Synthese
DOI 10.1007/s11229-015-0884-0

S.I.: VAGUENESS AND PROBABILITY

Lost in translation: unknowable propositions in probabilistic frameworks

Eleonora Cresto[1]

Received: 2 July 2014 / Accepted: 28 August 2015

Abstract Some propositions are *structurally unknowable* for certain agents. Let me call them 'Moorean propositions'. The structural unknowability of Moorean propositions is normally taken to pave the way towards proving a familiar paradox from epistemic logic—the so-called 'Knowability Paradox', or 'Fitch's Paradox'—which purports to show that if all truths are knowable, then all truths are in fact known. The present paper explores how to translate Moorean statements into a probabilistic language. A successful translation should enable us to derive a version of Fitch's Paradox in a probabilistic setting. I offer a suitable schematic form for probabilistic Moorean propositions, as well as a concomitant proof of a probabilistic Knowability Paradox. Moreover, I argue that traditional candidates to play the role of probabilistic Moorean propositions will not do. In particular, we can show that violations of the so-called 'Reflection Principle' in probability (as discussed for instance by Bas van Fraassen) need not yield structurally unknowable propositions. Among other things, this should lead us to question whether violating the Reflection Principle actually amounts to a clear case of epistemic irrationality, as it is often assumed. This result challenges the importance of the principle as a tool to assess both synchronic and diachronic rationality—a topic which is largely independent of Fitch's Paradox—from a somewhat unexpected source.

Keywords Knowability · Fitch's Paradox · Moore · Reflection Principle

✉ Eleonora Cresto
eleonora.cresto@gmail.com

1 Instituto de Filosofía, CONICET (National Council of Scientific and Technical Research) – University of Buenos Aires, Puan 470, 4th Fl., C1406CQJ Buenos Aires, Argentina

Journal: 11229-SYNT Article No.: 0884 ☐ TYPESET ☑ DISK ☐ LE ☐ CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

1 Introduction

There is a familiar paradox from epistemic logic—the so-called 'Knowability Paradox', or 'Fitch's Paradox'—which purports to show that if all truths are knowable, then all truths are in fact known; to put it differently, knowability (or weak verificationism) collapses with omniscience (or strong verificationism). This result is paradoxical if we think, as it seems natural to do, that there are indeed truths that nobody knows, and if we also think that the antecedent, the Knowability Principle ('all truths are knowable') is a substantive thesis whose truth or falsity should not be decided *a priori*.[1]

Fitch's paradox is intimately tied to so-called 'Moore's paradox'.[2] Moore calls our attention to the fact that no one can assert a statement such as 'here is a rabbit, but I don't believe it', without causing perplexity in the audience—although, *sensu stricto*, we do not have a formal contradiction, at least not yet. This situation is even clearer if we replace belief by knowledge. A standard way of putting Fitch's argument in motion is precisely by noticing that Moorean sentences of the type 'p, but I don't know that p', if true, can't themselves be known to be true: they are structurally un-knowable.[3]

Here is a brief reconstruction of Fitch's alleged 'modal collapse' between actual and possible knowledge:

1.	$K(\varphi \wedge \neg K\varphi)$	[assumption]
2.	$K\varphi \wedge K\neg K\varphi$	[distribution of K over conjunction]
3.	$\neg K\varphi$	[from 2, by the factivity of K]
4.	$\bot$	[from 2 and 3, by propositional logic]
5.	$\vdash \neg K(\varphi \wedge \neg K\varphi)$	[from 1–4]
6.	$\vdash \Box\neg K(\varphi \wedge \neg K\varphi)$	[from 5, because theorems are necessary]
7.	$\vdash \forall\psi(\psi \rightarrow \Diamond K\psi)$	[Knowability Principle: every truth is knowable][4]
8.	$\vdash (\varphi \wedge \neg K\varphi) \rightarrow \Diamond K(\varphi \wedge \neg K\varphi)$	[instance of 7]
9.	$\vdash \neg\Diamond K(\varphi \wedge \neg K\varphi)$	[from 6, by definition of alethic modalities]
10.	$\vdash \neg(\varphi \wedge \neg K\varphi)$	[from 8 and 9]

[1] The paradox appeared in press for the first time in Fitch (1963). In Salerno (2009) we can find a detailed account of the story of the paradox.

[2] Cf. Moore (1993).

[3] A word of caution. It is not completely obvious that the epistemic version of Moore's claim preserves the most interesting traits of Moore's original paradox without turning it into something different. I will comment very briefly on this point below.

[4] Arguably, a plausible version of this principle may require that the 'K' operator be read as 'someone, at some time, knows that'. However, we may well adapt the Knowability Principle so that it could be represented within models that deal with the knowledge of a single (ideally rational) agent. On the other hand, some attempts to solve the paradox would contend that (7) does not express the intuitive idea of the knowability of truth—say, because it fails to include an actuality operator [as in Edgington's proposal; cf. Edgington (1985)] or because, as it stands, (7) hides the relevant quantifiers, which should be understood as modal indexicals [as in Kvanvig's account; cf. Kvanvig (2006)]. For the most part, in this paper I will not be concerned with possible ways to block the paradox, although I will have something to say about this in the last section.

Journal: 11229-SYNT Article No.: 0884 ☐ TYPESET ☑ DISK ☐ LE ☐ CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

11. $\vdash \forall\psi(\psi \to K\psi)$ [generalization from 10]
12. $\vdash \forall\psi(\psi \to \Diamond K\psi) \to \forall\psi(\psi \to K\psi)$ [from 7–11][5]

Given that truth implies possibility, the converse holds as well, so we can strengthen the result to a biconditional:

13. $\vdash \forall\psi(\psi \to \Diamond K\psi) \leftrightarrow \forall\psi(\psi \to K\psi)$ [from 12]

Moreover, if the Knowability Principle is accepted, we obtain an even stronger conclusion:

14. $\vdash \forall\psi(\psi \to \Diamond K\psi) \wedge \forall\psi(\psi \to K\psi)$ [from 7 and 11]
15. $\vdash \forall\psi(\psi \to (\Diamond K\psi \leftrightarrow K\psi))$ [from 14]

For the most part, what I have to say applies indistinctly to sentences (or statement) and to propositions; if 'φ' is a sentence, the proposition expressed by 'φ' is the set of all and only those possible worlds in which 'φ' comes out true (in a given interpretation). As is customary, I will represent propositions by means of square brackets on the corresponding sentences. Moreover, I will speak loosely of agents knowing or believing sentences, even though this is not the usual way to go. The reader can just take it as an abbreviation for 'knowing or believing that the sentence is true'.

Let me introduce a bit of terminology. By a 'Moorean statement', or 'Moorean sentence', I mean any statement of the form '$\varphi \wedge \neg\Delta\varphi$', where '$\varphi$' replaces any sentence, and 'Δ' is some epistemic or doxastic operator in a wide sense; *mutatis mutandis* for 'Moorean proposition'. Notice that line (6) from the previous argument states that epistemic Moorean sentences cannot be known.

It should be noted that the connection between Moore's and Fitch's paradoxes works only to the extent that the belief operator can be credited with the satisfaction of the right combination of principles, such as distribution over conjunction ('$\vdash B(\varphi \wedge \psi) \to (B\varphi \wedge B\psi)$') and moderate factivity ('$\vdash B\neg B\varphi \to \neg B\varphi$'); alternatively, we can avoid moderate factivity but demand that distribution over conjunction be strengthened to a biconditional, while also demanding doxastic transparency ('$\vdash B\varphi \to BB\varphi$') and doxastic consistency ('$\vdash \neg B(\varphi \wedge \neg\varphi)$'); or perhaps we would rather have distribution over conjunction plus extended doxastic consistency ('if $B\varphi$ and $B\psi$, then $(B\varphi \wedge \psi) \nvdash \bot$'). We might also need to claim that 'sincere assertion bestows belief' (as in Tennant (1997)), insofar as it is the mere *assertion* of a (doxastic) Moorean statement which suffices to cause trouble. I am ready to concede that at least some such principles are fine as far as *rational* belief is concerned, and hence that we can indeed draw an analogy between Moore's paradox and step (6) from the Knowability argument.[6] However, it might be argued that the original Moorean statement need not be about rational belief at all, and hence that no interesting connection can be made with Fitch's Paradox.[7] I will not try to settle this debate here. If the reader feels inclined to reject

[5] To avoid quantification over propositions, one might treat (7) and (12) as sentence schemata and thus omit propositional quantifiers; the same applies to (13)–(15) below (thanks to an anonymous referee for pressing this point).

[6] For sympathetic approaches to the idea that there is an interesting connection between the two paradoxes cf. Tennant (1997), Linsky (2009), or van Benthem (2004), among other authors.

[7] Cf. Kvanvig (2006).

