of necessity as given in Halbach and Welch (2009) and can either be seen as a defence of the operator approach against the charge of expressive weakness or, as argued for in Stern (2015a), as a defence of the predicate approach against the backdrop of paradoxes, reducing such paradoxes to the notion of truth.

We finish the chapter with a presentation of an axiomatic theory that is intended to capture the semantics (Section 3.5). Such a theory is important because it allows one to reason about the semantics. As was discussed in Aumann (1999), when one gives a possible worlds framework to formalise a game theory context the question arises of what the players know about the framework itself and this question is best answered by providing a corresponding syntactic approach. The theory extends the system KF which is usually given as an axiomatisation of fixed points of the strong Kleene Kripke construction in the case of truth, but also needs to encode facts about the operation of probability and the probabilistic modal structures. Theorem 3.5.5 shows that our theory is complete in the presence of the ω-rule, which allows one to conclude $\forall x \varphi(x)$ from all the instances of $\varphi(\overline{n})$. This rule is needed to fix the standard model of arithmetic. To show the completeness when the ω-rule is present we construct a canonical model, which is of independent interest.

Chapter 4 — Supervaluational Kripke Construction

In Chapter 4, which is rather short, we will consider another Kripke-style semantics but now based on a supervaluational evaluation scheme. The particular reason that this version is interesting is that it ends up bearing a nice connection to *imprecise probabilities*. Imprecise probabilities is a model of probability which drops some particular assumptions of traditional probability by modelling belief states by sets of probability functions instead of a single probability function. It is a model that has been suggested for many reasons: for example because numerically precise credences are psychologically unrealistic, imprecise evidence may best be responded to by having imprecise credences, and they can represent incomparability in an agent's beliefs in a way that precise probabilities cannot.

As discussed in Chapter 2 we have that for many probabilistic modal structures there is no way of satisfying the intended semantics definition. We might alternatively describe this as saying there are no *stable states*: whatever probability evaluation function is chosen, some other probability evaluation function looks better from the original function's perspective. It turns out that this is not the case in the imprecise case. There are some imprecise probability assignments which *do* look best from their own perspective, i.e. there are some stable states. This can therefore be seen as an argument for imprecise probabilities that is very different from the existing arguments.

The development of this semantics also gives us the tools to provide (in Section 4.2) a semantics for groups of imprecise reasoners reasoning about one another. In this we are working in the expressively restricted operator language where no self-reference is expressible, but the more complicated setup being formalised, where agents have imprecise belief states, requires a more complicated semantics, and we can immediately provide this given the semantics developed for the expressively rich framework.

Chapter 5 — The Revision Theory of Probability

In the previous chapters we have developed semantics which drop certain traditional probability axioms and assumptions. In this chapter we will consider an alternative theory of probability where we have that the standard probability axioms are retained.

One influential theory for the liar paradox is the revision theory of truth. The revision theory of truth was independently developed by Gupta and Herzberger and the idea is to improve, stage-by-stage, some arbitrary model of the language. Unlike for the Kripkean construction, such a construction will never terminate but will instead result in a transfinite sequence of models. This lack of a "fixed point" is the price to pay for remaining fully classical. In this chapter we see how one can develop a revision construction for probability. Since the underlying logic is fully classical our probability notion will satisfy the usual axioms of probability (at least for finitely additive probability).

It is important that this process is continued into the transfinite in order to obtain natural probability and truth values, otherwise one might allow, for example, that $\forall n\, \overbrace{\mathsf{T}\ulcorner\mathsf{T}\ulcorner\ldots\mathsf{T}\ulcorner 0{=}0\urcorner\urcorner\urcorner}^{n}$ is false. In applying the revision theory to the case of probability we note that the usual definition of what to do at the limit stages of the revision theory won't be sufficient in the case of probability because there are natural notions of convergence in real numbers which we want to take account of. The limit stage is governed by a criterion for what hypotheses are legitimate and is given by the following idea:

> If a property of interest of the interpretations is brought about by the sequence beneath μ then it should be satisfied at the μ^{th} stage.

