# Metacognitive perspectives on unawareness and uncertainty

Paul Egré* Denis Bonnay

**Abstract**

A state of ignorance about a proposition can result from two distinct sources: *uncertainty* about what the correct answer actually is, and *unawareness* of what the answer might possibly be. Uncertainty concerns the strength of one's evidence, whereas unawareness concerns the conceptual components needed to articulate a proposition. This chapter discusses the implications of the distinction between uncertainty and unawareness for metacognition, and more specifically for the problem of what it takes to know that one knows and to know that one does not know. In particular, we relate the distinction between uncertainty-based unknowns and unawareness-based unknowns to the two-stage model proposed by Glucksberg and McCloskey for decisions about ignorance.

## 1 Introduction

What does it take to know that one does not know something, or to know that one knows? In epistemic logic, the ability to know that one does not know something whenever such is the case is referred to as 'negative introspection', and distinguished from the property of 'positive introspection', which concerns the ability to know that one knows whenever one knows (see Fagin et al. 1995).

Negative introspection is usually seen as a very demanding condition on knowledge, for two reasons: one reason, which will be examined more carefully below, is that we often fail to know that we don't know $p$ simply because we do not have access to the basic ingredients of the proposition $p$. Another reason is that systematic knowledge that one does not know would prevent us from believing that one knows in all cases in which we have false

*Corresponding author: paul.egre@ens.fr

beliefs: suppose that I falsely believe $p$, and therefore that I fail to know that $p$; in order to know that I don't know $p$ for a fact, I would need to believe that I don't know $p$, and so I would have to believe both that I know $p$ and that I don't know $p$, a contradiction.

There is a more active debate in the philosophy of knowledge regarding whether positive introspection should be seen as an analytic property of our knowledge. On the Cartesian view of knowledge, to know a proposition normatively requires us to be able to know that we know. This view was defended by Hintikka in particular and is considered as characteristic of epistemological internalism. Hintikka's argument rests on the idea that if knowing implies having a conclusive justification for one's belief, then one should thereby be in a position to know that the justification is conclusive (see Hintikka 1970, and Hemp 2006 for an overview). One criticism of Hintikka's analytic conception of the relation between knowing and knowing that one knows has been that knowing that one knows would require one to possess the concept of knowledge. This would deny knowledge to animals and infants, who presumably do not have the concept of knowledge. This, however, presupposes that knowing that one knows necessarily involves a metarepresentational attitude towards one first-order knowledge, an assumption that is far from obvious in the light of the idea that metacognitive processes need not always involve an abstract representation of the concept of knowledge (see Proust 2007 for a discussion). Another externalist attack on the analytic connection between knowing and knowing that one knows comes from considerations about the structure of justification. If to know a proposition is to have a good justification for it, then one may know without always knowing that one knows, for otherwise the risk is that of an infinite regress in the justifications. This, in a nutshell, is the gist of epistemological externalism about knowledge.

In this paper we shall not focus on normative considerations about the well-foundedness of either positive and negative introspection, in part because we contributed to that debate elsewhere (see in particular Egré 2008, Bonnay and Egré 2009, Bonnay and Egré 2011). Rather, the aim of this chapter will be to discuss some of the constraints on one's abilities to know that one knows and to know that one does not know in relation both to epistemology and to psychology. Obviously, there are many ways in which we can fail to realize our ignorance. On the other hand, there are also situations in which we have a clear access to our knowledge as well as to our ignorance. What makes the difference between those? Part of the present contribution will be an effort to interpret actual psychological data and relate them to more abstract models of knowledge and uncertainty (as used in epistemic

logic; see Egré and Bonnay 2010 for an example; in this chapter, we shall remain deliberately informal as far as possible).

Our leading thread in this discussion will be the distinction made in formal epistemology between two forms of ignorance, namely uncertainty and unawareness. Fundamentally, unawareness can be defined as a form of ignorance resulting from the lack of conceptual or representational resources needed to articulate a proposition. Uncertainty, on the other hand, concerns the lack of evidence needed to adjudicate the truth or falsity of a proposition that one can represent and articulate. Uncertainty, most of the time, is conscious, while unawareness, by definition, is unconscious. Our proposal is to examine the implications of the distinction between uncertainty and unawareness for metacognition. What we shall argue is that knowing that one knows or that one does not know is typically harder and less reliable in situations that require us to evaluate the strength of one's uncertainty. In contrast to that, knowing that one does not know is easier and more reliable for unknowns grounded in antecedent unawareness.

The chapter is structured as follows. In section 2, we introduce the distinction between uncertainty and unawareness and review the main differences between them. In section 3 and 4, we use the distinction to classify different ways in which one may fail to know that one knows or that one does not know a proposition. In section 5, we proceed to the discussion of two sets of experimental data concerning metacognition: experiments by Glucksberg and McCloskey 1981, and more recently by Hampton and colleagues, concerning the evaluation of one's ignorance, and experiments by Smith et al. concerning the monitoring of one's uncertainty. Consistently with the model of Glucksberg and McCloskey, we shall argue that appreciating the strength of one's evidence and appreciating the availability of specific conceptual resources in memory likely involve different mechanisms. More generally, we argue that the distinction between uncertainty-based unknows and unawareness-based unknowns can be subsumed under Glucksberg and McCloskey's two-stage model for decisions about ignorance. In section 6, finally, we focus on higher-order knowledge about one's uncertainty and discuss the case for an asymmetry between knowing that one knows and knowing that one does not know.

## 2 Uncertainty and unawareness

A fruitful way to approach the definition of metacognitive abilities such as knowing that one knows or that one does not know is to start by an

examination of the notion of ignorance. The object of this section is to argue that ignorance, understood as the failure to know some proposition, results from two importantly distinct sources, which are called *uncertainty* and *unawareness* in the epistemological literature.

## 2.1 Two forms of ignorance

While the notion of uncertainty has been at the center of epistemic logic since its inception, the clarification of the concept of unawareness is much more recent (see in particular Franke and de Jager 2008 for an excellent exposition and overview). As we shall see, uncertainty and unawareness themselves come in different varieties. However, the main opposition we can draw between them concerns the extent to which either form of ignorance is accessible to consciousness.

Let us consider uncertainty first. Suppose I am playing a version of the Monty Hall game. I am faced with two doors labelled A and B. Behind one of those, there is a goat, and behind the other there is a car. To win the game is to open the door with a car behind it. The quizmaster informs me of the situation and asks me which door I want to open. Clearly, I am in a state in which I do not know whether the car is behind door A or behind door B. In this case my ignorance about whether the car is behind door A or behind door B results from my incapacity to discriminate between the two doors. I entertain two possibilities about the true state of the world: that the car is behind door A and that the car is behind door B. The fact that these two possibilities are equally open to me is what we call *uncertainty*.

Contrast the previous situation with the following. Suppose I never heard of J.R.R. Tolkien, the author of *The Hobbit*, nor did I come across any of his books. In that situation I am ignorant of a number of facts about Tolkien besides his existence. In particular, I fail to know that Tolkien is the author of *The Hobbit*. My ignorance in that case is quite different from my ignorance in the previous case. The situation is not one in which I am uncertain as to who the author of *the Hobbit* might be. In particular, it is quite different from a situation in which I might have read *The Hobbit* and come across the name of Tolkien several times before, but in which, asked about who the author is, I would hesitate between J.R.R. Tolkien and C.S. Lewis. In the former situation, as opposed to the latter, I do not have the wherewithal to even represent the proposition that Tolkien is the author of the *Hobbit*. In a case like this I am simply *unaware* that Tolkien wrote *The Hobbit*, because I do not have the ingredients needed to entertain that proposition.