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

the connection, then just assume that I am talking about Moorean* sentences and Moore's Paradox*, where Moore's Paradox* enables us to show the inconsistency of '$B(\varphi \wedge \neg B\varphi)$'.

I will also say that a 'quasi-Moorean statement' captures the essence of Moorean sentences in a probabilistic framework. Here are some possible examples, where '$[\varphi]$' is the proposition expressed by 'φ', 'P' is a subjective, or perhaps an evidential, probability function, and r and s are real numbers in [0,1]:

$P([\varphi] \mid P([\varphi]) = r) \neq r$
$P([\varphi]) = r \wedge P(P([\varphi]) = r) \neq 1$
$\varphi \wedge P([\varphi]) \neq 1$
$\varphi \wedge P([\varphi]) = 0$
$\varphi \wedge P([\varphi]) < r$ for some acceptance threshold r

I will distinguish between *potential* and *genuine* quasi-Moorean statements; the second ones are the successful candidates. We know we have found a good translation of a Moorean sentence into a probabilistic framework if we are able to preserve essential features of Moorean statements in a probabilistic realm. I will call it 'the symmetry criterion'. Among other things, our candidate should preserve the ability to trigger a Fitch-like paradox.

In this paper I will explore the structure of several quasi-Moorean statements and use them to reconstruct a concomitant version of Fitch's paradox so as to get 'probabilistic un-knowability', so to speak. To carry out this project I will rely on a Kripke model enriched with probabilities. Reflecting on which candidates work, and which ones do not work, will prove useful to draw a number of morals that go well beyond the realm of epistemic or doxastic paradoxes. Ultimately, my purpose is to use this discussion to shed some light on the adequacy of certain epistemic principles. In particular, I will examine the putative existence of an appropriate link between the so-called Reflection Principle in probability and Moorean statements; this, in turn, will raise some doubts on the thought that satisfying the Reflection Principle is mandated by rationality. In addition, the project will highlight the convenience of adopting a possible refinement of the formalism, within which both strands of the paradox get a unified solution, in agreement with the symmetry criterion. I will not elaborate on the details here, as I have already presented the refinement somewhere else.[8] I believe the proposed model is adequate on independent grounds; the fact that it allows for a unified solution should reinforce our confidence in its adequacy.

What is the significance of Fitch's paradox? As is well known, there is no consensus on how to answer this question. To begin with, we can wonder in what sense we have a *bona fide* paradox, and not just a *reductio* of the Knowability Principle. Given that at least some versions of semantic anti-realism might want to embrace the claim that all truths are in principle knowable, Fitch's argument has sometimes been interpreted as a refutation of certain types of anti-realism (Hart and McGinn (1976)). Alternatively, it has been contended that Fitch's strategy is not problematic for the anti-realist once the Knowability Principle is correctly formulated (Edgington (1985), Edgington (2010))

[8] Cresto (2012).

or suitably restricted (Tennant (1997)), or once we realize that the anti-realist should be committed to the use of intuitionistic logic (Williamson (1982)). Yet other authors have suggested that Fitch's result is hard to swallow even for those who have no interest in semantic anti-realism whatsoever (Kvanvig (2006)). In this paper I will remain neutral on this controversy. Recall, moreover, that even though I believe a probabilistic version of the paradox can have an interest in itself, my main goal will be to use Fitch's paradox as a litmus test that will help us assess candidates for quasi-Moorean statements, following the symmetry criterion. This task is compatible with all major interpretations I have just mentioned. Whatever it is that is deemed problematic about the Knowability paradox, the problem should be inherited by a probabilistic setting; alternatively, if all Fitch's argument achieves is a *reductio* of verificationism, then we should find an analogous *reductio* within a probabilistic realm. (Presumably, within a probabilistic setting the verificationist should commit herself to a probabilistic knowability principle: true propositions should be able to have maximum evidential probability.)

2 A probabilistic setting

Consider a Kripke structure $S = \langle W, R, P_{prior}, v\rangle$ for a single agent, where 'W' is a countable set of possible worlds, 'R' is a suitable epistemic accessibility relation among worlds, 'P_{prior}' is a finitely additive prior probability function on subsets of W, and 'v' is a valuation function for the sentences of a suitably regimented language L. We shall assume regularity, *i.e.*, for any $A \subseteq W$, $P_{prior}(A) = 0$ iff $A = \varnothing$. This structure allows us to represent both knowledge and higher-order probability attributions. A similar account has been proposed by Timothy Williamson in recent years,[9] and is also reminiscent of other well known proposals that combine probabilities with epistemic operators, such as Halpern (2003).[10] Within Williamson's original framework, 'P_{prior}' is meant to capture the intrinsic plausibility of worlds before the evidence comes in, but we need not commit ourselves to this particular interpretation; we can simply conceive of it as embodying the priors of the agent, or perhaps the priors ascribed to the agent by the theoretician—the one who attempts to make both knowledge and probability attributions to the agent, from a third person point of view.

Within this setting, define '$R(w)$', for any world w, as the strongest proposition known by the agent in w:

$$R(w) = \{x : wRx\}$$

We will assume that R is reflexive. This guarantees that $R(w)$ is not empty, for any w; it also guarantees the factivity of knowledge, as is well known.

Define next the evidential probability of any proposition '$[\varphi]$' in a given world w, for any w, as the prior probability of '$[\varphi]$' conditional on the strongest proposition known by the agent in that world, *i.e.*:

[9] Williamson (2014).

[10] Halpern (2003), chapter 7.

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

$$P_w([\varphi]) = P_{prior}([\varphi] \mid R(w))$$

where 'φ' is any sentence of L. Recall that $R(w)$ is never empty and that we demanded regularity; hence (unconditional) evidential probabilities are always well-defined.

We can also profit from Williamson's device to refer to higher order probabilities. We shall say that proposition '$[P([\varphi]) = r]$', which tells us that the probability of '$[\varphi]$' is r, is the set of all worlds in which the evidential probability of '$[\varphi]$' is r (for some r in [0, 1]):

$$[P([\varphi]) = r] = \{w : P_w([\varphi]) = r\}$$

Then propositions such as '$[P([\varphi]) = r]$' can be plugged in as further arguments of the prior probability function of the model, thereby obtaining what could be understood (arguably) as a second order probability.[11]

Notice that expressions such as '$P_w([\varphi]) = r$' or '$P_{prior}([\varphi]) = s$' are metalinguistic, and hence they do not belong to the object language. For the most part, here I will leave the mechanism to build probabilistic sentences of L undetermined, *i.e.*, I will not be explicit as to how to build sentences of L that encode the probabilistic commitments of the agent. When needed, I will just underline the relevant proposition and use the underlined expression as a shortcut for some sentence of L that expresses exactly that proposition. Notice that, if it is true that $[\varphi]$'s evidential probability in world w is r, then w belongs to the set of worlds picked out by proposition '$[P([\varphi]) = r]$', and hence any sentence that expresses exactly that proposition will be true in w. In other words, we have:

$$P_w([\varphi]) = r \text{ iff } S, w \models \underline{[P([\varphi]) = r]}$$

To illustrate briefly how the model works, consider the following toy example. Suppose $W = \{w, x, y\}$, and suppose R is as shown in Fig. 1.
Here and elsewhere I use capital letters for sets of worlds, or propositions; when needed, I will also keep on using sentences of L between square brackets. If we assume a uniform prior probability function, we obtain:

$P_{prior}(A) = 2/3$
$P_w(A) = 1/2$
$[P(A) = 1/2] = \{w\}$
$P_w([P(A) = 1/2]) = 1/2$

In other words, A's prior probability is 2/3, while its evidential probability in w is just 1/2. Moreover, w is the only world in which the evidential probability for A is 1/2. Hence the (second order) evidential probability in w of the proposition stating that A's probability is 1/2 is, again, 1/2.

[11] In Sect. 9 I will address some worries on whether this analysis captures what we intuitively demand from a second order probability.

Journal: 11229-SYNT Article No.: 0884 ☐ TYPESET ☑ DISK ☐ LE ☐ CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

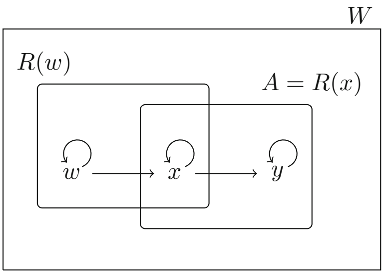

Fig. 1 A toy example for evidential probabilities

3 Quasi-Moorean statements

Are there any obvious candidates to formulate quasi-Moorean statements within framework S? What we might call the 'natural' proposal asserts the conjunction of 'φ' and a (probabilistic) statement telling us that the probability of '$[\varphi]$' is less than 1:

$$\varphi \wedge \underline{[P([\varphi]) < 1]}$$

We will examine this option with care in further sections. But first, we have some work to do. Interestingly, one of the most detailed discussion of the 'natural' candidate one can find in the literature seeks to establish a strong connection between such a statement and the so-called *Reflection Principle* in probability.[12] Moreover, the connection is meant to secure a possible line of defense for the Reflection Principle: violating the Reflection Principle would be analogous to committing Moore's paradox; this is in particular van Fraassen's position in van Fraassen (1995).

