From *Theoria* 68 (2002), 13–51.

**Invitation to Autoepistemology**

**LOYD HUMBERSTONE**  
(Monash University)

1. Autoepistemic Logic

The phrase ‘autoepistemic logic’ was introduced in Moore [1985] to refer to a study inspired in large part by criticisms in Stalnaker [1980] of a particular nonmonotonic logic proposed by McDermott and Doyle. Very informative discussions for those who have not encountered this area are provided by Moore [1988] and the wide-ranging survey article Konolige [1994], and the scant remarks in the present introductory section do not pretend to serve in place of those treatments as summaries of the field. A good deal of the material omitted here pertains to the specifically nonmonotonic nature of autoepistemic logic as standardly developed, but as we shall urge, there is from one point of view nothing distinctively nonmonotonic about the basic motivating ideas of the subject.

In autoepistemic logic one is interested in inferences made by an ideally rational and introspectively omniscient individual, reflecting not only on the contents of his/her—henceforth generally (following Stalnaker) ‘its’—beliefs but on the fact that they comprise its beliefs. (‘Belief’, throughout, should be taken as unqualified, fully confident belief, rather than the kind of tentative belief involved in saying “I believe that $p$, though I’m not sure”.) As Moore has acknowledged, the more appropriate name for the subject, at least as just characterized, would be ‘autodoxastic logic’, but we stick with the established name, following the practice of using ‘epistemic’ as a generic term covering both ‘epistemic’ (in the narrow sense) and ‘doxastic’. Similarly, we use the term ‘epistemology’ in a broad sense, encompassing not only the theory of knowledge but also the theory of belief, so that we have available the derived term ‘autoepistemology’ to substitute for ‘X’ in completing the proportionality: **epistemic logic** is to **epistemology** as **autoepistemic logic** is to **X**. Coincidentally, as important in autoepistemology as R. C. Moore has been in autoepistemic logic, is an earlier namesake, the philosopher G. E. Moore. The seminal role of the puzzle known as ‘Moore’s Paradox’ after him will come up in Section 2, in which the main subject of discussion is a constellation of arguments by Sydney Shoemaker roughly to the conclusion that the phrase used above, “an ideally rational and introspectively omniscient individual” involved a redundancy: knowledge, at least of what one’s own beliefs are, is actually an aspect of rationality rather than the result of ‘inner perception’ of one’s current mental states. If G. E. Moore is the father of autoepistemology, then Descartes—with his inescapably ‘first person’ *Cogito* argument—is perhaps its grandfather; we consider, from this perspective, some reactions to that argument in Section 3. In the meantime, we continue to introduce the logical side of the picture. We make no promise that the formal development definitively resolves any particular philo-

---

1 McDermott and Doyle [1980]. McDermott [1982].

2 We will concentrate on Shoemaker [1988]. An early presentation (originally published 1942) Moore’s paradox may be found in Moore [1952], p.543. *(Warning: The volume in which this appears, Schilpp [1952] also contains a paper by M. Lazerowitz entitled ‘Moore’s Paradox’, but it deals with an unrelated matter.)*

3 The ineliminably first-person character of much of Descartes’ reasoning is emphasized, for instance, in Anscombe [1975], p.55.
sophical issues, but invite the reader to enjoy an enhanced appreciation of these issues made available by having the framework of autoepistemic logic in the background. (A final section returns to logical matters, and an Appendix discusses a somewhat broader usage of the term “autoepistemology” from that envisaged here.)

Unlike traditional epistemic(-doxastic) logic, autoepistemic logic takes a first-person or ‘internal’ point of view, and when considering a set of statements containing both $\varphi$ and $B\psi$ we are to think of this as potentially recording the reflective agent’s beliefs to the effect that $\varphi$ and to the effect that it itself believes that $\psi$. Exploiting this internal perspective, autoepistemic logic can accommodate the fact that in some sense, on which more below, $B\varphi$ ‘follows from’ $\varphi$: the reflective agent can move from the presence of $\varphi$ amongst its beliefs, to the explicitly doxastic conclusion that ‘I believe that $\varphi$’ also belongs there—and so, given an appropriate rationality idealization, may be taken already to be there. There is, by contrast, no sense in which traditional epistemic logic with its external perspective sanctions such a transition: from the mere truth of (arbitrary) $\varphi$, nothing follows about its being believed that $\varphi$ by any given subject.\(^4\)

The most widely known legacy of Stalnaker [1980] is the concept of a stable set of beliefs, informally intended as a set beyond which “no further conclusions could be drawn by an ideally rational agent in such a state” (p.187); the formal definition, subject to mild re-notation (and relabelling of the defining conditions), and suited for application to sets of formulas of sentential logic is in terms of three conditions. We take ‘$B$’ as a 1-ary connective alongside the usual boolean connectives, and $\vdash$, in (DC) below, is the relation of tautological consequence as applied to the formulas of this language. The definition is as follows: set $\Gamma$ of formulas is stable just in case for all formulas $\varphi$, the following three conditions are satisfied:

\begin{align*}
(DC) & \text{ If } \Gamma \vdash \varphi \text{ then } \varphi \in \Gamma. \\
(+) & \text{ If } \varphi \in \Gamma \text{ then } B\varphi \in \Gamma. \\
(–) & \text{ If } \varphi \notin \Gamma \text{ then } \neg B\varphi \in \Gamma
\end{align*}