In the words of Heifetz et al. (2006), "unawareness refers to lack of conception rather than to lack of information". *Lack of conception* corresponds to a state in which we cannot verbally or conceptually articulate a possibility. Lack of conception can mean different things, however. The most radical form is what we may call *lack of acquaintance* with the concepts, when we do not even have the ingredients available in memory to articulate the proposition. This corresponds to the example we just discussed. In a lot of cases, however, lack of conception can result from *inattentiveness*, as discussed by Franke and de Jager, namely when the conceptual resources are available in memory, but when we are temporarily blind to them. Franke and de Jager, for instance, give the example of someone looking for his keys, and temporarily failing to even represent the possibility that the keys might be in his car. On their account, this is not lack of acquaintance with the car or the concept of the car, but temporary blindness to the possibility that the keys might be in the car, due to a temporary failure to activate the representation of the car in one's memory. As Franke and de Jager put it, the subject then would be able to utter truths like: "the keys are either on the desk or not on the desk", but would not be able to say in the same way: "the keys are in the car or not in the car".

This form of inattentiveness, finally, should be distinguished from a more common form of so-called unawareness, for cases in which we do have the conceptual resources available in memory, can activate them, but simply discard them as irrelevant. Such cases are certainly typical of most of our false beliefs. For a long time, for instance, one of us used to assume that teaine and caffeine were two chemically different substances. Later, he was told that they were the same molecule, and came to revise his earlier belief. So it could be said that he was *unaware* that teaine and caffeine were the same substance until he was told, not because he could not conceptually articulate the possibility that they were the same substance, but because that was a possibility he was failing to entertain as open. In the words of Franke and de Jager, he was *assuming* that teaine and caffeine were different substances and this blind assumption could be construed as some form of unawareness.

However, the situation is very different from those involving the two kinds of unawareness which were previously discussed. Lack of conception is symmetric. If I cannot represent the proposition that Tolkien is the author of *The Hobbit*, I cannot represent either the proposition that Tolkien is not the author of *The Hobbit*. Similarly, if I overlook the possibility that the keys might be in the car, I do not consider the possibility that the keys might not be in the car as a salient possibility (that is to say, I will not be

searching places *qua* places that are not places in the car). By contrast, the teaine *vs* caffeine example is asymmetric. I am overlooking the possibility that they are same but I am precisely not overlooking the possibility that they are different. From now on, we shall reserve the label 'unawareness' to symmetric cases, which stem from lack of conception or from inattentiveness.

## 2.2 Main differences

The difference between uncertainty and unawareness can now be characterized more abstractly. In ordinary ascriptions of knowledge, first of all, note that we say of someone that they do not know *whether* $p$, or that they are uncertain about *whether* $p$. By contrast, we report situations of unawareness by saying of someone that they do not know *that* $p$, or that they are unaware *that* $p$ is true. To fail to know whether or not $p$ is to entertain $p$ as well as its negation as two open possibilities. By contrast, cases of unawareness are cases in which the fact that $p$ is not entertained as a possibility, and in which the contradictory alternative, consequently, is not even represented in the agent's mind.[1]

Secondly, uncertainty and unawareness are resolved in different ways. Uncertainty reduces, and knowledge increases, as possibilities gradually get eliminated. Consider the Monty Hall game again, and suppose that I randomly pick out door A. The quizmaster opens it, unfortunately I see a goat behind it. Given my new evidence, however, I now eliminate the possibility that the car is behind door A, from which I can infer that it is behind door B. An increase in awareness, by contrast, is not adequately pictured as the narrowing down of a set of epistemic possibilities. Intuitively, it corresponds to the opposite. In a case in which I already have the conceptual resources but fail to attend to $p$ as a possibility, becoming aware means expanding the set of possibilities initially thought to be relevant. For instance, my becoming aware that teaine and caffeine were the same substance implied the consideration of a possibility previously excluded by my belief. In cases in which one is unaware of a proposition due to lack of the conceptual resources necessary to articulate the proposition, becoming aware will not quite mean expanding the set of possibilities. Rather, it involves adding structure to the space of possibilities. For instance, if I had never heard of Tolkien nor of the novel *The Hobbit*, and come across both names, I acquire the capacity to ask a new question such as: "Is Tolkien the author of *The Hobbit*?". I can thereby divide the space of logical possibilities by means of a division

[1] For more on the distinction between 'knowing that' and 'knowing whether' constructions, see Aloni et al. forthcoming.

that was previously unavailable to me (see Bromberger 1987/1992 on what it takes to articulate one's ignorance of a proposition, and Pérez Carballo 2010 for a recent account generalizing on that idea; we refer to the next section for an example).

The third and most relevant difference between unawareness and uncertainty, finally, concerns the status of consciousness in relation to epistemic possibilities. In a state of uncertainty, an agent is consciously entertaining possibilities as open, and consciously trying to get more information in order to reduce that uncertainty. In a state of unawareness, by definition, the agent cannot be conscious of the possibilities he or she is failing to take into account. The distinction between conscious and unconscious possibilities has a metacognitive import. As Franke and de Jager put it, unlike uncertainty, "unawareness is not introspective" (see Dekel, Lipman and Rustichini 1998 for a formal account of unawareness based on this important observation). This means that whereas one can know that one experiences uncertainty simultaneously with that uncertainty, one cannot know that one is unaware of a proposition at the moment one is unaware of it.[2] Knowing that one knows and knowing that one does not know are thus likely to obey different constraints depending on whether one's ignorance is a matter of uncertainty or of unawareness.

# 3 Unconscious ignorance and implicit knowledge

This section reviews some of the ways in which a state of unawareness precludes knowing that one does not know, but also knowing that one knows. In both cases, becoming aware implies a transition from implicit to explicit uncertainty, or relatedly, from implicit to explicit knowledge.

## 3.1 Unconscious ignorance

A situation in which an agent lacks the basic concepts necessary to articulate a proposition will necessarily prevent her from being conscious that she does not know a proposition at the moment she does not know it. Someone who never heard of Tolkien and *The Hobbit* cannot know that she does not know

[2] We do not rule out the possibility of states of unconscious uncertainty, but the point is that being uncertain is compatible with the consciousness of that state, whereas a state of unawareness is incompatible with the simultaneous consciousness of one's unawareness. In other words, I can be conscious that I *am* uncertain, whereas I can only realize that I *was* unaware. For more on what it means to dynamically realize that one was ignorant, see van Benthem (2004) and Bonnay and Egré (2011).

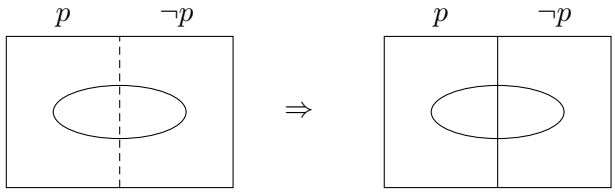


Figure 1: From implicit to explicit uncertainty

that Tolkien is the author of the Hobbit. However, if asked the question: "who is the author of the Hobbit?", or "is Tolkien the author of the Hobbit?", the mention of these names can be sufficient to trigger a change of state in the agent, from unawareness to uncertainty. This means that, at the moment the agent is asked the question, the agent's unawareness disappears, and the agent is now in a position to ask all sorts of questions about Tolkien. To the extent that the agent accommodates the information that expressions like "Tolkien" or "The Hobbit" have a reference, and that they belong to the expected referential categories ('Tolkien' refers to a person, 'the Hobbit' to some work of art) the agent is ipso facto in a position to know that she does not know whether or not Tolkien is the author of *The Hobbit*.

A convenient way of picturing the transition from unconscious to conscious ignorance is as follows. Let $p$ stand for the sentence "Tolkien is the author of the Hobbit", and $\neg p$ for its negation. Unawareness can be represented by the agent's incapacity to *delineate* between $p$ and non-$p$ possibilities. In the left figure below, the rectangle represents the space of conceptual possibilities, divided between $p$ and $\neg p$ as distinct possibilities. The absence of conscious delineation is represented by a dashed line between $p$ and non $p$ regions. Uncertainty, on the other hand, results from the agent's limited capacity to *discriminate* between possibilities. On both figures, this uncertainty is represented by the fact that the agent's information state, the set of possibilities available in principle to the agent on the basis of her evidence, overlaps on $p$ and $\neg p$ regions. In the lefthand diagram, uncertainty is only implicit, however. The transition from implicit to explicit uncertainty corresponds to the replacement of a dotted line by a solid line between $p$ and non-$p$ possibilities. The agent's information state is such that, once the agent gains awareness of the proposition at issue ("is Tolkien the author of *The Hobbit*?), she is in a state of conscious uncertainty about the answer to that question.