How is the connection between the Reflection Principle and Moore's paradox supposed to go, exactly? One possibility is to contend that a sentence that *negates* the Reflection Principle is itself a genuine quasi-Moorean statement. This is not exactly van Fraassen's position, but a close relative. A second, more cautious attitude would be to contend that the negation of the Reflection Principle mimics Moore's paradox because this negation is entailed by the assumption that a statement such as '$\varphi \wedge \underline{[P([\varphi]) < 1]}$' *has itself maximum probability*; this is actually the perspective endorsed by van Fraassen.[13]

In what follows I will begin by showing that the negation of the Reflection Principle is not a genuine quasi-Moorean statement. This result will clear the ground so that we can re-direct our efforts to more promising alternatives. In any case, a critical examination of why this identification fails will prove to be rewarding. I will discuss

[12] The Reflection Principle discussed in this section is not to be confused with 'epistemic reflexivity', most often referred to as 'epistemic transparency' or 'the KK Principle' (for any proposition φ :$\vdash K\varphi \rightarrow KK\varphi$). It is interesting to explore how the two senses of reflection interact with each other; I will take up this topic on board explicitly in further sections.

[13] To wit: Assume $P([\varphi] \cap P([\varphi]) < 1) = 1$. Then $P([\varphi] \mid P([\varphi]) < 1)P(P([\varphi]) < 1) = 1$, which means that both factors are 1. As we will see, '$P([\varphi] \mid P([\varphi]) < 1) = 1$' is a special case of [RP Failure], as will be stated below.

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

more relaxed connections between the Reflection Principle and Moore's paradox in later sections.

Some variants of the Reflection Principle are meant to refer to personal probabilities; others seek to connect personal probabilities with chances, as in David Lewis's *Principal Principle*. A possible version for evidential probabilities, within structure S, may go as follows:

$$P_w([\varphi] \mid [P([\varphi]) = r]) = r, \text{ for all } w \text{ in which the conditional probability is defined} \qquad \textbf{[RP]}$$

(*i.e.*, for all w such that $P_w([P([\varphi]) = r]) > 0$). That is to say, the evidential probability of $[\varphi]$ (in a particular world w), given that the probability of $[\varphi]$ is r, is also r. One of the main reasons why [RP] has attracted so much interest in the literature is its potential connection with discussions on diachronic rationality; we might acknowledge such a connection when we focus on versions of the principle in which we conditionalize on propositions that announce the probabilities held by the agent at later times, as in '$P_{t_0}(A \mid [P_{t_1}(A) = r]) = r$'.[14] In addition, we might endorse a version that makes room for vague probabilities, or vague partial beliefs. In this spirit, van Fraassen's *General Reflection Principle* actually goes like this:

> My current opinion about event E must lie in the range spanned by the possible opinions I may come to have about E at later time t, as far as my present opinion is concerned. (van Fraassen (1995), p. 16)[15]

In the rest of the paper I will focus exclusively on the synchronic case, and I will keep on working with precise real numbers, for the sake of simplicity. In further sections, however, I will seek to establish connections with other senses of vagueness.

Consider, then:

$$P_w([\varphi] \mid [P([\varphi]) = r]) \neq r \text{ for some world } w. \qquad \textbf{[RP Failure]}$$

I hope to show that, in spite of its initial plausibility, [RP Failure] is not a genuine example of a quasi-Moorean statement.

Notice that, within setting S, the validity of [RP] depends on the structure of R. We have a straigthforward counterexample in Fig. 2:

For simplicity, let us assume once again a uniform prior probability distribution; we then obtain:

$P_w(A) = 0$
$P_x(A) = 1/2$
$P_y(A) = 1/2$

[14] Actually, van Fraassen (1995) defends [RP] as a modest constraint on diachronic rationality, as opposed to full-fledged Bayesian conditionalization.

[15] *Pace* van Fraassen, it can be argued that the range of possible opinions I may come to have about E at a later time does not stand for a vague probability, but for a range of possible sharp probability assignments (thanks to an anonimous referee for pressing this point).

Journal: 11229-SYNT Article No.: 0884 TYPESET ☑ DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

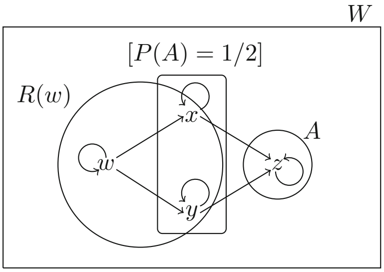

Fig. 2 A model for [RP Failure]

$P_w([P(A) = 1/2]) = 2/3$
$P_w(A \mid [P(A) = 1/2]) = 0$

As we can see, [RP] has been just violated.

As a matter of fact, the semantics of our system guarantee that [RP] holds iff R is an equivalence relation. For an informal account of why this is so, notice that, by definition, the evidential probability of any proposition A, in some world w, refers to the probability of A *given what the agent knows to be the case in that very world* w. Therefore, we should expect that [RP] be satisfied when what the agent knows in w (*i.e.*, '$R(w)$') does not have any chance of affecting her confidence in the proposition on which she is conditionalizing. This means, in turn, that the strongest proposition the agent knows in w is included in the proposition stating that $A's$ probability is (say) r. Now, if R is an equivalence relation, a (second-order) probability on '$[P(A) = r]$' will always be 0 or 1. So, if R is an equivalence relation, [RP] is either undefined or satisfied. As it turns out, this result can be strengthened to a biconditional. More precisely:

Proposition 1 *For any reflexive structure $S = \langle W, R, P_{prior}, v\rangle$: R is an equivalence relation iff for any world w, any proposition A and any r in [0,1], $P_w(A \mid [P(A) = r]) = r$, if it is defined.*[16]

Proof Left to right: By definition, $P_w(A \mid [P(A) = r]) = P_{prior}(A \cap \{w : P_w(A) = r\} \cap R(w))/P_{prior}(\{w : P_w(A) = r\} \cap R(w))$. Assume R is an equivalence relation. Then R partitions the domain in such a way that, for all $y \in R(w)$, $R(w) = R(y)$. Hence for any A, if $y \in R(w)$, $P_y(A) = P_w(A)$. This amounts to saying that, for any particular $r \in [0, 1]$: either $P_y(A) = r$ for all $y \in R(w)$, or $P_y(A) \neq r$, for all $y \in R(w)$. Hence either $R(w) \subseteq \{w : P_w(A) = r\}$, or $R(w) \cap \{w : P_w(A) = r\} = \varnothing$. In the last case [RP] is undefined. If [RP] is not undefined, by contrast, $P_w[P(A) = r] = 1$. Our formula then reduces to $P_{prior}(A \cap R(w)/P_{prior}(R(w)) = P_w(A) = r$.

Right to left: R is reflexive, by assumption. We will prove that it is also symmetric and transitive.[17] To show transitivity, assume there are w, x and y in W such that wRx

[16] This proposition has already been proven by Williamson (2014). By a 'reflexive structure' I mean any S with a reflexive R; recall that we need to assume the reflexivity of R anyway in order to account for the factivity of knowledge.

[17] For this part of the proof I follow closely Proposition 1 in Williamson (2014).

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

and xRy. As xRy, we have $P_x(\{y\}) = b > 0$. Hence $x \in [P(\{y\}) = b]$. Moreover, as $x \in R(w)$, we can guarantee that $P_w(\{y\} \mid [P(\{y\}) = b]) = P_{prior}(\{y\} \cap [P(\{y\}) = b]) \cap R(w))/P_{prior}([P(\{y\}) = b]) \cap R(w))$ is defined, by regularity, and by [RP] it is equal to b. Hence $y \in R(w)$, and so wRy, as desired.