In Gupta and Belnap's *locus classicus* on revision theories (Gupta and Belnap, 1993) the authors just consider the properties "φ is true" and "φ is false", and understand "brought about" according to what they call a stability condition. Note that this notion of stability is not connected to that in Chapter 4. For Gupta and Belnap, if φ is true stably beneath μ, meaning that from some point onwards, φ is always true, then φ should also be true at the stage μ; and similarly for falsity. This is a weak way to make this criterion precise and they show that even such a weak characterisations leads to an interesting construction. We will instead present a strong limit stage criterion, the particular change being that we consider more properties. For example we will also be interested in properties like

> The probability of φ is equal to the probability of ψ.

This stronger limit rule allows us to obtain nice models at the limit stages which may have different kinds of properties to the models obtainable at the successor stages. We also suggest that in the case of probability such strenthenings of the limit rule are required otherwise important information about probability that is only obtained through convergence rather than settling on particular values is lost.

In this chapter we consider two different methods of revising probability, which result in different revision constructions for this language. In the first construction we will develop Leitgeb's work from Leitgeb (2008, 2012). This construction cannot apply to general interpretations of probability but instead fixes it to something that might be considered as semantic probability. The second will be based on possible world structures and can be used to give a theory for probabilities in the form of subjective probabilities or objective chances. This is because it is based on background probabilistic modal structures.

That concludes Part I and the development of semantics.

Part II — Rationality Requirements

Chapter 6 — Introduction

In the second part of the thesis we will turn to a different, but related, question:

> What rationality requirements are there on agents in such expressively rich frameworks?

and, relatedly:

> To what degree should an agent believe a sentence that says something about her own degrees of belief?

In this section we are therefore focusing on the particular interpretation of probability as subjective probability, or degrees of belief of an agent.

There has been a large body of work trying to develop justifications for particular rationality constraints on agents, particularly focused on justifying probabilism. There are two main influential styles of argument: an argument from *accuracy*, initially presented in Joyce (1998), and a so-called *Dutch book argument*, originating from Ramsey (1931). The argument from accuracy says an agent should have credences that are as *accurate*, or as close to the truth, as possible. The Dutch book argument says that agents should have credences which, if they bet in accordance with these credences, will not lead them to a guaranteed loss of money. In this part we also consider the question of whether the semantics we have developed can model agents who are doing well from an accuracy or Dutch book point of view.

Michael Caie has recently argued (Caie, 2013) that accuracy and Dutch book criteria need to be modified if there are self-referential probabilities, and that appropriately modified they in fact lead to the requirement that a rational agent must have degrees of belief which are *not* probabilistic and which are also not representable in any of the semantics we have proposed in Part I. If it turned out that Caie's suggested modifications of the criteria were correct, then this would be a blow to our proposed semantics. Perhaps the appropriate response in that case would be to admit that our semantics are unable to model rational agents, so the notion of probability embedded in these semantics could not be interpreted as subjective probability. This question is thus very important for us; however we will be arguing that Caie's suggestions are wrong and that the semantics we have proposed is compatible with both accuracy and Dutch book considerations.

Chapter 7 — Accuracy

Caie argues that rationality should be concerned with how accurate a belief state would be *if it were to be adopted.* Chapter 7 starts with a systematic study of Caie's proposal and will show a number of undesirable consequences of it: It will lead to agent being rationally required to be probabilistically incoherent, have negative credences, fail to be introspective and fail to assign the same credence to logically equivalent sentences. We will also show that this accuracy criterion depends on how inaccuracy is measured and that the accuracy criterion differs from the Dutch book criterion (which will be studied in Chapter 8).

These will give us more motivation to consider rejecting his modification and instead consider something much closer to the usual accuracy criterion: we follow Konek and Levinstein (ms) in arguing that the agent should consider how accurate the considered credences are from the perspective of her current credences instead of considering how accurate they would be if they were to be adopted.