### 3.2 Implicit knowledge

From this diagram we can account for dual cases, in which the agent lacks the conceptual resources to even figure out a particular question, but such that, if she were given the appropriate concepts, she would correctly discriminate between right and wrong answers. One speaks of *tacit* or *implicit* knowledge for such cases. A good example of such tacit knowledge is linguistic knowledge. In the tradition of generative grammar, most of our linguistic competence is characterized as a form of tacit knowledge of regularities about sentence formation. Halle in particular describes phonological knowledge as "knowledge untaught and unlearned" (Halle 1978). To illustrate it, he gives the example of phonological rules that competent speakers of English master without difficulty, but are unaware of, such as plural formation in English. There are three kinds of morphophonological realization of the plural in English, namely plurals in -iz- as in *buses*, plurals in -s- as in *cats*, and plurals in -z- as in *dogs*. A regularity about the choice between these plural suffixes is for instance that: "if a noun ends with a sound that is non-voiced, the plural is formed with -s-". For instance, the noun "dog" ends with a voiced stop, whereas the noun "cat" ends with a non-voiced stop. The knowledge of such a generalization is obviously implicit in most speakers, simply because the concepts of a stop or of a voiced consonant need not be available.

If an agent were given those concepts, she may still not be able to thereby state the generalization of course. However, consider a simpler instance of this generalization. There is obviously a sense in which every competent speaker of English knows that *the plural of "cat" is "cats" rather than "catz" or "catiz"*. Most speakers of English will be unaware of this proposition, simply for failing to attend to the possibility that the plural of "cat" might have been "catz" or "catiz". When asked, however, they would obviously respond correctly to the question: "is the plural of "cat" "cats", "catz", or "catiz"?". There is a sense, therefore, in which competent speakers of English know implicitly that the plural of "cat" is not "catz", but are unaware of this fact, and are unaware that they know it. When asked explicitly, however, they are put in a position to correctly eliminate the possibility that the plural of "cat" is "catz".[3] In Figure 2, the lefthand-side diagram represents a situation in which the agent's informational state implies such a proposition $p$, but in which the agent cannot initially delineate between $p$ and $\neg p$ possibil-

[3]See Schaffer (2008) for a discussion of the epistemological implications of contrastive knowledge attributions of the form 'knowing that $p$ rather than $q$'. See Aloni and Egré (2010) for a discussion of Schaffer's view on epistemological contrastivism.

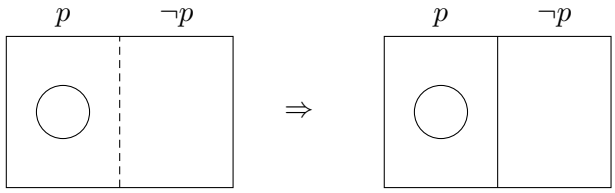


Figure 2: From implicit to explicit knowledge

ities. Once the delineation is made, however, the informational state of the agent puts them in a position to explicitly know that $p$, and also to know that they know.

The example of linguistic knowledge given here to illustrate this transition from implicit to explicit knowledge may still raise some questions. It may be argued that linguistic competence is a form of knowing-how, irreducible to the knowledge of a body of propositions. This view is not uncontroversial (see in particular Stanley and Williamson 2001 for a defense of knowing-how as propositional knowledge). Even if we grant it, however, we may still be able to find examples of explicit propositional knowledge that fit exactly the sort of transition intended by Figure 2. Maybe the oldest example is Plato's discussion, in *Meno*, of what it takes for a young child to know that the length of the diagonal of the square is not commensurable with the length of the sides. Plato's argument, in a nutshell, is that the child could not come to the knowledge of that proposition if he did not know it previously. One way to make sense of Plato's views would be in the terms of Figure 2: the child has all the discriminative resources to recognize that that proposition is true. What the child misses, however, are the concepts and intermediate constructions that allow him to delineate logical space in a way that will allow him to identify the relevant state of affairs (in the dialogue, Socrates teaches the concepts and constructions to the young child). A recent account of mathematical knowledge along exactly those lines is elaborated in the work of Pérez Carballo (2010), based on the discussion of other actual examples of mathematical discoveries.

## 4 Underconfidence and overconfidence

Cases of unawareness of the kind discussed in the previous section, in which we fail to realize what we know and what don't know, cannot be characterized as cases of *misrepresentation* of one's knowledge, but rather, as cases of *unrepresentation* of the structural components of a proposition. In this

section we move to cases in which negative and positive introspection fail not because of unawareness, but because of a misrepresentation of the structure of one's first-order uncertainty. This corresponds to cases in which an agent has a wrong appreciation of her discrimination capacities, or of what her evidence allows her to conclude. Situations of that kind are generally described as cases of underconfidence or of overconfidence.

### 4.1 Overconfidence

Consider overconfidence first. Many are the occurrences in which an agent holds an incorrect belief about whether $p$, yet represents herself as knowing $p$. For instance, at the time I used to believe that teaine and caffeine were different molecules, my knowledge of chemistry did not actually allow me to rule out that they were the same. Yet I represented my evidence as conclusive enough to rule out this possibility. The situation is depicted on the lefthand diagram of Figure 3 below. The $p$ area represents the proposition that teaine and caffeine are different, and $\neg p$ the proposition that they are the same. The circle on that figure represents the possibilities among which I am able to discriminate on the basis of my actual knowledge of chemistry. The ellipsis, on the other hand, gives the representation I have of my evidence. In this case, I believe that I know $p$, because the ellipsis is included in the set of $p$ possibilities. This represents a case in which I have a wrong appreciation of my actual evidence: I underestimate some possibilities, and overestimate others. Incidentally, this diagram also represents one form of unawareness, distinct from what we characterized as 'lack of conception'. That is, it represents a situation in which I am able to articulate the difference between $p$ and $\neg p$ possibilities and have them in my conceptual apparatus (as materialized by the solid line between $p$ and $\neg p$), but in which I mistakenly fail to entertain $\neg p$ possibilities as open.

### 4.2 Underconfidence

Cases of underconfidence are exactly symmetric. In a situation of underconfidence, I am in a position to know $p$, but I don't adequately represent myself as knowing $p$. This is represented on the righthand diagram of Figure 3. This would be a situation in which I am in a position to exclude that teaine and caffeine are different substances, but in which I misrepresent my actual evidence, and believe that I don't know whether teaine and caffeine are the same or not. From a psychological point of view, underconfidence may have many different sources. From an epistemological point of view, it

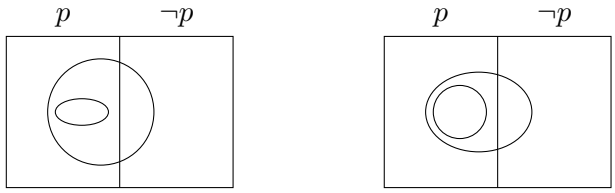


Figure 3: Overconfidence and underconfidence: circle = actual first-order uncertainty ; ellipsis = representation of one's first-order uncertainty

implies considering irrelevant possibilities as relevant.