To show symmetry, assume xRy. Hence $y \in R(x)$, and so $P_x(\{y\}) = a > 0$. Thus $x \in [P(\{y\}) = a]$. Moreover, by reflexivity we have $x \in R(x)$. Hence $P_{prior}(\{y\} \cap [P(\{y\}) = a] \cap R(x))/P_{prior}([P(\{y\}) = a] \cap R(x)) = P_x(\{y\} \mid [P(\{y\}) = a])$ is defined, by regularity, and by [RP] it is equal to a. Now suppose, for reductio, that $a < 1$. For transitivity, for every z, if $z \in R(y)$, then $R(z) \subseteq R(y)$. So $P_z(R(y)) = 1$. Then $[P(R(y)) = a] \cap R(y) = \varnothing$. So $P_{prior}(R(y) \cap [P(R(y)) = a] \cap R(x))/P_{prior}([P(R(y)) = a] \cap R(x)) = 0$. Then, by [RP], $a = 0$. Contradiction. Hence $a = 1$. Thus $P_x(R(y)) = 1$. Thus we obtain that $x \in R(x) \subseteq R(y)$, so $x \in R(y)$, and hence yRx, as desired. □

The upshot is that, given that we have not demanded that R be an equivalence relation, violations of [RP] can be true in some structure S. Moorean statements can be true as well, of course. However, doxastic Moorean statements cannot be themselves believed, while epistemic Moorean statements cannot themselves be known to be true (even if they are, as a matter of fact, true). By contrast, it is not difficult to see that if R is not an equivalence relation, violations of [RP] can be given maximum probability in S. So [RP Failure] is not the probabilistic equivalent of a Moorean statement.

Someone could object at this point that all we have shown is that R should be an equivalence relation, out of rationality considerations. I will address this objection with some detail in Sect. 5. But first, let us make sure that [RP Failure] can indeed have maximum evidential probability in S.

4 Longing for Fitch's paradox: paradox unduly lost

Evidently, if [RP Failure] is a genuine quasi-Moorean statement, and [RP Failure] is sometimes justified, then quasi-Moorean statements are sometimes justified, which presumably means that quasi-Moorean statements, unlike their non-probabilistic counterparts, can sometimes be legitimately asserted. Indeed, we can prove that, within the present setting, violations of particular instances of [RP] can in fact receive maximum probability (actually, they are knowable) and hence that quasi-Moorean statements so understood do not trigger a probabilistic version of Fitch's paradox: the modal collapse[18] is suitably blocked. Therefore, either there is no probabilistic analogue of Fitch's paradox—at least not in a Kripke setting like the one presented here—or [RP Failure] is not a genuine quasi-Moorean statement.

Let $S^* = \langle W, R, R^*, P_{prior}, v\rangle$ be a structure where W, R, P_{prior} and v are as before, and where R^* is an alethic accessibility relation. Consider the following version of a Probabilistic Knowability Principle, formulated in the metalanguage:

$$\forall\varphi\forall w\text{: if } \models_w \varphi, \text{ then } \models_w \Diamond\underline{[P([\varphi]) = 1]} \qquad \textbf{[P Knowability]}$$

[18] Or whatever it is that we think Fitch's paradox shows. Cf. the last paragraph of Sect. 1.

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

Or, equivalently:

$$\forall\varphi\forall w\text{: if } \models_w \varphi, \text{ then } \exists x\; wR^*x : P_x([\varphi]) = 1 \qquad \textbf{[P Knowability}'\textbf{]}$$

Consider also the following instance of [P Knowability], for $r \neq s$:

$$\text{If } \models_x \underline{[P([\varphi] \mid [P([\varphi]) = r]) = s]}, \text{ then}$$
$$\models_x \Diamond\underline{[P([P([\varphi] \mid [P([\varphi]) = r]) = s]) = 1]}$$

Essentially, it says that, if a particular violation of [RP] is true in a given world, then it is possible for an agent to assign probability 1 to such a violation. It is easy to check that there are models in which both the antecedent and the consequent of this conditional come out true. All we need to do is find some proposition A such that $\underline{[P(A \mid [P(A) = r]) = s]} \subseteq R(x)$, for $r \neq s$, for some world x that can reach to itself through R^*—which is a fairly reasonable and modest requirement to impose on alethic modalities. To illustrate this situation, let us just enrich our prior toy example from Fig. 2 with the assumption that wR^*w. As the intersection between A and $[P(A) = 1/2]$ is empty, we obtain:

$P_w(A \mid [P(A) = 1/2]) = 0$
$P_y(A \mid [P(A) = 1/2]) = 0$
$P_x(A \mid [P(A) = 1/2]) = 0$

Indeed, the set of worlds in which '$P_w(A \mid [P(A) = 1/2])$' is 0 coincides with $R(w)$; hence $P_w([P(A \mid [P(A) = 1/2]) = 0]) = 1$, and, consequently, $\models_w \underline{[P([P(A \mid [P(A) = 1/2]) = 0]) = 1]}$. Insofar as wR^*w, '$\Diamond\underline{[P([P(A \mid [P(A) = 1/2]) = 0]) = 1]}$' is true in w.[19]

As we can see, we obtained a true instance of [P Knowability] in which we assigned maximum probability to a true (potential) quasi-Moorean statement, so the probabilistic analogue to Fitch's paradox could not get off the ground. This should be disturbing. Paradoxes do not just dissolve in the air; we cannot get rid of a paradox without at least explaining what went wrong the first time we presented it. Regardless of what our favorite diagnosis concerning Fitch's original argument is, we should expect some basic structural features of knowledge attribution to be preserved in a more general setting. Notice that maximal evidential probability behaves as a knowledge-like concept, in the sense that, for a finite W, $P_w([\varphi]) = 1$ iff $\models_w K\varphi$, for any φ and w[20] (if W is infinite, then '$P_{prior}([\varphi] \cap R(w))/P_{prior}(R(w))$' could be 1 even when $R(w) \not\subseteq [\varphi]$, so while the conditional 'if $\models_w K\varphi$, then $P_w([\varphi]) = 1$' is bound to be true, the converse need not hold).

It might be that we cannot do better within the present framework. Or, more likely, we can suspect that our candidate for a quasi-Moorean statement was not good enough. Let me put it as a dilemma: Either there is a serious, insurmountable asymmetry

[19] As a matter of fact we have obtained something stronger, to wit, we have obtained that the sentence stating that [RP Failure] has probability 1 is actually true in w.

[20] Recall that $P_w([\varphi]) = P_{prior}([\varphi] \mid R(w))$ is always well defined, due to the regularity of P_{prior}.

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

between the attribution of knowledge and that of maximal evidential probability, or something has been lost in translation, so to speak.

5 Discussion: reflection, rationality and vagueness

The dilemma from the previous section could be rejected if we restrict acceptable models to those which adopt an equivalence accessibility relation, *i.e.*, to S5. I do not think we should demand such a restriction. Let me elaborate on this point.

Consider the following objection to the idea that violations of [RP] can be successfully known. Some authors have contended that doxastic versions of Moorean statements are not *logically* problematic, in the sense that a (deeply troubled) agent could well believe, *as a matter of fact*, that p and that he doesn't believe that p.[21] To the extent that this is the case, there might be room to develop models in which believing such abnormal statements is possible; we may well rely on a paraconsistent framework, or impose restrictions on the B operator, among other options. Analogously (so the objection goes) we may well find models—such as those presented in the previous section—in which violations of [RP] not only come out true, but where agents may also know that this is the case. However, this does not make such violations any less troubling, in the same sense that building a model that makes it possible for agents to believe Moorean assertions does not dispel the air of paradox we feel in the original Moorean case. Rather, what we should say is that such models do not capture the behavior of *rational* attitudes. An ideally rational agent is one whose R is an equivalence relation, just as a rational agent is one who does not believe conceptual impossibilities. Under this perspective, a rational agent just *cannot* violate the Reflection Principle, let alone *know* that he has violated it. To put it differently, our fan of [RP] can object along these lines: trying to convince a supporter of [RP] that one can be rational and still know [RP Failure] is akin to trying to convince a classical logician that believing a doxastic Moorean statement is also not irrational, because we can always find consolation in paraconsistent logic.[22]

To address this objection let me re-assess whether an equivalence accessibility relation is indeed mandatory for rational agents. As I can see it, we have good reasons to resist this claim. The reasons I have in mind differ from the usual complaint according to which an S5 model, though simpler and more tractable, depicts an unacceptably strong version of ideal rationality. In general, I am not a big fan of attempts to debunk a particular formal account on the grounds that it is too much idealized—idealizations can play an important role at the time of clarifying myriads of notions. Rather, the problem in this case is that S5 models give us *the wrong kind of idealization*. The problem is not that real agents typically do not verify, say, transparency claims (such as those embodied in the *KK* Principle) but, rather, that *ideal* agents should be sensitive to vagueness considerations in a way that S5 agents cannot be.

[21] This is one of the reasons why Kvanvig (2006) claims that there is no interesting analogy between Fitch's and Moore's paradox.

[22] Paraconsistent analyses of Fitch's paradox (such as Beall (2009)) argue that agents can indeed know epistemic Moorean statements; on similar grounds, a paraconsistent logician may well disagree with the claim that an agent just cannot, as a matter of logic, rationally believe a doxastic Moorean statement.