To clarify the idea that a subject’s total stock of initially given beliefs, $\Gamma_0$, say, leads by the process of reflection on what is in and what is not in that stock to a particular stable set ‘grounded in’ $\Gamma_0$, Moore introduced the latter concept with the definition that $\Gamma$ is grounded in $\Gamma_0$ when

$$\Gamma \subseteq \text{Cn}(\Gamma_0 \cup \{B\varphi \mid \varphi \in \Gamma\} \cup \{\neg B\varphi \mid \varphi \notin \Gamma\})$$

where $\text{Cn}$ is the consequence operation associated with the consequence relation $\vdash$ featuring in the deductive closure condition (DC) above. Replacing the inclusion here by an equality gives the definition of Moore’s concept of $\Gamma$’s being a stable expansion of $\Gamma_0$. (This is indeed a stable set, as the reader may verify.) Several consequence-like notions can be defined in terms of the latter, though we cannot get any notion of $\varphi$’s following from $\Gamma_0$ by defining this to hold when $\varphi$ is an element of the stable expansion of $\Gamma_0$ because there is in general no such thing as ‘the’ stable expansion of a given set of formulas. (There may be more than one such expansion, and there may be none; see the Moore and Konolige references given earlier.) One way to fix this situation is to consider the relation – let us call it “$\models$” – defined to hold between $\Gamma_0$ and $\varphi$ just in case $\varphi$ is an element of every stable expansion of $\Gamma_0$. This relation is nonmonotonic (and therefore not in the traditional sense a consequence relation) in that

\(^4\) As is clear from the sentence to which this note is appended, we use symbols such a ‘$\varphi$’ here in two different ways: on the one hand to occupy the place of names of sentences (or formulas) and on the other to occupy the place of sentences; there is a similar double employment of sentence letters ($p,q, \ldots$) below.
\[ \Gamma_0 \subseteq \Gamma_1 \text{ and } \Gamma_0 \models \phi \text{ do not imply } \Gamma_1 \models \phi. \]

An example given by Moore is the following, in which \( p \) is some sentence letter:
\[ p \rightarrow \text{B}p \models \neg p, \text{ but it is not the case that } p, p \rightarrow \text{B}p \models \neg p. \] (Strictly we should write “\( \{p, p \rightarrow \text{B}p\} \models \neg p\),” etc., but use the same abbreviative conventions as are customary for consequence relations; ’\( \rightarrow \)’ stands for material implication.) Roughly, the reason that we have \( p \rightarrow \text{B}p \models \neg p \) is that there is no way to reason from \( \{p \rightarrow \text{B}p\} \) to \( p \), so \( p \) should be missing from a stable set grounded in \( \{p \rightarrow \text{B}p\} \), so \( \neg \text{B}p \) goes into such a set, which then by (DC)—essentially, *Modus Tollens* here—gives \( \neg p. \)

In addition to the nonmonotonic \( \models \) just considered, several monotonic relations can be defined in terms of the notion of stability. To introduce these, let us call a set \( \Gamma \) **positively stable** if it satisfies (for all formulas \( \phi \)) Stalnaker’s conditions (DC) and (+) above, and **negatively stable** if it satisfies the conditions (DC) and (–); thus the stable sets are those which are both positively and negatively stable. Moore [1985] criticized McDermott and Doyle [1980] for what amounts defining a variant notion like that of stable expansion except with negatively stable sets in place of stable sets, “McDermott and Doyle’s agents are omniscient as to what they do not believe, but they may know nothing about what they do believe” (p.86). We are not here especially interested in the nonmonotonic inference relations, such as \( \models \) above and this variant, however, because the epistemological connections to be attended to below do not require the complexities they bring with them. (These complexities arise mainly over the notion of grounding and are given a full and subtle discussion by Moore and Konolige, the latter introducing several refinements of this notion to avoid various anomalies.) Consider, then, the three consequence relations—whose monotonicity is evident from the form of the definitions—given by:

\[
\begin{align*}
\Gamma_0 &\models^{\text{st}} \phi \text{ if and only if for every stable } \Gamma \supseteq \Gamma_0, \phi \in \Gamma. \\
\Gamma_0 &\models^{\text{st}+} \phi \text{ if and only if for every positively stable } \Gamma \supseteq \Gamma_0, \phi \in \Gamma. \\
\Gamma_0 &\models^{\text{st}−} \phi \text{ if and only if for every negatively stable } \Gamma \supseteq \Gamma_0, \phi \in \Gamma.
\end{align*}
\]

The relation \( \models^{\text{st}} \), in which we shall take the greatest interest below, obviously differs from \( \models \) since the latter is not monotonic; this reflects the fact that not every stable superset of a set (here \( \Gamma_0 \)) is a stable expansion of that set. (Since stable expansions are indeed, as noted above, stable sets, we do have the inclusion \( \models^{\text{st}} \subseteq \models \).) On the other hand, in the case of \( \models^{\text{st}+} \), if we had instead defined a notion of the positively stable expansion(s) of a set following Moore’s lead but replacing stability by positive stability, in other words, calling \( \Gamma \) a **positively stable expansion** of \( \Gamma_0 \) when \( \Gamma = \text{Cn}(\Gamma_0 \cup\{\text{B}p \mid \phi \in \Gamma\}) \), then the resulting \( \models \)-style relation holding between a set \( \Gamma_0 \) and those formulas in each of its positively stable expansions, would be monotonic and would coincide with \( \models^{\text{st}+} \). Since this is not so for the corresponding negative version of \( \models \), as is evident from the remarks on McDermott and Doyle above, one may tempted to conclude that it is Stalnaker’s “negative” condition (–) which bears the responsibility for nonmonotonicity. But since \( \models^{\text{st}} \) as well as \( \models^{\text{st}−} \) employs this condition in the background and these are monotonic, that would be an overswift reaction: it is rather the interaction between the negative condition and the ‘grounding’ idea that is