This leaves open the possibility that accuracy considerations do in fact support the semantics we provided in Part I, and in Section 7.3.2 we will show that one way of understanding the accuracy criterion does in fact lead to the semantics developed. In doing this we will still need some additional generalisations and considerations in formulating the accuracy criterion because the semantics we developed, at least in Chapters 3 and 4, dropped certain assumptions implicit in the traditional accuracy criterion by dropping classical logic and the assumption that credences assign single real numbers to each sentence. We will briefly consider how one might apply these considerations in such a setting in Section 7.3.4. This connects to work by Robbie Williams (2012; 2014) on non-classical probabilities. In the semantics developed in those chapters we were able to find *fixed points*, which in Chapter 4 we called *stable states.* It will turn out that for the way we suggest to formulate the rational constraints in Section 7.3.4, these will be exactly the credences that are *immodest*, or look the best from their own perspective, so these are desirable credal states.

Chapter 8 — Dutch book Criterion

In Chapter 8 we will consider the Dutch book argument and will work with the assumption that an agent in such a situation does bet in accordance with her credences, and under that assumption try to develop a Dutch book criterion which is applicable in a wide range of circumstances. In developing this criterion we are expanding a suggestion from Caie (2013). We will show that the proposal that we finally settle on is in fact a version of the modified accuracy criterion that we considered in Chapter 7. It therefore inherits a number of undesirable characteristics. This will therefore lend more weight to our proposal to in fact reject this criterion by rejecting the assumption that an agent bet with her credences.

We finish the thesis with a short conclusion chapter; Chapter 9.

References

Robert J. Aumann. Interactive epistemology II: Probability. *International Journal of Game Theory*, 28(3):301–314, 1999.

Michael Caie. Rational probabilistic incoherence. *Philosophical Review*, 122(4): 527–575, 2013.

Anil Gupta and Nuel Belnap. The revision theory of truth. *MIT Press, Cambridge*, 1(99):3, 1993.

Volker Halbach and Philip Welch. Necessities and necessary truths: A prolegomenon to the use of modal logic in the analysis of intensional notions. *Mind*, 118(469):71–100, 2009.

James M. Joyce. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, 65:575–603, 1998.

Jason Konek and Ben Levinstein. The foundations of epistemic decision theory. Accessed Oct 31, 2014., ms.

Saul Kripke. Outline of a theory of truth. *The journal of philosophy*, 72(19): 690–716, 1975.

Hannes Leitgeb. On the probabilistic convention T. *The Review of Symbolic Logic*, 1(02):218–224, 2008.

Hannes Leitgeb. From type-free truth to type-free probability. In Greg Restall and Gillian Russel, editors, *New Waves in Philosophical Logic edited by Restall and Russell*, pages 84–94. Palgrave Macmillan, 2012.

Frank P. Ramsey. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198, 1931.

Johannes Stern. Modality and axiomatic theories of truth I: Friedman-Sheard. *The Review of Symbolic Logic*, 7(02):273–298, 2014a.

Johannes Stern. Modality and axiomatic theories of truth II: Kripke-Feferman. *The Review of Symbolic Logic*, 7(02):299–318, 2014b.

Johannes Stern. Necessities and necessary truths. Proof-theoretically. *Ergo, an Open Access Journal of Philosophy*, 2, 2015a.

Johannes Stern. *Toward Predicate Approaches to Modality*, volume 44 of *Trends in Logic*. Springer, 2015b. ISBN 978-3-319-22556-2.

J. Robert G. Williams. Gradational accuracy and nonclassical semantics. *The Review of Symbolic Logic*, 5(04):513–537, 2012.

J. Robert G. Williams. Probability and non-classical logic. In Christopher Hitchcock and Alan Hájek, editors, *Oxford Handbook of Probability and Philosophy.* Oxford University Press, 2014.