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

Consider worlds w_i, for $i = 1 \ldots n$, such that the length of a particular table is 50 + i cm in each world w_i. Let 'φ' be 'the table is less than 100 cm long'. 'φ' is true in worlds w_1 to w_{49}. It can be intuitively appealing to assume that we can determine the length of the table with an error of ± 3 cm (of course the example can be modified to make it as realistic as we want, but for our current purposes this will suffice). In this scenario, transitivity is violated: in world w_{46} we know that φ, but we do not know that we know this, since in world w_{47}, which is epistemically accessible to w_{46}, we no longer know that φ. In general, demanding transitivity would make it impossible to account for cases in which we would be inclined to attribute knowledge 'locally', so to speak; this is one of Williamson's points in Williamson (2000), ch.7. As is well known, Williamson also uses examples of this sort to explain the erosion that occurs when we go up to higher-order levels of knowledge. Our knowledge of 'φ' becomes more uncertain the closer we get to w_{50}, so at some point it becomes natural to say that we do not know that we know that φ, even if we in fact do know that φ: a naïve sense of epistemic transparency (related to the truth of the *KK* Principle) is bound to fail. Now, nothing in my example so far indicates the presence of vague terms at work;[23] however, we can change the story slightly so that vagueness becomes the main issue. Just take 'φ' to be 'the table is large', or change the scenario to fit your favorite example of vagueness. Let me bracket here the problem of how to assign truth values to such a sentence in intermediate worlds (as this would go beyond the purposes and goals of this paper). Still, it is clear that ideal agents should be discriminative enough so as to know what to say in extreme cases of application of the vague term.

One philosopher's *modus ponens*, however, is another one's *modus tollens*: A fan of S5 can object here that all this argument shows is that, in this and similar scenarios, ideal agents should discriminate better—so it should not be true that $(w_i, w_{i+1}) \in R$ in the first place. But this will often lead us to say that R should be the identity relation, out of rationality considerations. However, it is not clear why we should assume that ideally rational agents are *empirically* omniscient beings, so demanding that R be the identity relation will be just inadequate for most cases.

There is a slightly different argument we can give here to reinforce this conclusion. It may well be that a possible explanation for the existence of vague terms in the language is that we often do not need more precise devices, given, among other things, that our discrimination powers are not perfect. This is related to the failure of empirical omniscience, rather than to a failure of reasoning capabilities. Insofar as we are trying to model perfect reasoners endowed with languages designed to encode imperfect discrimination capabilities, an identity accessibility relation is not the most appropriate tool, as it makes it impossible to use the resources of the language to their full potential. In short, a case can be made for the claim that it is the very existence of vague terms in the language which gives support to the convenience of treating ideal reasoners as exhibiting non-transitive accessibility relations.

An analogous phenomenon takes place within higher-order probabilities: here, too, agents may become increasingly more uncertain. Formally speaking, this is once again

[23] Actually, Williamson's own diagnosis is that this particular phenomenon is not due to any putative vagueness related to the concept of knowledge. I agree; just to be clear, although I do think there is an interesting connection with vagueness here, it is not due to the vagueness of *knowledge*.

Journal: 11229-SYNT Article No.: 0884 ☐ TYPESET ☑ DISK ☐ LE ☐ CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

a straightforward consequence of allowing for non-transitive Rs. Thus, in the light of the previous paragraph, there are no grounds for demanding that second order probabilities be always 0 or 1 *on pain of irrationality*. It is, again, a type of failure of self-knowledge – motivated, among other things, by sensitivity to vagueness-related considerations.

Here is another way of making the same point. There are two senses of reflection at stake: (i) the first one relates to self-knowledge, while (ii) the second one refers to a particular sense of probabilistic coherence, which is at the same time a potentially useful device to bridge the gap between prior and posterior probabilities. As it turns out, once sense (i) is elaborated in a way that makes room for the fact that ideal reasoners can be creatures endowed with vague languages (and less-than-perfect discrimination powers), it exerts immediate constraints on sense (ii). I will come back to these two senses of reflection in Sect. 8.

Incidentally, let me point out that these observations are compatible with the defense of a moderate version of epistemic transparency; to do so we may resort to models in which the behavior of higher order knowledge operators is governed by different accessibility relations R^i, for $i = 1, 2 \ldots$ Then the lack of transitivity in our epistemic accessibility relations can still lead to (weaker) versions of the *KK* Principle (such as '$K^i\varphi \rightarrow K^{i+1}K^i\varphi$'), provided R^i and R^{i+1} are related in an appropriate way.[24] Such models can also verify moderate versions of [RP]. I am not going to dig any deeper on this point here, though the topic will come up again in later sections.

Further considerations might speak in favor of the convenience of abandoning symmetry as well, at least for some scenarios. Hence, once again, it is not clear it is a rationality requirement. Suppose in w_1 Sasha is not feeling quite well, even though she is not running a fever; by contrast, she does have a fever in w_2. In w_1 Sasha is in doubt as to whether the real world is w_1 or w_2; if she were in w_2, by contrast, she will be certain of her having a fever (though in w_1 she does not know that she would be so certain, say, because she is not aware of the accessibility structure of this framework).

In short, there are good reasons not to demand that R be always an equivalence relation, out of rationality. In any case, for our present purposes a milder claim will suffice: we just need to agree that failing to demand such an R (and hence failing to demand that [RP] be satisfied) is not as hard to digest as acknowledging the possibility of rationally believing a (doxastic) Moorean statement. In other words, it does not amount to the same radical departure from usual pre-theoretical notions of what we should expect from ideal reasoners. I trust we can secure at least this basic agreement.

6 In search of a probabilistic version of the knowability paradox

As it happens, there *are* alternative formulations for quasi-Moorean statements within setting S, which yield the probabilistic equivalent to structural unknowability and give rise to probabilistic versions of Fitch's paradox—so the symmetry can be restored. Consider, as we did in Sect. 3, a statement such as '$\varphi \wedge \underline{[P([\varphi]) < 1]}$'. It is easy to see

[24] See Sect. 9.

Journal: 11229-SYNT Article No.: 0884 ☐ TYPESET ☑ DISK ☐ LE ☐ CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

that, although this conjunction can be true in some worlds, its evidential probability can never be 1, in any world. In other words, we can prove that:

Proposition 2 *Let S be as before. Then,* $\forall w : P_w([\varphi] \cap [P([\varphi]) < 1]) < 1$.

Proof Take any world $w \in W$. By definition, $P_w([\varphi] \cap [P([\varphi]) < 1]) = P_{prior}(([\varphi] \cap [P([\varphi]) < 1] \mid R(w)) = P_{prior}([\varphi] \cap [P([\varphi]) < 1] \cap R(w))/P_{prior}(R(w))$. For this probability to be 1, proposition '$R(w)$' needs to be included both in '$[\varphi]$' and in '$[P([\varphi]) < 1]$'. But, although the intersection between '$[\varphi]$' and '$[P([\varphi]) < 1]$' need not be empty (hence the conjunction of the relevant sentences can be true in w), and even though '$R(w)$' could in principle be included in any of the two, it cannot be included in both. If '$R(w)$' is included in '$[\varphi]$', then $P_w([\varphi]) = 1$, and hence, insofar as R is reflexive, $w \in [\varphi]$ (so $w \notin [P([\varphi]) < 1]$). On the other hand, if $R(w)$ is included in $[P([\varphi]) < 1]$, this means that all worlds in $R(w)$ can reach at least one not-φ world. So '$\varphi \wedge \underline{[P([\varphi]) < 1]}$' cannot have maximal evidential probability: it is probabilistically unknowable. □

Incidentally, notice that the truth of Proposition 2 is independent of how we choose the alethic relation R^*. Notice also that this inequality already incorporates the possibility to account for vague probabilities, to the extent that they can be cashed out by intervals $[r, s] \subseteq [0, 1]$.