---

5 Moore [1988] illustrates what is going on here by considering his own perspective on the choice of \( p \) as ‘Nixon is dead’, reasoning that if Nixon were (as of 1988) dead, he (Moore) would know about it, so one of his beliefs is represented by \( p \rightarrow \text{B}p \); we assume—without here going into exactly what this amounts to—that he no other relevant beliefs. In particular, not having the belief that Nixon is dead, by (–) Moore arrives at \( \neg \text{B}p \), and hence by *Modus Tollens*, at the conclusion that Nixon is still alive. For further discussion of such reasoning on the basis of one’s own ignorance, see Humberstone [1988]. (See note 23 below for a correction to note 5 of this paper.)
The three relations $\vdash^{st}$, $\vdash^{st+}$, and $\vdash^{st–}$ are all distinct; for example taking $\vdash^–$ as $\vdash^{st+}$, we have (1) but not (2), while for $\vdash^–$ as $\vdash^{st–}$ we have (2) but not (1), while with $\vdash^–$ as $\vdash^{st}$ we have both (1) and (2):

(1) \quad p \vdash^– Bp

(2) \quad \vdash^– Bp \rightarrow p \quad (\text{i.e., } \emptyset \vdash^– Bp \rightarrow p)

(1) is obviously correct for $\vdash^– = \vdash^{st}$, $\vdash^{st+}$, by (+); for $\vdash^– = \vdash^{st}$, $\vdash^{st–}$, we have (2) because if $p$ belongs to negatively stable $\Gamma$, we get $Bp \rightarrow p \in \Gamma$ by (DC), and if not, then by $\neg(–)$, $\neg Bp \in \Gamma$, so, again by (DC), $Bp \rightarrow p \in \Gamma$. (A further comment on (2) will be made below.) A case of non-distinctness should also be noted. For each of the above choices of $\vdash^–$, one can imagine a variant definition which replaces the reference to ‘every stable (positively stable, negatively stable) $\Gamma \supseteq \Gamma_0$’ in the above definition by one to ‘every consistent stable (positively stable, negatively stable) $\Gamma \supseteq \Gamma_0$’. Now although stability does not imply consistency—the set of all formulas is certainly stable—it is not hard to see that the variant definitions do not give consequence relations distinct from the respective originals. For, to take the case of $\vdash^{st}$ by way of example, if every consistent stable superset of $\Gamma_0$ contains $\varphi$, then every stable superset of $\Gamma_0$ contains $\varphi$, since the inconsistent (stable) set certainly contains $\varphi$. Using this fact one sees that we have not only (3) but also (4), for every formula $\varphi$:

(3) \quad \varphi \vdash^{st} B\varphi

(4) \quad B\varphi \vdash^{st} \varphi

For, à propos of (4), if $B\varphi$ belongs to a consistent stable set, then so does $\varphi$, or else (by $\neg(–)$) $\neg B\varphi$ would, contradicting consistency (given (DC)).

As the justification—which did not exploit the status of $p$ as a sentence letter—for (2) above reveals, we can also write (4) in a form with nothing on the left and ‘$B\varphi \rightarrow \varphi$’ on the right. For (3) this is not the case: $\varphi \rightarrow B\varphi$ certainly does not belong to every stable set for arbitrary $\varphi$. For example, take $\varphi$ as $p$. The easiest way to see that this does not belong to every stable set (and so is not a consequence by $\vdash^{st}$ of the empty set) would be to exploit the fact that a consistent set of formulas is stable if and only it comprises all those formulas true throughout a Kripke model (for modal logic, thinking of ‘B’ as an alternative notation for ‘$\Box$’) in which every point is accessible to every point.\(^6\) (More generally: $\Gamma \vdash^{st} \varphi$ just in case for any such model, if all formulas in $\Gamma$ are true throughout the model, then $\varphi$ is true throughout the model. This is a special case of the semantic description offered in Section 4 below.) Since we can have more than one point in the model, this knocks out $p \rightarrow Bp$. The moral of this example is that $\vdash^{st}$ does not satisfy the

---

\(^6\) This was proved in the 1980’s independently by several people, including Halpern and Moses [1985]; see Theorem 3 on p.112 of Moore [1988], and the footnote on that page, for further information and references. Moore’s formulation does not involve the restriction to consistent stable sets we have imposed, because he allows Kripke models in which the set of points (‘worlds’) is empty. Halpern has suggested to me that the simplest way of understanding the relation $\vdash^{st}$ is in terms of the usual preservation-of-truth-at-a-point-in-a model consequence relation associated with $SS$; calling this simply $\vdash$, then we have $\psi_1, \ldots, \psi_n \vdash^{st} \varphi$ just in case $B\psi_1, \ldots, B\psi_n \vdash B\varphi$. (See the remark which follows in the main text as to why such an equivalence obtains.) I have preferred to relegate this characterization to a footnote since, while technically correct, it seems to suggest a kind of reduction of autoepistemic to alloepistemic logic which is not consistent with the perspective the present paper advocates. Halpern’s most recent (somewhat negative) assessment of autoepistemic logic in the style of Moore may be found in Halpern [1997].
Deduction Theorem: we cannot, that is, given the information that \( \Gamma, \varphi \vdash^{\text{st}} \psi \), conclude that \( \Gamma \vdash^{\text{st}} \varphi \to \psi \). We shall return to this point in Section 2 and again in Section 4.