Note that agents with a perfect capacity to evaluate their actual first-order uncertainty would be agents for whom the regions delimited by circle and ellipsis would exactly coincide. By way of consequence, the lack of coincidence here between the two levels gives counterexamples to the principles of positive and negative introspection. On the lefthand diagram, the agent does not know $p$, but believes she knows $p$. If our agent is consistent, this implies that she does not know that she does not know (for knowing that one does not know would imply believing that one does not know). On the righthand diagram, the agent knows $p$, but believes she does not know $p$. This implies that she does not know that she knows in at least one sense, the sense in which knowing that one knows would imply believing that one knows.[4]

Inclusion or even overlap between ellipsis and circle as in our figure may not always happen, except to represent specific properties relating first-order knowledge and the representation of one's knowledge. For instance an agent may completely misrepresent the evidence underlying her first-order knowledge, and believe that she knows $\neg p$ when she is actually in a position to know $p$. Intuitively, this would be a case of delusion, rather than of underconfidence, namely a case in which the agent's representation of her evidence is entirely dissociated from what one might consider as conclusive evidence (see for instance Feinberg & Roane 2005 for a review of clinical cases, in which an agent, for instance, reports that her arm or leg, which she cannot move, belongs to someone else. This is a case in which, arguably, the

[4]Importantly, however, in order to correctly believe that one knows $p$ or that one does not know $p$ on a particular occasion, a perfect representation of one's first-order uncertainty is not needed. What suffices for adequately believing that one knows $p$ or that one does not know that $p$ is for the two levels to draw the same distinctions with regard to $p$ and $\neg p$ possibilities. This means that the agent may still misrepresent her knowledge about other propositions, but in a way that may be irrelevant for what concerns $p$.

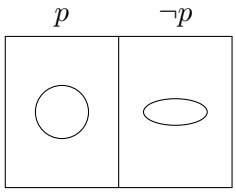


Figure 4: A situation of delusion: the agent believes she knows $\neg p$ (ellipsis); her first-order evidence (circle) should lead her to conclude $p$

agent believes her arm to belong to someone else, and believes she knows it, where she should rationally conclude that the arm really is her arm).[5]

Another case of dissociation between first-order evidence and the representation of one's first-order evidence that might be used to illustrate such dissociations is blindsight. In blindsight, the agent who suffers brain damage has very good reasons to believe that he won't navigate a course of obstacles, but his behavior shows he can navigate (see the overview of de Gelder 2010). The agent's internal evidence, in this case, puts him in a position to believe he cannot navigate. His behavior, on the other hand, leads one to conclude that he has the ability to navigate. Note that blindsight is a case of dissociation between a practical ability (what one can do) and the representation of that practical ability (what one believes one can do). Whether practical abilities can be analyzed in terms of propositional knowledge is a much debated issue (see Ryle 1971 and Stanley and Williamson 2001 for opposing views, as well as Lihoreau 2008). If we accept this reduction, then blindsight can be described as a case of dissociation between knowing-how and second-order knowledge of that knowing-how. A difficult question that we shall not attempt to discuss any further here concerns the relation between what we call first-order evidence and the representation of that first-order evidence. Obviously, in blindsight the internal evidence of blindness is distinct from the external evidence and feedback that one can reliably detect obstacles, but both can eventually be accommodated by the patient. Possibly, agents

[5]Agents whose arm of leg is paralyzed due to hemiplegia have some internal evidence that the arm might not belong to them, for instance inasmuch as they cannot move it. However, while some patients are ready to take other evidence into account to the effect that the limb still is theirs, others persistently deny the evidence to that effect. See Feinberg & Roane (2005: 667): "Patients who have asomatognosia may attribute ownership of the limb to the examining doctor. This simple misattribution often can be reversed when the error is demonstrated to the patient. In other patients, the misidentifications are truly delusional, and patients maintain a fixed belief in the misidentifications when they are confronted with evidence of their errors."

with blindsight who are told that they can navigate a course of obstacles are able to rely on that further evidence to correct their initial belief that they cannot navigate. The case, in this regard, is very distinct from cases of delusion in which patients are resilient against accepting new evidence against their initial belief.

A caveat is in order regarding the way the examples we reviewed were described. For these various cases of underconfidence and of delusion to count as counterexamples to positive introspection, they need to be so analyzed that the agent is taken to know but not to know that she knows. But the ascription of first-order knowledge is debatable. In the asomatognosia case and in the blindsight case, one might be tempted to resist the ascription of first-order knowledge and rather say that the agent possesses the relevant first-order evidence but somehow fails to gain first-order knowledge on the basis of that evidence. A similar analysis of the modified teaine *vs* caffeine example, where my knowledge in chemistry is in principle sufficient for me to exclude the possibility that they are different, could be proposed. According to that alternative analysis, I would be in position to know that teaine and caffeine are the same but I would not thereby know that they are the same. The underlying claim would be that being in position to know does not amount to knowing. This line of thought would of course be welcome to the advocates of positive introspection who hold that positive introspection cannot fail to hold because it is a characteristic property of knowledge.

## 5 Evaluating one's ignorance

In this section we turn to the discussion of the metacognitive mechanisms by which we appreciate whether we know that we know or know that we don't know a proposition. Based on the main distinction between conceptual unawareness and informational uncertainty, we will argue that we should distinguish two sets of metacognitive abilities. One concerns the ability to monitor one's uncertainty, the other concerns the ability to appreciate one's acquaintance with the ingredients of a proposition. The section is organized as follows: first, we review in a bit more detail aspects of the modeling of knowledge in epistemic logic. Then, we consider experiments done by Glucksberg and McCloskey, and more recently by Hampton et al., suggesting that the distinction between two forms of ignorance has a metacognitive correlate.

### 5.1 Informational content and conceptual content

Ignorance, we said so far, can result from two different conditions: lack of discrimination between competing alternatives on the one hand, lack of conceptualization of the alternatives on the other. In epistemic logic, standard models of knowledge generally incorporate the notion of discrimination. Relative to a context $w$, the information available to agent $i$ is represented by a set $R_i(w)$ of possibilities compatible with $i$'s evidence, called the information set. The information set is what we represented by a circle in the Figures above, and it represents the possibilities among which the agent cannot discriminate. The agent $i$ is then said to know $p$ in the context $w$ if and only if his or her information set $R_i(w)$ entails the proposition expressed by $p$.

Note that on that view, a knowledge state is defined purely in terms of the objective information available to the agent. A main limitation of this approach of knowledge is that it makes knowledge coarse-grained: two sentences with logically equivalent informational contents are such that an agent who knows the one is predicted to know the other. Likewise, an agent who knows $S$ is predicted to know $S'$ whenever the informational content of $S$ entails that of $S'$.[6] For instance, the model predicts that someone who knows that the keys are in the house thereby knows that the keys are in the kitchen or in some other room. Arguably, however, one may be aware that the keys are in the house without being aware that they are in the kitchen or in some other room, because the keys being in the kitchen is not a salient possibility.

Models of unawareness offer to solve this problem by enriching the purely informational definition of knowledge. On the resulting view, knowledge is relative not only to the agent's informational set $R(w)$, namely to a set of epistemic alternatives, but it is also relative to an awareness set $A(w)$, consisting of the concepts the agent is aware of. For instance, in order to be aware that they keys are in the car or not in the car, the concept of a "car" must be present in the agent's awareness set (see Franke and de Jager 2008). A knowledge content, on that perspective, is no longer defined in purely informational terms, but has a syntactic or linguistic structure (viz. Fagin and Halpern 1988, Franke and de Jager 2008, Hill 2010, Cozic 2011).

From a psychological point of view, therefore, the upshot of those models is that knowledge is a function of two distinct parameters: one concerns the occurrence in memory of the concepts necessary to articulate a proposition.

[6] See Stalnaker 1990 for a discussion of this prediction, and Fagin et al. 1995 for various ways of dealing with this problem, known as the problem of logical omniscience in epistemic logic. See also Cozic 2011 for more on the connection with unawareness issues.

The other concerns the quality of the information or evidence received to adjudicate whether that proposition is true or not. A reasonable prediction, therefore, is that metacognition, understood as the ability to know whether we know or don't know a proposition, will in turn be constrained differently, depending on whether we need to check on informational or conceptual content regarding our first-order knowledge.