Let us see now how a Fitch-like argument could go, using the quasi-Moorean statement just suggested, and a principle such as:

$$\forall\varphi\forall w : \text{ if } \models_w \varphi, \text{ then } \exists x : P_x([\varphi]) = 1 \qquad \textbf{[P Knowability 2]}$$

The proof will proceed in the metalanguage, insofar as expressions such as '$P_w([\varphi]) = r$' are formulated by the theoretician. For any structure S:

1.	$\exists w : P_w([\varphi] \cap [P([\varphi]) < 1]) = 1$	[assumption][25]
2.	$\neg\exists w : P_w([\varphi] \cap [P([\varphi]) < 1]) = 1$	[by Proposition 2]
3.	$\forall\varphi\forall w :$ if $\models_w \varphi$, then $\exists x : P_x([\varphi]) = 1$	[P Knowability 2]
4.	If $\models_w (\varphi \wedge \underline{[P([\varphi]) < 1]}$, then $\exists x P_x([\varphi] \cap [P([\varphi] < 1]) = 1$	[instance of [P Knowability 2]]
5.	If $\models_w \varphi$ and $P_w([\varphi]) < 1$, then $\exists x P_x([\varphi] \cap [P([\varphi] < 1]) = 1$	[from 4]
6.	If $\models_w \varphi$, then $P_w([\varphi]) = 1$	[from 5, 2]
7.	$\forall\varphi\forall w$: if $\models_w \varphi$, then $P_w([\varphi]) = 1$	[generalization from 6, given that 'φ' was any proposition whatsoever, and 'w' was also any world whatsoever.]
8.	If for all worlds and sentences, if $\models_w \varphi$, then $\exists x P_x([\varphi]) = 1$, then for all worlds and sentences, if $\models_w \varphi$, then $P_w([\varphi]) = 1$	[3 – 7]

[25] Notice that here I am trying to mimic standard proofs of Fitch's result. Such proofs typically start by assuming that a Moorean statement can be known. Likewise, here I start by assuming that a quasi-Moorean statement (i.e., '$\models_w (\varphi \wedge \underline{[P([\varphi]) < 1]})$') can receive maximum evidential probability

Journal: 11229-SYNT Article No.: 0884 TYPESET ☑ DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

In other words, if all propositions that are true in a particular world are given maximal evidential probability in *some* world, then all propositions that are true in a world are given maximal evidential probability in that very same world. So R is bound to be the identity relation.

Notice that we have not required Necessitation in step (2), even though it is easy to show that (2) is equivalent to a statement with a modal operator:

$2'$. $\forall x \neg \exists wxR^*w :\models_w \underline{P([\varphi] \cap [P([\varphi]) < 1]) = 1}$, *i.e.*:
$2''$. $\forall x \models_x \Box \underline{P([\varphi] \cap [P([\varphi]) < 1]) < 1}$

Notice, moreover, that [P Knowability 2] is weaker than [P Knowability] as I used it in Sect. 4; there are no diamonds in step (3) of the proof as I have just presented it. [P Knowability 2] demands that every truth have maximal evidential probability in some world (rather than: in some of the worlds that relate alethically to a given world). So we need not consider an alethic R^* at all (or, equivalently, we can say that R^* is the universal relation). This simplifies the proof a bit, without loss of generality. Actually, as we relied on a weaker Probabilistic Knowability Principle, the result we obtained is stronger than the one we would have attained with the aid of [P Knowability]. In any case, notice that a similar simplification could have been applied to the original, non-probabilistic Fitch's paradox.

7 Generalizing the knowability principle

We can generalize what we have presented so far and consider an even weaker formulation for the Probabilistic Knowability Principle, along the following lines:

$$\text{If } \models_w \varphi, \text{ then } \exists x : P_x([\varphi]) \geq r; \qquad \textbf{[P Knowability 2}'\textbf{]}$$

where r is a threshold for, say, 'the agent is confident enough' (given what she knows). Once again, we will obtain a Fitch-like paradox when 'φ' is replaced by '$\psi \wedge \underline{[P([\psi]) < r]}$' To see this, notice that we can prove the following:

Proposition 3 *Let S be as before.* $\forall w$: *If* $w \in [P([\varphi]) < r]$, *then* $P_w([\varphi] \cap [P([\varphi]) < r]) < r$

Proof Assume $w \in [P([\varphi]) < r]$. Then by definition $P_{prior}([\varphi] \cap R(w))/P_{prior}(R(w)) < r$. Hence, we also have $P_{prior}([\varphi] \cap [P([\varphi]) < r] \cap R(w))/P_{prior}(R(w)) = P_w([\varphi] \cap [P([\varphi]) < r]) < r$. □

Now we can again obtain a Fitch-like claim to the effect that, if all truths are such that we could become confident of them, then all truths are such that we are currently confident that they are indeed true. As before, the proof proceeds in the metalanguage. Then, for any structure S:

1. $\forall \varphi \forall w$: if $\models_w \varphi$, then $\exists x : P_x([\varphi]) \geq r$ — [[P Knowability 2′], for some threshold $r \neq 0$]
2. If $\models_w (\varphi \wedge \underline{[P([\varphi]) < r]}$, then $\exists x P_x([\varphi] \cap [P([\varphi] < r]) \geq r$ — [instance of [P Knowability 2′]]

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

3. If $\models_w \varphi$ and $P_w([\varphi]) < r$, then $\exists x P_x([\varphi] \cap [P([\varphi] < r]) \geq r$ [from 2]
4. If $\models_w \varphi$ and $P_w([\varphi]) < r$, then $\neg\exists x P_x([\varphi] \cap [P([\varphi] < r]) \geq r$ [by Proposition 3]
5. If $\models_w \varphi$ and $P_w([\varphi]) < r$, then $\perp$ [from 3 and 4]
6. If $\models_w \varphi$ then $P_w([\varphi]) \geq r$, [from 5]
7. If for all worlds and sentences, if $\models_w \varphi$, then $\exists x P_x([\varphi]) \geq r$, then for all worlds and sentences, if $\models_w \varphi$, then $P_w([\varphi]) \geq r$ [1–6]

In short, potential confidence entails *actual* confidence. To put it in a somewhat different terminology, the ability to acquire confirmation collapses into actual confirmation. Notice that here we are no longer dealing with probability 1—which is, arguably, a knowledge-like notion. So we are no longer dealing with a suitable translation of Fitch's result to a probabilistic realm, but with a genuine probabilistic perplexity, in Fitch's spirit.

8 The reflection principle again: Van Fraassen's integrity defense

In Sects. 3, 4 and 5 I argued that genuine quasi-Moorean statements cannot be identified with [RP Failure]. It might be contended, however, that the uneasiness we experience towards the violation of [RP] is part of the *explanation* of why certain *other* statements, such as those considered in Sects. 6 and 7, are genuine quasi-Moorean sentences. According to van Fraassen:

> [L]et us note the formal connection, at least, between Moore's paradox and the Reflection Principle. If we try to generalize Moore's sentence schema to a probabilistic form, we arrive at:
>
> It seems certain [likely, very likely] to me that: A and it seems unlikely to me that A;
>
> or, less qualitatively:
>
> It seems likely to me to degree y that (A and it seems likely to me to degree x that A): $P(A \wedge p(A) = x) = y$
>
> where the number x is lower than the number y and 'p' describes present (current) opinion. The synchronic form of the Special Reflection Principle …is violated unless y is less than or equal to x. (van Fraassen (1995), p. 19)

Clearly, '$P(A \mid P(A) = r) = r$' entails that $P(A \cap p(A) = r) \leq r$. To adjust the terminology to our current framework, if r is less than 1, attributing maximum probability to '$\varphi \wedge \underline{[P([\varphi]) = r]}$' entails that [RP] has been violated. Thus we might suggest that the paradoxicality of genuine quasi-Moorean statements lies in the fact that they yield violations of [RP]. However, *pace* van Fraassen, *we do not need [RP] to account for the Moorean blindspot*. As we have seen, '$P_w([\varphi] \cap [P([\varphi]) < 1]) = 1$' is an impossible claim in S, for any w and φ (Proposition 2); [RP] does not play any role at the time of determining this impossibility. This does not mean to say that there

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

cannot be other reasons to defend [RP], of course—but a putative connection with Moore's paradox does not seem to be one of them.

The last point can be contested on the following grounds. Ultimately, van Fraassen's contention is that satisfying [RP] is crucial to our *integrity*. Moore's paradox and [RP Failure] share in this respect the same type of 'inconsistency in a broad sense' which is pragmatic rather than semantic:

> It seems to me therefore that the correct notion of probabilistic incoherence must take its inspiration from the notion of inconsistency made manifest by Moore's paradox…. It is not inconsistent in the sense of 'unsatisfiable', 'incapable of being true', which is the semantic notion of inconsistency. But I cannot have a coherent state of opinion which I could express by a statement of the form [$\varphi \wedge \neg B\varphi$] (van Fraassen (1995), p. 27).

This is indeed an appealing defense, but the problem is that integrity can be interpreted in many ways. There is a different sense of integrity we might also feel pressed to honor—and there is a potential conflict between the two. The tension appears more forcefully once we take epistemic changes into account, as in diachronic versions of the Reflection Principle.

Let us consider once again the distinction drawn in Sect. 5 between two senses of reflection: Reflection as transparency (related to self knowledge), and reflection as a type of (probabilistic) coherence. These two senses of reflection relate in turn to two senses of integrity. According to the first one, integrity requires that we take pride in feeling accountable for our present and future actions; they are all (robustly) ours. It is precisely because we feel we ought to be so accountable that we better know who we are—where who we might become is also part of who we are.[26] It might happen that we do not approve of the way in which we foresee we will change our minds, in which case we might be able to take measures to alleviate the mistake (from our current point of view) as Ulysses did when he imagined himself facing the sirens. Identity, and responsibility within identity, overrides coherence over time. Let me call it 'Ulysses integrity'. Ulysses-type integrity can account for a temperate version of the *KK* Principle, as well as for temperate versions of [RP],[27] including also diachronic counterparts of such temperate versions, but not for the full-blown [RP]—not even for the synchronic case, as we have seen.