There are a few points to make about (3) and (4). (3) appears to say that with respect to some notion of ‘following from’ which we are prepared to take seriously enough to have baptized with special notation, \( B\varphi \) follows from \( \varphi \). And, it might be objected, something has gone seriously wrong if we are willing to entertain any sense in which its being believed that \( \varphi \) follows from its being the case that \( \varphi \) (for arbitrary \( \varphi \)). Nor, the objection would continue, against (4), is there any sense in which the truth of an arbitrary belief follows from that belief’s being held. Such complaints would be appropriate against a venture in traditional epistemic(-doxastic) logic, but they overlook the internal perspective characteristic of autoepistemic logic. The transition, in thought, from the left-hand side of (3) to its right-hand side is perfectly legitimate when it involves passing from a \( \varphi \) which the reasoner accepts to a conclusion of the form “I believe that \( \varphi \)”. (Cf. the example at the end of this section.) The starting point for such a transition is not the mere hypothesis that \( \varphi \), but rather the status of \( \varphi \) as accepted by the reasoner. A similar consideration applies in the case of (4). This distinction between premisses as suppositions (mere hypotheses) and premisses as (putatively) given data will receive further attention in Sections 3 and 4 below.

We can re-express the conjunction of (3) and (4) as a strengthening of Stalnaker’s condition (+):

\[
(++) \quad \varphi \in \Gamma \ if \ and \ only \ if \ B\varphi \in \Gamma.
\]

There is a similar strengthening available for (–):

\[
(––) \quad \varphi \not\in \Gamma \ if \ and \ only \ if \ \neg B\varphi \in \Gamma.
\]

But whereas the condition (++) is satisfied by arbitrary stable sets, (––) is only satisfied by consistent stable sets. Indeed neither of the two directions of (––) corresponds to a \( \vdash^{\text{st}} \) statement in the way that the two directions of the equivalence in (++) correspond to (3) and (4), which suggest that a certain generalization of this consequence relation may be worthy of consideration. The following paragraph, which may be omitted by readers interested primarily in the applications presented in succeeding sections, is devoted to this matter.

The idea would be to use generalized or ‘multiple-conclusion’ consequence relations which allow there to be other than exactly one formula on the right; appropriate defining conditions—which can be thought of as structural rules—may be found in Shoesmith and Smiley [1978], q.v. for further historical and motivational details, as well as references to versions of the generalized framework by Gentzen, Carnap, and Scott. The one we are interested in we call \( \models^{\text{st}} \), though of course one could define also generalized consequence relations \( \models^{\text{st}+} \) and \( \models^{\text{st}–} \), by analogy with \( \vdash^{\text{st}+} \) and \( \vdash^{\text{st}–} \) above. We define:

---

7 Given that we have not defined \( \vdash^{\text{st}} \) in terms of deduction using formula-to-formula rules, the label ‘Deduction Theorem’ for the property here seen to fail is perhaps less than ideal. If we were working explicitly with sequents and sequent-to-sequent rules, we would instead be speaking, in a sequent-calculus proof-system, of a failure of the rule for inserting \( \to \) on the right, or, in a natural deduction approach, of a failure of \( \to \)-Introduction (“Conditional Proof”). Since we need a term for the corresponding property of consequence relations independent of any particular proof-theoretic (or indeed semantic) presentation of them, we choose the label likely to be familiar to the largest number of readers, and take it that ‘Deduction Theorem’ is that label.
\( \Gamma_0 \models^{st} \Delta \) if and only if for every stable \( \Gamma \supseteq \Gamma_0, \Gamma \cap \Delta \neq \emptyset \).

Note that this coincides in meaning with the claim that \( \Gamma_0 \vdash^{st} \varphi \) for the case in which \( \Delta = \{ \varphi \} \). By exploiting the fact that \( \Delta \) can have more than one element, we can reflect Stalnaker’s condition \((-\rangle \) ( = the ‘only if’ direction of \((\neg –\)) directly, as (5), in which \( \Gamma_0 \) is empty:

\begin{equation}
\Gamma_0 \models^{st} \varphi, \neg B \varphi
\end{equation}

Our discussion of (2) above revealed that (for any \( \varphi \)) \( B \varphi \to \varphi \) belonged to every stable set, making it a \( \vdash^{st} \)-consequence of the empty set, so that we have, writing out the \( \to \) in favour of negation and disjunction, and replacing \( \vdash^{st} \) by \( \models^{st} \), since these are equivalent when there is a single formula on the right:

\begin{equation}
\Gamma_0 \models^{st} \varphi \lor \neg B \varphi
\end{equation}

But (6) is a weaker claim than (5). (6) only says that every stable set must contain \( \varphi \lor \neg B \varphi \), whereas (5) says that every such set must contain either \( \varphi \) or else \( \neg B \varphi \). Since a disjunction can belong to a stable set without either disjunct belonging, though not conversely, the latter is the stronger claim, and it is one that cannot directly be expressed without recourse to the generalized consequence relation \( \models^{st} \). (Commas on the right do not, then, amount to occurrences of \( \lor \), by contrast with what the use of multiple right-hand sides in more conventional settings may lead one to expect.) \( \models^{st} \) also allows us, à propos of the ‘if’ direction of \((\neg –\)) to contrast (7) and (8):

\begin{align}
\neg B \varphi, \varphi &\models^{st} \psi \\
\neg B \varphi, \varphi &\models^{st}
\end{align}