### 5.2 Glucksberg and McCloskey's two-stage model

Empirical support for the hypothesis formulated above can be found in work by Glucksberg and McCloskey (1981), and more recently by Hampton et al. (2011) on the psychology of "known unknowns". Glucksberg and McCloskey put forward the supposition that there are two types of don't know decisions. The first type, on their account, concerns cases in which "no potentially relevant information is known" (1981: 312). The second type, by contrast, concerns cases in which "some potentially relevant information is retrieved, but this information turns out to be insufficient to permit an answer to the question".

An imaginary example of a don't know of the first kind they give is:

(1) What is the name of the largest store in Budapest?

For such a question, they consider that "most people [not living in Hungary] would probably say something like, "I have absolutely no idea", or "I know nothing about *any* of the stores in Budapest"." (ibid.). To illustrate a don't know of the second kind, they give the example of the following question:

(2) Is Kiev in the Ukraine?

This time, they imagine that the subject knows Kiev to be a city in Russia, knows Ukraine to be in Russia[7], but is "just not sure whether or not Kiev is in the Ukraine".

Both examples are examples in which the agent, once asked the question, is in a state of ignorance about the answer. In the first case, however, the agent lacks the conceptual information needed to answer the question (a wh-question, or constituent question). In the second case, by contrast, the agent simply lacks conclusive evidence to ascertain a yes vs. no answer to the question (a polar question, or yes-no question). An intuitive difference

[7] It is debatable, to say the least, whether in 1981 "Ukraine is in Russia" was true, as assumed by Glucksberg and McCloskey. Fortunately, this does not matter to the present discussion.

between the two cases is that, in the former case, the agent can be confident that he or she does not know the answer. In the latter case, by contrast, the agent should be less confident that she doesn't know the answer.

The difference between the two situations points to a metacognitive difference concerning the reliability of our don't know judgments. Thus, what Glucksberg and McCloskey predicted and tested for was that unknowns of the first kind should give rise to 'rapid don't know responses', and unknowns of the second kind to comparatively slow responses. A rapid don't know answer, on their account, is the output of a preliminary phase in which relevant information is searched for in memory. If no item is found, a don't know answer is produced. A slow don't know answer, by contrast, is the output of two phases: in the first memory search phase, some relevant information is found in memory about the ingredients of the question; in the second phase, evidence is examined to establish the relation between those ingredients.

In a first experiment, Glucksberg and McCloskey presented subjects with 24 sentences such as 'John has a pencil', 'Bill doesn't have a magazine'. They first trained subjects to remember the content of the sentences to a reliable extent. In the test phase, subjects were then presented with sentences of six different types (true affirmative/negative, false affirmative/negative, don't know affirmative/negative) and given the choice to respond by True, False, or Don't Know. In this setting, a sentence such as 'John doesn't have a pencil', for instance, is a false negative, because 'John has a pencil' was among the 24 sentences presented. 'John has a magazine', on the other hand, is a don't know affirmative, assuming that neither that sentence, nor its negation, were initially among the 24 sentences presented.

According to the two-stage model, correct 'I don't know' responses are the output of a process terminating earlier. Correctly responding 'I don't know' only involves going through the first phase of memory search (no evidence is found, the process terminates). By contrast, correctly responding 'Yes' or 'No' involves going through the first phase of memory search (evidence is found, which needs to be examined) and also through the second phase during which evidence is examined. Therefore, Glucksberg and McCloskey's model predicts that response times for correct 'I don't know' should be faster. Indeed, Glucksberg and McCloskey found response times for correct Don't Know answers to be significantly shorter than response times for correct True and correct False.

What happens with incorrect Don't Knows in this experiment? Glucksberg and McCloskey are silent about this. What the model predicts is indeed not so clear. It could be that no evidence has been found. In this case, the process would have terminated at the end of the first phase and shorter

response times are to be expected. Or it could be that some evidence has been found, which was deemed to be inconclusive (for example, the subject has stored some relevant information, but she is not able on the basis of that information to tell whether the sentence in the database was affirmative or negative). In that case, the process would have terminated at the end of the second phase and response times comparable to those for 'Yes' and 'No' answers are to be expected.

In order to test their model further, Glucksberg and McCloskey devised another experiment (experiment 3 in the 1981 paper), in which the two different kinds of Don't Know answers were made easier to distinguish. In that experiment, subjects' use of Don't Know answers was tested for statements pertaining to general knowledge. They made a distinction between two kinds of questions: questions for which they expected subjects to have relevant information in their memory (such as 'does Ann Landers have a degree in journalism?') even though that information was bound not to be sufficient to answer the question; questions for which they expected subjects not to be able to find relevant information (such as 'does Bert Parks have a degree in journalism?')[8]. As predicted by the two-stage model, Don't Know responses of the second kind, were found to be matched by faster reaction times.

The main interest of Glucksberg and McCloskey's study from our perspective is that it rests on a division between two forms of 'don't know' answers that can be related to the division between what we called uncertainty and unawareness. On their account, a first category of don't know answers corresponds to cases in which the agent has *no idea* of what the answer might possibly be (because subjects never had access to the answer). Cases of unawareness can be subsumed under that category. Suppose I am asked 'is Tolkien the author of the Hobbit?', and never heard of either Tolkien or the Hobbit before. From Glucksberg and McCloskey's model we expect that lack of acquaintance with the concepts will give rise to a fast and reliable 'don't know' answer. Glucksberg and McCloskey's second category of don't know answers corresponds to cases in which the agent does have relevant information, but hesitates between competing possibilities. This fits what we characterized as uncertainty proper.

It is important however to stress that not all cases that Glucksberg and

[8]Ann Landers and Bert Parks were two public figure widely known to most Americans in 1981. Ann Landers is the pen name of an advice columnist. Parks was a television announcer and the emcee of the Miss America pageant. What Ann Landers is known for is relevant to the question whether she has a degree in journalism. What Bert Parks is known for is not directly relevant to the question whether he has a degree in journalism.

McCloskey describe as cases for which the agent has no idea of what the answer might be can be adequately described as cases for which the agent has no acquaintance with the conceptual ingredients relevant to the question (the latter is certainly a sufficient condition for lack of any idea about the answer, but not a necessary one). For instance, in a question such as 'does Margaret Thatcher use an electric toothbrush?', the expectation is indeed to get a don't know answer of the first kind, but not because subjects never heard of Margaret Thatcher or do not have the concept of an electric toothbrush. Rather, the idea is that subjects will simply lack any appropriate evidence on which they can base their answer. In this regard, the dichotomy we propose between uncertainty-related unknowns and unawareness-related unknowns does not exactly coincide with Glucksberg and McCloskey's dichotomy, although it can be subsumed under theirs as a particular case.

### 5.3 Hampton et al.'s 2011 study

In a more recent study, Hampton et al. (2011) have proposed a distinct measure in order to evaluate what people know about their ignorance. Instead of comparing response times attached to don't know answers, they investigated whether the possibility of using an ignorance answer, as opposed to just True or False, increases the consistency of the subjects' use of True and False. In their study, subjects had to complete a questionnaire twice: one group of subjects could use only True/False answers, and another group could use '100% sure it's true', '100% sure it's false', and 'Not sure either way' (henceforth, Unsure). Subjects were invited to fill the questionnaire during a first session and then at a second session one or two weeks later. In the first experiment, subjects in each condition were presented with three distinct kinds of questions: questions of either general knowledge (viz. 'Texas is the size of Oklahoma'), autobiographical facts (viz. 'I have used a blue notebook') or about category statements (viz. 'Darts is a sport').

Hampton et al.'s basic finding is a more pronounced use of the 'Unsure' answer for questions of general knowledge, as opposed to questions about categories or autobiographical matters. For autobiographical questions or categorization questions, subjects were not significantly more consistent in their answers when they had the possibility to use the 'Unsure' answer, as opposed to just 'True' and 'False'. The 'Unsure' answer only increased consistency for questions about general knowledge. In a distinct experiment, Hampton et al. replicated the same pattern by comparing answers to questions of general knowledge to answers given about personal aspirations (viz. 'I aspire to be on TV') and questions about moral beliefs ('animal testing is

wrong'). There too they found consistency to increase only for questions of general knowledge.