A second sense of integrity, by contrast, requires that we take pride in coherence. If our future self, as we foresee it, is not quite the person we believe we should become, we give up on such a future person. Diachronic coherence overrides personal identity over time (and notice that the thesis works both ways: the present self could give up on his or her past self). It is the sense of integrity that leads the idealist young man to say to his lover: "if in the future I abandon my ideals, I beg you to think that the person you now know and love does no longer exist; in such event *I* will be dead". Let me

[26] Of course, we cannot demand knowledge of the future on rationality grounds. What can be demanded, however, is that we take active steps to be able to make accurate predictions about our future temporal slices.

[27] Cf. Cresto (2012).

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

call it 'Parfit integrity'.[28] A consequence of Parfit integrity is that I consider a future person to be identical with me only to the extent that I can have trust in that future person's probability assignment. Parfit-type integrity can then account for diachronic versions of [RP], and, by extension, for their synchronic, more restricted versions.

In short, according to van Fraassen Moore's paradox exhibits a type of 'inconsistency in a broad sense' which is also the type of inconsistency we can identify in some cases of integrity failure—what I have called 'Parfit integrity'. However, there are other types of integrity failure we might worry about. Since Parfit integrity is not all we demand from agents, it is not clear whether [RP] is a principle we should always enforce; at any rate, not out of Moorean-type considerations.

9 Beyond epistemic paradoxes. Motivations for hierarchic languages

One of the main goals of this paper was to explore the structure of successful quasi-Moorean statements, and, concomitantly, to build a probabilistic version of the Knowability Paradox. The results obtained helped us draw a number of morals that go well beyond Fitch's argument. We have shown that violations of [RP] can be attributed maximal evidential probability, so [RP Failure] is not itself a quasi-Moorean sentence: the Moorean spirit has been 'lost in translation'. More generally, the 'integrity defense', which seeks to compare violations of [RP] with Moorean-type irrationality, should be taken *cum grano salis*, since [RP] may actually conflict with other senses of integrity. Of course, there might well be alternative strategies to show the putative illegitimacy of [RP Failure]—Dutch Books, calibration arguments, etc. But it should be clear by now that at least some lines of argument will not do.

In addition, we have confirmed that the chosen formal framework maintains a healthy symmetry between knowledge attribution and the attribution of maximal evidential probability. Of course, symmetry as I understand it here is just a *necessary* condition for a satisfactory framework, but it is not sufficient. Moreover, if we take the symmetry criterion seriously, any adequate *solution* to the paradox should be unified as well. In order to do so, a more sophisticated setting might be desirable. In what follows I will outline such a setting, without entering into the details. As we will see, it leads naturally to a unified solution to both strands of Fitch's paradox. Given the symmetry criterion, this fact can be turned into indirect evidence in favor of the formalism.

The system I favor is largely based upon the one we have been working with so far, but assumes a hierarchy of K and P operators, together with certain restrictive rules for building well-formed formulas. The motivation for this proposal goes as follows.

To begin with, a case can be made for the claim that second order probabilities demand that we conditionalize on second order evidence. Suppose we have information about the state of the weather tomorrow. We have read the forecast in the newspaper, watched the weather channel, etc. On the basis of all this information, we conclude that the probability of rain tomorrow in our city is 0.3. Now suppose a friend asks us

[28] See Parfit (1973), pp. 145–6. I do not mean to say that Parfit himself supports what I dubbed 'Parfit Integrity', but just that his description of what he calls 'The Complex View' comes close to capturing what I have in mind. Parfit's example is re-elaborated in Elster (1984), Part II.

Journal: 11229-SYNT Article No.: 0884 TYPESET ☑ DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

how probable it is that our rational degree of belief that there is rain tomorrow is in fact 0.3. As I see it, in this case our friend is no longer interested in the probability of a proposition about meteorology, but in the probability of a proposition *about the degree of confirmation* possessed by our original meteorological statement. Which is the relevant evidence to answer this question, then? Intuitively, what we have to assess is how good we are at the time of engaging in confirmation theory. Thus the relevant total evidence is no longer $R(w)$: the evidence for our second-order probability should consist in what we know about our capabilities to adequately confirm propositions; the strongest proposition that expresses this idea is in fact $KR(w)$.[29] Hence when we calculate a second order evidential probability we should conditionalize on $KR(w)$. The proposal then generalizes to increasingly higher levels.[30]

Consider now a metalinguistic statement such as '$P_w([P([\varphi]) = r])$'. As we have seen, this statement is meant to capture the intuition that we are calculating a *second* order probability. However, the argument of the probability function is just a set of worlds; we could have well referred to it by other means—for example, by means of a suitable sentence of L without epistemic operators (say, '$[\psi]$'), as in a regular first order probability. Therefore, propositions understood as sets of worlds seem to be too coarse grained for what we want.

A possible suggestion at this point is to let the arguments of our probability functions be *sentences* of well-regimented languages; we can define a sequence of languages $L^0, L^1 \ldots L^n \ldots$, with probability operators $\underline{P^0}, \underline{P^1} \ldots \underline{P^n} \ldots$ that apply to sentences of lower lever languages. Thus, the probability of a set of worlds will depend crucially on the way we refer to it. In other words, at the time of calculating evidential probabilities, the 'mode of presentation' matters. For structural reasons of internal coherence, we should also demand a corresponding sequence of knowledge operators $K^0, K^1 \ldots K^n \ldots$, each with its own accessibility relation.[31] It can be shown that, if we choose our accessibility relations carefully—so as to make sure that knowledge and probability attributions cohere with each other—the model can validate a moderate version of the *KK* Principle: to wit, if the agent has first order knowledge that φ, then she has *second order knowledge* that she has first order knowledge that φ ('$K^1\varphi \rightarrow K^2K^1\varphi$'). Such a moderate version of *KK* can be satisfied without actually demanding that any of the accessibility relations be transitive. Hence the model can

[29] By definition, $KR(w) = \{x \in W\colon \text{if } xRy, \text{ then } y \in R(w), \text{ for all } y\}$. Hence $KR(w) \subseteq R(w)$.

[30] It might be contended that agents need not be aware of '$KR(w)$'—in which case they would not know which proposition they should conditionalize on (thanks to an anonymous referee for giving me the opportunity to clarify this point.). However, if there is a problem here, it is not exclusive of the enhanced framework, as similar considerations can be make for the standard setting; to wit, it might be contended that agents need not be aware of '$R(w)$' either. There are at least two ways out, which relate to two very different interpretations of the formalism. On one hand, we can conceive of the framework as a tool for the theoretician (or the interpreter), who seeks to make knowledge and probability attributions from a third person point of view. She is the one who assesses, to the best of her knowledge, what the agent knows or ignores in each possible situation. On the other hand, we can think of the framework as 'viewed' from the inside, as it were, *i.e.*, as structured from the first person perspective. In this case we can take '$R(w)$' to refer to the information the agent has consciously gathered. '$KR(w)$' could then capture the subset of $R(w)$ which the agent takes to be the result of extremely reliable research methods, among other possibilities.

[31] In a nutshell, by having the right sequence of knowledge operators we guarantee that statements with evidential probability 1 will be known by the agent.

Journal: 11229-SYNT Article No.: 0884 TYPESET ☑ DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

still be sensitive to vagueness related considerations, as discussed in Sect. 5. All we need is the weaker assumption that relation R^{i+1} composed with R^i is included in R^i.

Within this enriched setting we are no longer able to express Moorean or quasi-Moorean statements in our sequence of languages.[32] Expressions such as '$(*) K^2(\varphi \wedge \neg K^1\varphi)$', as well as their probabilistic counterparts, are not well-formed formulas to begin with. Therefore, the two versions of the Knowability Paradox dissolve. Unlike other attempts to address Fitch's paradox, in this case the syntactic restrictions on well-formed formulas respond to principled reasons that are completely independent of the discussion on Knowability. Once the restrictions are in place, however, we obtain a unified answer to Fitch-like paradoxes as a nice side effect.

To sum up, the fact that we were able to develop a probabilistic Fitch-like argument in the simpler setting tells us that it was a setting worth exploring; the fact that we could not find straightforward solutions to the paradoxes within that framework suggests that a more sophisticated structure might be welcome. Finally, the fact that we can obtain a unified solution to the paradox within a richer framework, which has been originally developed for independent reasons, reinforces the thought that we are on the right track.