(7) is a correct claim, which could equally well have been made with \( \vdash^{st} \), whereas (8), which because we have \( \emptyset \) on the right could not have been made with \( \vdash^{st} \), is not correct as it stands and would require for its correctness that above definition of \( \models^{st} \) be replaced by one in which the reference to stable supersets gives way to a reference to consistent stable supersets—a distinction without a difference, as we noted, for the case of traditional single-suceedent consequence relations.

Those familiar with the literature on autoepistemic logic may feel the label to be misappropriated for application to the study of such monotonic consequence relations as \( \vdash^{st} \), \( \vdash^{st+} \) and \( \vdash^{st–} \). The following quotation from Stalnaker [1980], p.186, uses the term ‘non-monotonic’ in place of the (yet to be coined) ‘autoepistemic’, but conveys an important aspect of the flavour of the autoepistemic logic it inspired, an aspect which is quite separable from the nonmonotonicity of the usual development of the subject in the hands of Moore and others:

The crucial assumption that motivates the distinctive features of non-monotonic logic is that the agent makes inferences, not only from the given information, but from the fact that it has the information that it has.

The phrase ‘from the fact that it has the information that it has’—setting to one side an irrelevant reading on which the fact in question is the rather dull fact that would be reported by ‘I have the information that I have’—remains ambiguous between ‘from the fact that it has all the information that it has’ and ‘from the fact that it has all and only the information that it has’. It is the latter interpretation that leads (when coupled
with the idea of groundedness) to the nonmonotonicity of Moore’s autoepistemic logic, whereas even with
the former we have the distinctive ‘auto’ (‘internal’, ‘first-person’) feature. For example, if one of my
beliefs is that if I believe that \( p \), then \( q \), and another is that \( p \), then I can reach the conclusion that \( q \) in two
steps from these beliefs. First I draw \( p \) out of my stock of beliefs; then I pass to ‘I believe that \( p’ \); a Modus
Ponens then completes the journey. The characteristic moving between levels – the level of belief, and the
level of the content of the belief in question – has here nothing to do with nonmonotonicity. Such a move
may also legitimately occur in the reverse direction (from ‘I believe that \( p’ \), to \( p \) ), as is embodied in (4)
above. Varying the case just given, then, if one of my beliefs is that if \( p \) then \( q \), and another is that I believe
that \( p \), then I can reach the conclusion that \( q \) again by two steps from these beliefs, the second of which is
Modus Ponens and the first is the removal of the ‘I believe that’ prefix. The internal perspective here
continues to exhibit no nonmonotonicity (since (3) and (4) above are given for monotonic \( \vdash_{st} \)), so we
conclude that there is nothing inherently nonmonotonic about autoepistemic logic conceived as the formal
exploration of that perspective.

2. Shoemaker on Knowing One’s Own Mind

Shoemaker [1988] develops an argument along approximately the following lines for the conclusion that
however it may be that we become aware of our own mental states, it is certainly not by means of the
exercise of a special quasi-perceptual faculty of introspection. For—the argument runs—if such
introspection were required, we could imagine a fully rational and cognitively sophisticated individual in
whom this faculty was nonetheless defective. Such an individual, Shoemaker argues, would not, even
though fully conversant with the concepts deployed in such ascriptions, as well as fully rational,
conceptually sophisticated and in no way cognitively impaired, be able to self-ascribe propositional
attitudes except on the basis of evidence of the same kind as one person ascribes attitudes to another
(“third-person evidence”). Someone meeting this description—including not only the ‘privative’ part but also
the condition of sophisticated rationality—Shoemaker calls self-blind. But, he goes on to argue (as we
shall see in the following paragraph), in fact there could not be such a self-blind individual. Therefore, the
normal (“first-person”) way we have of ascribing propositional attitudes to ourselves is not via the exercise
of any such quasi-perceptual introspection. For any perception-like way of coming to know things must be
logically capable of failing to work, or of being absent, in such a way as to leave a person—whose
rationality and cognitive sophistication we may suppose to be quite unimpaired—ignorant of the facts on
which it is supposed to deliver knowledge: to be in the state the argument concerned shows to be the
impossible state of self-blindness.

We turn to the main argument offered against the possibility of self-blindness. Shoemaker himself
actually presents a somewhat different argument in this capacity—something he calls the ‘argument from
Moore’s Paradox’—to which we shall return below, and gives the present argument as a subsidiary

---

8 Others may take the core idea of autoepistemic reasoning differently; it is clear that Moore emphasizes the
nonmonotonic aspects of the relation we have called \( \vdash_{st} \) (though somewhat confusingly, he uses the term
‘nonmonotonic’ only to apply to the logic(s) of McDermott and Doyle he is criticizing). Levesque [1990]
develops a monotonic modal-style logic with an additional epistemic operator for “all I know is that ___” to
capture the ‘negative’ aspects of Moore’s autoepistemic logic, and subtitles his paper ‘A Study in
Autoepistemic Logic’, although what we regard as distinctive—the internal perspective—is absent. (The
germs of this idea is already present in Halpern and Moses [1985].)