Hampton's explanation for this contrast also relies on Glucksberg and McCloskey's model. For questions about general knowledge, Hampton et al. suggest that a reason for the greater stability of 'don't know' answers might be the more frequent lack of relevant information in memory. For instance, subjects may be more inclined to respond twice 'Unsure' to 'Texas is the size of Oklahoma' because they don't find any relevant information in memory (in order to be 100% confident either way). In contrast to that, the lesser stability of 'Unsure' answers for autobiographical facts or personal aspirations on their view is that 'you will always have some relevant basis in memory on which to base your answer. In this case it is a question of trying to retrieve evidence and argument in favor of the statement being true or not'. For instance, one may be less prone to saying 'I don't know' to 'I aspire to be on TV', because one can find reasons either way, and eventually in a way that will decide one for a True or for a False answer.

Hampton et al.'s study allows us to elaborate on the use of 'don't know' answers. To a question of general knowledge such as 'Texas is the size of Oklahoma', a typical way in which one would issue a stable don't know answer is when one does not even have a clear representation of where Oklahoma is located in the US and of what size it might be. This would correspond to a case in which one is initially unaware of what the size of Oklahoma might be: I never came across that information. Faced with the question, I can therefore move to a stable state of conscious uncertainty, in which I know that I lack the basic evidence to adjudicate the question. By contrast, faced with an autobiographical question such as: 'I have used a blue notebook', I can think of many occasions in which I have used a notebook. I then try to find evidence for whether or not, in at least one of those occasions, the color of the notebook was blue. In a case like this, intuitively, it is harder to be in a stable state of uncertainty, in particular because I know that as a matter of principle, I have had this information available to me.

### 5.4 Relation between the two don't know answers

The upshot of the two sets of experiments we discussed is that there appears to be two kinds of 'don't know' answers. The first kind includes cases in which we can be fairly confident that we do not know the answer, because no answer even comes to mind (as in 'what is the name of the largest store in Budapest?'). The second kind of 'don't know' answer includes cases in which we are able to articulate the answer in principle, but fail to be confident that

it is the correct answer. For such cases, a judgment of uncertainty is less reliable, simply because there is a competition between alternatives, based on available evidence in that case (as in 'did you ever use a blue notebook?').

A lot of cases may be mixed, however. Compare, for instance, the two questions: 'Is Alabama the size of Oklahoma?' and 'Is Texas the size of Oklahoma?'. At the moment I am writing, I would not be able to locate Alabama and Oklahoma on a US map, but would be able to locate Texas. The names 'Alabama' and 'Oklahoma' are familiar, however, just like the name 'Texas', but I happen to know more things about Texas than about the other two states (for instance, I know Texas shares a border with Mexico). Intuitively, the first question is a question for which I lack any potential evidence, since from my perspective, 'Alabama' and 'Oklahoma' are merely names of indistinct US states. Faced with the first question, I would therefore respond 'I don't know' without hesitation. I have *no evidence for a 'yes' as opposed to a 'no'*. In the case of 'Is Texas the size of Oklahoma?', I do have some potentially relevant evidence in my memory such as: 'Texas is a fairly large state in the US'. This is a case in which, though I do not know anything directly about the size of Oklahoma, and cannot remember exactly how big Texas looks like on a map, I would be tempted to make a guess (here a guess in favor of the hypothesis that Texas is not the size of Oklahoma).

An important aspect to this example is that while part of the question concerns an item about which I have hardly any direct evidence, I have at least some potentially relevant evidence coming from the other half of the question. Intuitively, this would be a case in which I am less than 100% sure that the answer is true, and also less than 100% sure that it is false, but also in which I am more than 50% sure that it is true. The Alabama/Oklahoma case, however, is one in which I am distinctively close to 100% sure that I don't know. In the Texas/Oklahoma case, this rather is a situation in which, though I am strictly speaking less than 100% sure either way, I am no longer indifferent between the yes and no answer. The difficulty about such cases is that, because we do have partly relevant evidence, we cannot be sure that we won't do better than chance. Equivalently, we cannot be sure that we don't know the answer, in the weak sense of being able to give the correct answer so as to do better than chance.

In concrete cases, therefore, the distinction between Glucksberg and McCloskey's two kinds of don't know answers will not be pure. On the other hand, attention to pure cases is revealing of the structure of higher-order knowledge. Thus, pure cases of conceptual unawareness are cases that will give rise to a clear perception of our ignorance. Franke and de Jager, for

instance, point out that an important feature of unawareness is that it is fragile. This means that, in a situation in which I am unaware of who Tolkien might even be, the very asking of the question 'is Tolkien the author of the Hobbit?' breaks the unawareness, and puts us in a state of uncertainty regarding whether the answer is yes or no (see section 3). By contrast, most situations of uncertainty are not situations that result from antecedent unawareness states. In the case of 'did you ever use a blue notebook?', I easily find positive evidence for a 'yes' as well as for a 'no'.[9] A report that one does not know will be less stable, then, because in principle, one would report uncertainty, rather than yes or no, only when the yes-evidence and the no-evidence balance each other. In the next section, we propose to focus on such cases: that is, we will set aside cases of ignorance resulting from unawareness, to focus on the appreciation of one's uncertainty in cases in

[9]D. Spector (p.c.) points out an interesting connection between the two kinds of don't know answers that we distinguish here and the distinction originally due to F. Knight (1921) in economic theory between 'risk' and 'uncertainty'. He also invites us to clarify the link between our notion of uncertainty and Knight's notion. Uncertainty in our sense and as used in epistemic logic is a generic notion, compatible with both what Knight calls risk and what he calls uncertainty. In economic theory since Knight, 'risk' is associated to the idea of an uncertainty that can be sufficiently precisely quantified, based on a priori or a posteriori statistical knowledge. 'Uncertainty' on the other hand is an uncertainty for which the agent may lack the resources to make any adequate probabilistic quantification (Knight 1921: III.VII.47-48). An illustration of 'uncertainty' in that Knightian sense is given by the classic example from Ellsberg (1961), in which you know that there are 90 balls in an urn, 30 red, and 60 blue or yellow. The ignorance of the exact proportion of blue balls is a case of uncertainty. Another example, this time related to unawareness, appears in a paper by Gilboa et al. (2009):

> "There is a semi-popular talk at your university, titled, "Cydophines and Abordites". You are curious and may listen to the talk (...) however, before the talk you have no idea what the terms mean. (...) You are asked whether all cydophines are abordites. Obviously, you have no idea. But if you are Bayesian, you should have probabilistic beliefs about this fact. How would you be able to come up with the probability that all cydophines are abordites?".

In our terms, what the example illustrate is a case of uncertainty resulting from antecedent unawareness – here lack of conception–, corresponding to a transition of the kind we illustrated in Figure 1 (except that, in this case, and unlike in our Tolkien and Hobbit example, there is not even a stable representation of what the expressions might possibly mean). Gilboa et al.'s remark about probabilities in a sense supports our observation that for uncertainty resulting from conceptual unawareness, there is no competition between relevant sources of evidence, hence no obvious way in which an agent could assign probabilities to the alternatives at issue, in contrast to cases of uncertainty resulting from a competition between alternatives informed by memory and by an adequate representation of conceptual space.

which we do find evidence for competing answers.

# 6 Evaluating one's uncertainty

The experiments reported in the previous section suggest the following picture of the relation between first-order knowledge and knowledge about that knowledge: in cases in which one lacks any evidence relevant to adjudicate whether a proposition is true or not, one can be confident that one does not know whether the proposition is true or not. Cases of prior lack of acquaintance with the constituents of the question are cases for which one will typically issue a reliable 'don't know' judgment. By contrast, in cases in which one does have evidence regarding the status of the proposition, but only partial evidence, knowing whether one knows or not is typically harder to adjudicate, for it depends on an evaluation of the weight of one's evidence. In this section we propose to examine more closely the relation between the strength of one's first order evidence and the adequacy of metacognitive evaluations of one's own knowledge.