10 Concluding remarks

Along these pages I presented genuine examples of quasi-Moorean statements, and I used them to build a probabilistic version of the Knowability Paradox. I also showed that violations of [RP] do not share the relevant traits of Moorean-type irrationality: Moorean and quasi-Moorean statements are true blind spots, whereas [RP Failure] is not. Finally, I outlined a formal framework within which both strands of the paradox received a uniform solution.

Let me address a few final concerns on the very structure and goals of this paper. As I have already anticipated in the Introduction, someone might object that I have assumed all along that the relevance of Fitch's paradox lies in the fact that it reveals a worrisome modal collapse between possible and actual knowledge,[33] and, analogously, between possible and actual evidential probability. But it is far from clear whether this is the right description of the problem.

The objection, however, would be misguided. All I have required is an endorsement of what I dubbed 'the symmetry criterion' between knowledge and probability. Whatever it is that we deem problematic with Fitch's result, we should obtain a similarly problematic result within a probabilistic framework. Hence, *if* Fitch's original argument involves a modal collapse of some sort, there should be an analogous modal collapse within a probabilistic version of the argument, with the aid of a quasi-Moorean statement. But this paper is neutral concerning the existence of such a modal collapse.

[32] The proposed solution shares a family resemblance with other attempts to solve Fitch's paradox with the aid of typed languages, such as Linsky (1986), Linsky (2009), or Paseau (2008). However, the sequence of languages that I have in mind is more restrictive than usual hierarchic proposals, in the sense that 'K^i' is only meant to apply to sentences of the form '$K^{i-1}\varphi$' or their negations; analogous restrictions apply to higher order probability statements. A rationale for this demand can be found in Cresto (2012).

[33] As suggested by Kvanvig (2006).

Journal: 11229-SYNT Article No.: 0884 TYPESET DISK LE CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

Actually, although my own diagnosis is that all versions of Moore's and Fitch's Paradox arise out of a confusion between levels of operators, the probabilistic version that I have offered here is compatible with many different interpretations of the source of the problem. Thus, for example, my probabilistic reconstruction of the Knowability argument is perfectly compatible with the claim that true verificationists should switch to intuitionistic logic in order to block the potential *reductio* of the Knowability Principle.[34] Alternatively, if we are convinced that there are good reasons to demand that the Knowability Principle be restricted to Cartesian Propositions,[35] we should equally demand that [P Knowability 2] be only instantiated by sentences that can be assigned probability 1 without contradiction.

On the other hand, it is not clear how a request to restrict the Knowability Principle to actual truths, as in Dorothy Edgington's proposal,[36] could work in the probabilistic scenario of Sects. 2, 3, 4, 5, 6, 7 and 8. Consider the following amendment to [P Knowability 2],[37] as a possible way of capturing the idea that only actual truths are knowable:

$$\forall\varphi\forall w: \text{ if } \models_w \varphi, \text{ then } \exists x: P_x(\{w\}) > 0, \text{ and } P_x([\varphi]) = 1 \qquad \textbf{[P Knowability A]}$$

(or, equivalently, '$\forall\varphi\forall w$: if $\models_w \varphi$, then $\exists x : w \in R(x) \subseteq [\varphi]$'). We can see that the paradox still runs:

4′. *If* $\models_w (\varphi \wedge \underline{[P([\varphi]) < 1]})$, then $\exists x : P_x(\{w\}) > 0$ and $P_x([\varphi] \cap [P([\varphi]) < 1]) = 1$ [Instance of [P Knowability A]]

As '$P_x([\varphi] \cap [P([\varphi]) < 1]) = 1$' is an impossible claim, we still obtain:

5′. If $\models_w \varphi$, then $P_w([\varphi]) = 1$

In other words, if there is no world x such that '$[\varphi] \cap [P([\varphi]) < 1]$' can be included in $R(x)$, then the same is true for whichever world we can pick out in $R(x)$ as the actual world. So the rest of the argument follows without changes. We can take this result either as evidence that Edgington's amendment is not effective in this sort of probabilistic setting, or as a motivation to find a different translation for Edgington's intuition. I am not going to take a stance on this point here.

In any case, I submit that, once we adopt the symmetry criterion, finding a unified solution for both strands of Fitch's paradox is mandatory, if some solution is offered at all. In this sense, the symmetry criterion can be used not only to test the adequacy of a given probabilistic setting, but also to test the adequacy of different solutions to the knowledge version of the paradox. By itself, the criterion will not succeed in singling out a best answer, though some proposals can be ruled out as formally inadequate. This leaves us the interesting task of revising well known proposals in a systematic manner to see whether they pass the test. For the moment, we have learnt that there are formal refinements of Kripke settings that work just fine.

[34] Williamson (1982), Dummett (2009).

[35] Cf. Tennant (1997).

[36] Edgington (1985), Edgington (2010). See also Rabinowicz and Segerberg (1994).

[37] Thanks are due to Wlodek Rabinowicz for this suggestion.

Journal: 11229-SYNT Article No.: 0884 ☐ TYPESET ☑ DISK ☐ LE ☐ CP Disp.:2015/9/14 Pages: 23 Layout: Small-X

Acknowledgments Previous versions of this paper have been presented at the *II ALFAN Congress* (The Latin American Analytic Philosophy Association, Buenos Aires, August 2012), at the *Conference on Probability and Vagueness* (The University of Tokyo, Tokyo, March 2013), and at the *First Oxford-Buenos Aires Colloquium* (The Argentine Society for Analytic Philosophy, Buenos Aires, April 2013). I am grateful to the participants for their questions and comments. In particular, I want to thank Richard Dietz, Paulo Faria, and two anonymous referees for *Synthese*. Special thanks are also due to Ramiro Caso for his invaluable help at the time of editing the final version of this document.

References

Beall, J. (2009). Knowability and possible epistemic oddities. In J. Salerno (Ed.), *New essays on the knowability paradox* (pp. 105–125). Oxford: Oxford University Press.
Cresto, E. (2012). A defense of temperate epistemic transparency. *The Journal of Philosophical Logic*, *41*, 923–955.
Dummett, M. (2009). Fitch's paradox of knowability. In J. Salerno (Ed.), *New essays on the knowability paradox*. Oxford: Oxford University Press.
Edgington, D. (1985). The paradox of knowability. *Mind*, *94*, 557–568.
Edgington, D. (2010). Possible knowledge of unknown truth. *Synthese*, *173*, 41–52.
Elster, J. (1984). *Ulysses and the Sirens: Studies in rationality and irrationality*. Cambridge: Cambridge University Press.
Fitch, F. (1963). A logical analysis of some value concepts. *The Journal of Symbolic Logic*, *28*, 135–142. Reprinted in *New essays on the knowability paradox*, pp. 21–28, by J. Salerno, Ed., 2009, Oxford: Oxford University Press.
Halpern, J. (2003). *Reasoning about uncertainty*. Cambridge MA: MIT Press.
Hart, W., & McGinn, C. (1976). Knowledge and necessity. *Journal of Philosophical Logic*, *5*, 205–208.
Kvanvig, J. (2006). *The knowability paradox*. Oxford: Oxford University Press.
Linsky, B. (1986). Factives, blindspots and some paradoxes. *Analysis*, *46*, 10–15.
Linsky, B. (2009). Logical types in some arguments about knowability and belief. In J. Salerno (Ed.), *New essays on the knowability paradox* (pp. 163–179). Oxford: Oxford University Press.
Moore, G. (1993). Moore's paradox. In T. Baldwin (Ed.), *G. E. Moore: Selected writings* (pp. 207–212). London: Routledge.
Parfit, D. (1973). Later selves and moral principles. In A. Montefiore (Ed.), *Philosophy and personal relations*. London: Routledge and Kegan Paul.
Paseau, A. (2008). Fitch's argument and typing knowledge. *Notre Dame Journal of Formal Logic*, *49*, 155–176.
Rabinowicz, W., & Segerberg, K. (1994). Actual truth, possible knowledge. *Topoi*, *13*, 101–115.
Salerno, J. (2009). Knowability noir: 1945–1963. In J. Salerno (Ed.), *New essays on the knowability paradox* (pp. 29–48). Oxford: Oxford University Press.
Tennant, N. (1997). *The tamning of the true*. Oxford: Clarendon Press.
van Benthem, J. (2004). What one may come to know. *Analysis*, *64*, 95–105.
van Fraassen, B. (1995). Belief and the problem of Ulysses and the Sirens. *Philosophical Studies*, *77*, 7–37.
Williamson, T. (1982). Intuitionism disproved? *Analysis*, *47*, 154–158.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Williamson, T. (2014). Very improbable knowing. *Erkenntnis*, *79*, 971–999.

Journal: **11229-SYNT** Article No.: **0884** ☐ TYPESET ☑ DISK ☐ LE ☐ CP Disp.:**2015/9/14** Pages: **23** Layout: **Small-X**