9 See also the other papers in Shoemaker [1996], as well as Gallois [1996], for further developments of this
theme.
consideration. Suppose that we have a self-blind individual; following Shoemaker, we call him George. Recall that part of what is meant in describing George as self-blind is that George is a rational and cognitively sophisticated individual; the rest—called the ‘privative’ part of the notion above—is that these abilities notwithstanding, George lacks knowledge, except when it is available on the basis of third-person evidence, of his own propositional attitudes, and in particular his beliefs. George is accordingly perfectly capable of entertaining the thought that he has this or that particular propositional attitude, even though he cannot settle its truth-value in a certain way. About the ‘rationality’ part of the heading, we have to be a little less specific; certainly included is the ability to follow out the logical consequences of one’s beliefs, and perhaps to detect and avoid inconsistencies amongst them, as well as to avoid various ‘pragmatic paradoxes’ (Shoemaker’s term) involved in one’s assertions. When we ask self-blind George what he believes about what’s on the table in front of him, he will be unable to answer, there being, we may suppose, no evidence of the third-person type available to him. But when we ask him what’s on the table in front of him, he unhesitatingly replies that there is a bowl containing two apples and a banana.

Now the crucial consideration deployed by Shoemaker is that we can plausibly treat as one aspect of the rationality part of self-blindness, the condition that George conforms to the following principle, which we may call the Question-Linking Principle:

If asked “Do you believe that P?” he ought to answer “yes” just in case he would answer “yes” to the question “Is it true that P?” (and this should be so whether his intention is to tell the truth or to lie). (Shoemaker [1988], p.194)

If we ask George, in the situation envisaged, whether he believes that there is a banana in front of him, he will be unable to answer, in the absence of evidence of the requisite type. But if we ask George instead whether there is a banana in front of him, he will answer “Yes”, since he has all the evidence he needs for this. Thus George is violating the Question-Linking Principle, contradicting the assumption that he suffers from no defect of rationality. It would appear, then, that it is impossible for there to be anyone satisfying all the conditions we have imposed on this hypothetical figure, George, and that therefore our self-acquaintance is not of the quasi-perceptual sort we imagined absent from his case.

There is a conspicuous resemblance between the Question-Linking Principle and the condition (++) noted in Section 1 to be satisfied by all stable sets (to the effect that $B\varphi$ is in such a set just in case $\varphi$ is). Since autoepistemic logic was described in part in terms of introspective omniscience, it may seem unclear that Shoemaker’s intent here is autoepistemological; as he notes himself, the etymology of the word ‘introspect’ suggests exactly the account of knowledge of one’s own beliefs that he is concerned to repudiate: the quasi-perceptual model. But we can use the term non-committally, its etymology notwithstanding. In fact, familiarity with the study of autoepistemic logic may well tend to reduce the quasi-perceptual associations of ‘introspection’ because thinking of such transitions as that from $\varphi$ to $B\varphi$ as in some sense logical, as that study suggests, is very much in line with Shoemaker’s intent: to reveal a preparedness to make these transitions as the exercise of one’s power of reason rather than of a special inwardly directed sensory faculty. Quite how to describe the transitions is somewhat problematic, however, and here we shall just briefly explain the problem without attempting to solve it. At the end of Section 1, we quoted a passage from Stalnaker to the effect that the subject “makes inferences (...) from the fact that it has the information that it has” and it is very natural to speak in this way of inferring $B\varphi$, or perhaps $\neg B\psi$ from the fact that one believes that $\varphi$ (in keeping with (+)) or, respectively, does not believe that $\psi$ (in keeping with (−)). Natural or not, a difficulty arises. In order to make an inference from the fact that such-and-such, the subject would first have to register the fact that such-and-such; for example one could not infer anything from the fact that one’s house was one fire unless one noticed – or was told, or
somehow or other gathered – that one’s house was on fire.\textsuperscript{10} Not to beat about the bush: one would have first to believe that one’s house was on fire before one was in any position to make any inference whose starting point was the fact that one’s house was on fire. Now our envisaged agent is supposed to infer from the fact that it believes that $\varphi$ to the conclusion $\text{B}\varphi$ ( = “I believe that $\varphi$”). The required starting point for this inference, if what was just said about cases such as that of one’s house being on fire was correct, is already the terminus for the inference. There is presumably some solution to this problem—essentially a problem in how best to conceptualize autoepistemic reasoning—but we leave the task of finding it to others.\textsuperscript{11}