## 6.1 Discrimination tasks

A relevant paradigm for the investigation of the evaluation of one's uncertainty is given by discrimination tasks in which the difficulty in discriminating is gradual and can be modulated. This paradigm, in particular, plays a central role in Smith et al.'s theory of uncertainty monitoring (Smith et al. 2003). Typically, subjects are assigned a task of discrimination in which they have to report a particular condition ($s$, for signal), or its absence ($n$, for no signal). Subjects are allowed to give three kinds of response, either a Signal Response ($S$) or a No Signal response ($N$) or they can opt out and use a third response ($U$, for "Uncertainty").

For instance, Smith et al's. density discrimination task is one in which subjects are shown a box with a number of pixels illuminated. The signal condition in their display is when the number of pixels is exactly 2950. No signal corresponds to the case of fewer than 2950 pixels. The signal condition is matched with a Dense response, and the No signal condition with a Sparse response. As expected of such psychophysical tasks, what Smith et al. report is that when the number of pixels is sufficiently below 2950, discrimination is easy and subjects make correct use of the Sparse response. Close to 2950 pixels, subjects make larger use of the Dense response. Between those two levels, there is a range of pixel configurations for which the use of the third response gradually increases (in humans and certain species of animals).

Typically, the use of the third response is highest where the response curves for Sparse and Dense cross each other, namely are at a ratio of 1:1. What appears, in particular, is that subjects make the heaviest use of the third response in the region where the competition between signal and noise is at its maximum, namely where subjects are equally drawn to the Sparse or the Dense response.

A noteworthy feature of the curves shown by Smith et al. is that when the third response is at its maximum use, it is only used about 70% of the time (ibid., Fig. 3, subject D). In contrast to that, when the Sparse response is at its maximum use (for easy Sparse configurations), it represents close to 100% of the responses. One may wonder about the sources of this asymmetry.

It is possible, first of all, to conceive of experimental set-ups in which optimal use of the Uncertainty response would reach 100%. As signal detection theory has made clear, how much a response is used in a task is a function not only of the subject's sensitivity, but also of the structure of rewards and penalties attached to the task (see McNicol 1972/2005). Because of that, a subject could be in a state of less than perfect discrimination about the correct response, that is a state of uncertainty, but still be rationally motivated to use the response Sparse instead of the response Uncertain because the former is a more profitable strategy.

Consider however a structure of rewards is such that the expected value of the Uncertainty response in certain configurations is the same that the expected value of the Sparse (or Dense, for that matter) response on other configurations. If one observed a tendency for subjects to use the Uncertainty response in the former configurations to a lesser extent than Sparse (or Dense) on the latter configurations, this would provide evidence that it is harder to adequately perceive one's being uncertain when one is uncertain than it is to perceive that one is certain when one is certain. The response curves presented by Smith et al. are compatible with this hypothesis: they indicate that the uncertainty response is less stable for intermediate cases than the Sparse and Dense responses are for clear cases.

### 6.2 Higher-order knowledge and imperfect discrimination

Incidentally, situations of imperfect discrimination have been used by Williamson (1990, 1994, 2000) as typical exemplifications of cases in which agents ought to lack an adequate representation of their knowledge and uncertainty. Williamson's argument is a normative argument about the failure of positive introspection in principle, but it is interesting to try and relate

it to actual tasks of discrimination. Williamson does not use the framework of signal detection to give a model of uncertainty, though some links can be made between his model and the SDT model (see Egré and Bonnay 2010). In particular, he does not rely on a probabilistic representation of uncertainty, as in SDT, but instead he uses qualitative models of uncertainty of the sort informally presented above, in which agents can be basically in three states: either they know that $p$, or they know that not $p$, or they are uncertain either way.

For Williamson, situations of imperfect discrimination can be characterized as situations in which the relation of epistemic possibility between alternatives is non-transitive.[10] For instance, for configurations with more than 1000 pixels illuminated (and less than 4000, say), an agent with limited discrimination may not be able to reliably discriminate between pixels configurations that differ from each other by fewer than 25 points. This means that if the pixel configuration were 2950 pixels, the agent could not accurately discriminate it from a configuration of 2925 pixels, nor from a discrimination of 2975 pixels. However, the agent may be able to discriminate a configuration with 2925 pixels from one with 2975 pixels.

In Williamson's approach, the minimum difference reliably detectable between two configurations can thus be viewed as setting a margin of error: between 1000 pixels and 4000 pixels, for instance, the agent estimates the status of configurations with a margin of error of about 25 pixels. This means that if the configuration is exactly 2950 pixels, the agent should be uncertain whether it is Sparse or Dense. If the configuration is 2920 pixels, however, the agent is in a position to know that the configuration is Sparse, because it is represented as having at most 2945 pixels, which still counts as Sparse.

An important assumption of Williamson's account is that margins of error constrain first-order knowledge and higher-order knowledge in a uniform way. This means that if the margin of error were 25 pixels, then in order to be certain that the configuration is Sparse, the configuration must be below (2950-25)=2925 pixels. But in order to be certain that one is certain that the configuration is Sparse, the configuration must be below (2950-2×25)=2900 pixels. And similarly at higher levels: each new iteration of knowledge is represented by the addition of a new margin of error. This model makes the following interesting prediction regarding the relation between first-order discrimination and metacognition. The prediction is: the further away a

[10] This assumption is common to a number of other frameworks. See also Halpern 2008, Luce 1956 and van Rooij 2010 among others.

signal is from the boundary, that is, the higher the signal to noise ratio, the more confident the agent should be that there is signal. On the other hand, the model makes a counterintuitive prediction, which is that given a fixed margin of error, there will always be a failure of higher-order knowledge at some point (for instance, suppose the agent sees a configuration of 2875 pixels: this is three margins of error away from the Dense condition. Here the agent is predicted to know that he knows that the configuration is Sparse, but not to know that he knows that he knows this, a very counterintuitive prediction). Arguably, however, some pixel configurations are such that the signal to noise ratio will be so high that subjects will be in a position not only to issue a correct answer, but also to be certain that it is the correct answer.

One possibility to amend Williamson's basic model in this regard is simply to assume that perceptual margins of error do not constrain first-order knowledge and metacognitive levels in the same way (see Dokic and Egré 2009, Bonnay and Egré 2009 and Egré and Bonnay 2010 for a conceptual and logical elaboration, as well as Loussouarn 2010 for a discussion of the cognitive basis of a such a distinction). If that assumption is relaxed, one can preserve the idea that the stronger the signal to noise ratio will be, the more confident an agent should be that he has first-order knowledge, while making some cases such that the agent is in principle fully confident about those. Some other options have been suggested: Mott (1997), Halpern (2008), Dutant (2007) and more recently Spector (2011) basically consider that Williamson's margin for error principles are not sound, as a result of which they preserve a similar idea, which is that positive introspection can be maintained for the internal perceptions of the agent.

# 7 Conclusion

Our initial observation in this paper has been the idea that in order to examine what it takes to know that one knows or to know that one does not know something, one should carefully distinguish between different forms of ignorance. Based on a central distinction made in formal epistemology, we have argued that there are two fundamentally distinct sources of ignorance, namely ignorance based on uncertainty, and ignorance based on unawareness.

Both uncertainty and unawareness are sources of unknown unknowns and of unknown knowns. In cases of unawareness, unknown unknowns are the most typical cases, whereas unknown knowns correspond to cases of

implicit knowledge that an agent cannot represent to herself for lack of the relevant concepts. In cases of uncertainty, unknown unknowns and unknown knowns are better described as forms of overconfidence and underconfidence respectively, that is as cases in which the agent does not have an adequate representation of the structure of his or her first-order uncertainty and evidence.