Another complaint against the above appeal to the Question-Linking Principle has been raised by Wlodek Rabinowicz (p.c.): the plausibility we find in this principle as a norm of rationality may be due to its being addressed to beings – namely us – who are self-aware by exactly the ‘inner perception’ route Shoemaker argues is not available, but we take its availability so much for granted that we assume anyone would on pain of irrationality abide by the principle, failing to take seriously the possibility that such a route might not be available to the subject under consideration. This would indeed be a fatal flaw in Shoemaker’s argument, and it is hard to tell whether the flaw is present because the Question-Linking Principle is essentially wheeled in in the warm light of a favourable intuitive reaction – and intuitions, the objector would maintain here, are hardly to be trusted when they have been honed against a background (of massively widespread quasi-perceptual self-awareness) absent from the critical test cases. Sidestepping what appears, then, to be something of a stand-off, we pause to note another way of taking “beings like us” in the suggestion that the Question-Linking Principle has appeal as a normative principle only for beings like us. Instead of considering this as amounting to “beings who are quasi-perceptually self-aware”, we could take it as “beings who have the concept of belief”. Of course, employing, as it does, the concept of belief, the Question-Linking Principle would not be intelligible to anyone lacking that concept, but the present proposal is that a necessary condition for a subject $S$ to have the concept belief would be that $S$ self-ascribes precisely those beliefs that $S$ possesses. (We need not think of such self-ascription as overtly linguistic – \textit{cf.} note 14 below – and indeed the Question-Linking Principle itself could be put in terms of raising, in thought, the two questions “Is it the case that $P$?” and “Do I believe that $P$?”. The Principle demands that the either should be answered “Yes” just in case the other is.) The prototype for moves along these lines would be a suggestion from Peacocke [1987], p.171, to the effect that “it is partially constitutive of a thinker’s possessing the concept pain that he is disposed to judge the first-person, present-tense content that he is in pain precisely when (and for the reason that) he is in pain”. A more thorough discussion than we intend to embark on here would investigate the merits of the analogous suggestion that that partially constitutive of possessing the concept of belief that one be disposed to judge the first-person present tense content that one believes that P precisely when (and for the reason that) one does believe that P.

\textsuperscript{10} The same point applies if instead of speaking of inferring something from the fact that such-and-such, one spoke of inferring it from one’s house being on fire, from one’s believing that so-and-so, etc. It is not being claimed that one can only make inferences from what one believes to be the case, because certainly one can make inferences from mere suppositions—something we shall be keen to emphasize in Section 3 below—but these are hardly to the point here.

\textsuperscript{11} Presumably such a solution would characterise and then invoke some non-inferential notion of a rational transition. Writing of the transition from $p$ to ‘I believe that $p$’ Gallois [1996], p.46, says that it constitutes a “peculiar type of inference-like move [which] is justified”: what would be desirable is a more informative characterisation.
Although we have given, as Shoemaker’s ‘main argument’ above, an argument based on the Question-Linking Principle, what Shoemaker himself seems to regard as the main argument—with that described above having something of the status of a variation—is his ‘argument from Moore’s Paradox’. The appeal to Moore’s Paradox by Shoemaker occupies a rather obscure dialectical role in his paper, however, and is not related to the Question-Linking Principle in quite the way he seems to think, which is why it is here relegated to a secondary position. Recall that a Moore-paradoxical assertion is one of the form ‘ϕ but I do not believe that ϕ’. Shoemaker [1988], p.193, remarks that “there are, it seems offhand, conceivable circumstances in which such an utterance might be expected from someone who was self-blind...”, since the person could have evidence favouring each conjunct. But rationality should preclude anyone from making such an (assertive) utterance; therefore self-blindness is not compatible with rationality. This is the argument from Moore’s Paradox. But in the first place its status is, as I say, obscure, because later (p.201f.), Shoemaker writes:

There is a contradiction involved in the idea that the total evidence available to someone might unambiguously support the proposition that it is raining and that the total third-person evidence might unambiguously support the proposition that the person does not believe that it is raining. For the total third-person evidence concerning what someone believes about the weather should include what evidence he has about the weather—and if it includes the fact that his total evidence concerning the weather points unambiguously towards the conclusion that it is raining, then it cannot point unambiguously towards the conclusion that he doesn’t believe that it is raining. So the situation I said seems “offhand” to be conceivable is not really conceivable.

Setting to one side this dialectical complication, there are further difficulties with trying to use the argument from Moore’s Paradox in the way that Shoemaker proposes to use it.

The first of these is that it is far from clear how the argument is supposed to work. Given that George believes that ϕ, the mere pressure coming from his supposed cognitive sophistication to avoid being Moore-paradoxically committed to ‘ϕ but I do not believe that ϕ’ should certainly lead him to avoid committing himself to “I do not believe that ϕ”: but he can do this by remaining non-committal rather than by committing himself to its negation, “I believe that ϕ”.12 (Note that “I do not believe that ϕ” is to be taken as having the force of “¬Bϕ” rather than, as it may in other contexts often mean, “B¬ϕ”. This well known ambiguity – analysed in the heyday of transformational grammar by appeal to the rule of neg-raising to obtain the latter reading – gives rise to the question posed by the title of Williams [1979].) Actually we can distinguish a weaker from a stronger sense in which one might be non-committal as between ϕ and ¬ϕ. The stronger sense is that one might explicitly believe “I do not believe that ϕ and I do not believe that ¬ϕ”. The weaker sense—we could call it implicit non-commitment—is that while not having the latter

12 Gallois [1996], p.78, maintains that an explicitly non-committal attitude is not available to the self-blind individual. He claims—with what merit, we do not propose to ask—that such an individual, believing that either Tom or Mary is at the party, though having no idea as to which of them is there, will have to say that (“from my point of view”) it is true that Tom or Mary is at the party even though it is not true that Tom is there, and not true that Mary is there”. This means that the distinction in the parenthetical comment immediately following in the text would not be available to one who is self-blind. There is a special issue about the non-commitment in respect of what George believes suggested in the text: one whose beliefs form a stable set cannot be non-committal about anything of the form Bϕ since either this or its negation belongs to every such set. But the idea that a rational subject’s beliefs form a stable set is—a somewhat extended form of—what Shoemaker is arguing for here and not something available at this stage of the dialectic to argue from. (The need to say something about this was drawn to my attention by Steve Gardner.)
belief, one simply fails to believe either that $\varphi$ or that $\neg\varphi$. The implicit form of non-commitment seems to recommend itself especially in the present context, since something of a higher-level Moore-paradoxicality would be apparent on the part of someone who believes this ‘$\varphi$ but I have no belief as to whether or not I believe that $\varphi$’. (Such higher-level versions of Moore’s Paradox were drawn to my attention by Brian Weatherson.)