Based on that, following Glucksberg and McCloskey's two stage model of answer processing, we have argued that the principled distinction between uncertainty and unawareness has a metacognitive correlate. Deciding whether one knows or does not know the answer to a question appears to give rise to stabler and faster verdicts for cases in which one lacks basic acquaintance with the answer, that is for cases in which one can realize one's unawareness prior to the question. By contrast, deciding whether one knows or does not know the answer to a question is typically harder and more demanding for cases in which one has some evidence available in memory, but competing evidence, that is for situations of uncertainty. Fundamentally, the reason why it is harder to know whether one knows or does not know a proposition in cases of uncertainty is due to the fact that one needs to weigh the strength of available evidence. This stage is simply not needed for cases in which no evidence is found in memory.

Little has been said here finally about the psychological processes that enable us to weigh our first-order evidence and to decide about the strength of our first-order uncertainty. In the case of Smith et al.'s discrimination paradigm, however, we pointed out that there likely is an asymmetry between the confidence that one is certain, and the confidence that one is uncertain. This implies, on our view, that high degrees of first-order certainty should tend to go with high confidence that one is certain. By contrast, high degrees of uncertainty do not appear to give rise to high confidence that one is uncertain to the same extent.

## Acknowledgments

We are very grateful to J. Proust for detailed comments on the first version of this paper, and for a number of helpful exchanges on the topic of metacognition. We are also indebted to J. Hampton for directing our attention to Glucksberg and McCloskey's model of ignorance and to his recent work on metacognition. We also thank D. Spector for feedback and for stimulating discussions on the epistemology of higher-order knowledge, and M. Cozic for valuable comments and discussions on the topic of unawareness.

## References


[1] Aloni M. & Egré P. (2010). Alternative questions and knowledge attributions. *The Philosophical Quarterly* 60 (238):1-27.

[2] Aloni M., Egré P., de Jager T. (forthcoming) Knowing whether A or B. *Synthese*.

[3] van Benthem J. (2004). What one may come to know. *Analysis* 64 (282): 95105.

[4] Bonnay D. and Egré P. (2009). Inexact knowledge with introspection. *Journal of Philosophical Logic*, 38 (2), pp. 179-228.

[5] Bonnay D. and Egré P. (2011). Knowing one's Limits. An analysis in Centered Dynamic Epistemic Logic. In P. Girard, M. Marion and O. Roy (eds), *Dynamic Formal Epistemology*, Springer.

[6] Bromberger S. (1987) What we don't know when we don't know why. Repr. in Bromberger S. *What we know we don't know*. University of Chicago Press and CSLI Publications. 1992.

[7] Cozic M. (2011). Probabilistic Unawareness. *Cahiers de recherche, Série Décision, Rationalité, Interaction*, IHPST: Paris. Under review.

[8] Dekel E., Lipman B.L., Rustichini A. (1998). Standard State-Space Models Preclude Unawareness. *Econometrica* 66 (1): 159-173.

[9] van Ditmarsch H., Kooi B., van der Hoek W. (2007), *Dynamic Epistemic Logic*.

[10] Dutant J. (2007). Inexact Knowledge, Margin for Error and Positive Introspection. in D. Samet (ed.) *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge* (TARK XI), (Louvain-la-Neuve: Presses Universitaires de Louvain) pp.118-124.

[11] Egré P. (2008). Reliability, Margin for Error and Self-Knowledge. In V. Hendricks and D. Pritchard (eds), *New Waves in Epistemology*:215-250, Palgrave MacMillan

[12] Egré P. (2011). Epistemic Logic. In L. Horsten and R. Pettigrew eds. *Continuum Companion to Philosophical Logic*, chap. 16. Continuum.

[13] Egré P. & Bonnay D. (2010). Vagueness, Uncertainty and Degrees of Clarity. *Synthese* 174, vol. 1, pp. 47-78.

[14] Ellsberg D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics* 75 (4): 643669.

[15] Fagin R. & Halpern J. (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence* 34: 39-76.

[16] Fagin R., Halpern J., Moses Y., Vardi M. (1995) *Reasoning about Knowledge*, MIT Press.

[17] Feinberg T.E. & Roane D. M. (2005) Delusional Misidentification. *Psychiatrics Clinic of North America* 28: 665D683.

[18] Franke M. & de Jager T. (2010). Now that you mention it: Awareness Dynamics in Discourse and Decisions. In A. Benz et al. (eds), *Language, Games, and Evolution*. Lecture Notes in Computer Science, Volume 6207, 60-91.

[19] de Gelder B. (2010). Uncanny sight in the blind. *Scientific American*, 302(5): 60-65.

[20] Gilboa I., Poslewaite A.., Schmeidler D. (2009). Is it always rational to satisfy Savage's axioms? *Economics and Philosophy*, 25: 285-296.

[21] Glucksberg, S. & McCloskey, M. (1981). Decisions about Ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory* 7: 311-325.

[22] Halle M. (1978), Knowledge Unlearned and Untaught: what speakers know about the sounds of their language. In M. Halle, J. Bresnan and G. A. Miller, *Linguistic theories and psychological reality* (1978), pp. 294-303.

[23] Halpern J. (2008). Intransitivity and Vagueness, *Review of symbolic logic*, 1 (4), pp. 530-547.

[24] Hampton J. A., Aina B., Mathias Andersson J., Mirza H. Z., Parma S. (2011), The Rumsfeld Effect: the unknown unknown. *Journal of Experimental Psychology: Learning Memory & Cognition*, in press.

[25] Hemp D. (2006). The KK (Knowing that one Knows) Principle. Internet Encyclopedia of Philosophy.

[26] Heifetz A. & Meier M. & Schipper B. C. (2006). Interactive unawareness. *Journal of Economic Theory* 130: 7894.

[27] Hill B. (2010). Awareness Dynamics. *Journal of Philosophical Logic* 39 (2): 113-137.

[28] Hintikka J. (1962). *Knowledge and Belief. An Introduction to the Logic of the Two notions*. Cornell: Ithaca.

[29] Hintikka, J. 1970. Knowing that One Knows reviewed. *Synthese* 21: 141-62.

[30] Knight F. (1921). *Risk, Uncertainty and Profit*. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Co.

[31] Lihoreau F. (2008). Knowledge-how and ability, *Grazer Philosophische Studien* 77 (1):263-305

[32] Loussouarn, A. 2010. *De la métaperception à l'agir perceptif*. Thèse, EHESS.

[33] Luce D. (1956). Semiorders and a Theory of Utility Discrimination. *Econometrica* 24 (2): 178-191.

[34] McNicol D. (1972/2005). *A Primer of Signal Detection Theory*, Psychology Press, reissued 2005, Lawrence Erlbaum and Associates.

[35] Mott P. (1998). Margins for Error and the Sorites Paradox. *Philosophical Quarterly* 48 (193):494-504.

[36] Pérez Carballo A (2010). Structuring logical space, manuscript, MIT, under review.

[37] Proust J. (2007) Metacognition and Metarepresentation: Is a Self-Directed Theory of Mind a Precondition for Metacognition? *Synthese*, Vol. 159, No. 2, Self-Ascriptions of Mental States (Nov., 2007), pp. 271-295.

[38] R. van Rooij (2010). Vagueness and Linguistics. In G. Ronzitti (Ed.), *The Vagueness Handbook*. Dordrecht: Springer.

[39] Ryle G. (1971). Knowing How and Knowing That, in Gilbert Ryle: *Collected Papers*, Volume 2 (New York: Barnes and Nobles, 1971): 212-225.

[40] Schaffer J. (2007). Knowing the answer. *Philosophy and Phenomenological Research* 75 (2): 383-403.

[41] Smith J. D., Shields W.E , Washburn D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences* 26, 317373.

[42] Spector D. (2011). Margin for error semantics and signal perception. Manuscript, Paris School of Economics. Under review.

[43] Stalnaker R. (1990). Mental content and Linguistic Form. *Philosophical Studies* 58. Repr. in R. Stalnaker, *Context and Content*, Oxford University Press, 1999, chap. 12.

[44] Stanley and Williamson (2001). Knowing how. *The Journal of Philosophy*, 98 (8), 411-444.

[45] Williamson T. (2000). *Knowledge and its Limits*. Oxford UP.