Perhaps one might attempt the following, less direct, route from the considerations about Moore’s Paradox to the Question-Linking Principle. Often the idea that a set $\Gamma$ of premisses entails a conclusion $\varphi$ is explained in terms of inconsistency by saying that this amounts to the inconsistency of the set $\Gamma \cup \{\neg\varphi\}$. Call a set of statements Moore-paradoxical when commitment to the truth of all of its elements is Moore-paradoxical in the familiar sense; thus for instance $\{p, \neg Bp\}$ as well as $\{p \land \neg Bp\}$ and any superset of either of these, are Moore-paradoxical. Then define a set $\Gamma$ to Moore-entail $\varphi$, by parity with the notion of entailment simplciter, just when the set $\Gamma \cup \{\neg\varphi\}$ is Moore-paradoxical. (This is of course quite unfair to Moore, who was after all responsible for introducing the word ‘entail’ into philosophy in its currently accepted sense, but it is appropriately suggestive for the present short-term usage.) We notice then that the transition from any $\varphi$ to $B\varphi$ is the transition from $\varphi$ to something Moore-entailed by $\varphi$, perhaps hoping that this is how the need to avoid Moore’s Paradox might bring the Question-Linking Principle in its train. (What the above definitions give directly from the Moore-paradoxicality of $\{\varphi, \neg B\varphi\}$ is that $\varphi$ Moore-entails $\neg\neg B\varphi$, rather than $B\varphi$, but this is a nicety we can ignore.) The suggestion does not work well, however. Take our other example above of a Moore-paradoxical set, $\{p \land \neg Bp\}$. Applying the definition of Moore-entailment here, we get that the empty set Moore-entails $\neg(p \land \neg Bp)$, or, as we may re-write this, $p \rightarrow Bp$. This is certainly not the kind of thing we expect our cognitively sophisticated subject George to have to be committed to, and, as we saw in the preceding Section while discussing the failure of the Deduction Theorem à propos of (3) there, it is not something whose presence in every stable set is guaranteed.

The way in which Moore-entailment was defined in terms of Moore-paradoxicality unfortunately guarantees, taken against a background of classical logic for the boolean connectives, a kind of contraposition: if $\varphi$ Moore-entails $\psi$, then $\neg\psi$ Moore-entails $\neg\varphi$. For if $\varphi$ Moore-entails $\psi$, then $\{\varphi, \neg\psi\}$ is Moore-paradoxical, in which case so is $\{\neg\psi, \neg\neg\varphi\} (= \{\neg\neg\neg\varphi, \neg\psi\})$, and accordingly $\neg\psi$ Moore-entails $\neg\varphi$. Contraposition of this kind is something we do not want in our account of the conditional commitments the rational subject is prepared to make. We want the commitment to $B\varphi$ given the commitment to $\varphi$, but we do not want a commitment to $\neg\varphi$ given a commitment to $\neg B\varphi$. The Question-Linking Principle was appropriately stated by Shoemaker in terms of Yes as an answer for George to the question “Do you believe that $p$?” if and only if this was his answer “Is it the case that $p$?”; he very sensibly did not suggest that George should answer one of these questions with No if and only if that was his answer to the other. The appropriate ‘negative’ form of the condition $(\leftrightarrow)$ which the Question-Linking principle makes vivid is the condition $(\rightarrow \rightarrow)$ in Section 1, and not the following condition on stable $\Gamma$: $\neg\varphi \in \Gamma$ if and only if $\neg B\varphi \in \Gamma$, which such sets in general do not satisfy, and which the ‘No’ version of the Question-Linking Principle would endorse.\(^{13}\) This example shows, incidentally, that the consequence relation $\vdash_{st}$ of Section 1 is not (as it is sometimes put) congruential—one cannot replace equivalents by equivalents in arbitrary contexts and secure equivalent results—for as we have just noted, $p \vdash_{st} B\varphi$, yet it is not the case that $\neg p \vdash_{st} \neg B\varphi$. (We write ‘$\varphi \vdash_{st} \psi$’ for ‘$\varphi \vdash_{st} \psi$ and $\psi \vdash_{st} \varphi$’.) In fact, the failure of contraposition just noted is another aspect of the failed Deduction Theorem already remarked upon, as we

---

\(^{13}\) For this reason, the gloss given by Gallois [1996], p.50, on the Question-Linking Principle is potentially misleading: “for me the question whether $p$ is not distinguishable from the question whether I believe that $p$”.

11
can see if we write $\neg \varphi$ as $\varphi \rightarrow \bot$ (for any formula, $\bot$ a special *Falsum* constant or the abbreviation of some fixed contradiction $q \land \neg q$ – in the latter case we exempt $\neg q$ from the range of those formulas $\neg \varphi$ envisaged as rewritten). For suppose that

(i) $\varphi \vdash_{st} \psi$;

by (DC) we have

(ii) $\psi, \psi \rightarrow \bot \vdash_{st} \bot$.

From (i) and (ii) there follows by a suitable structural principle (essentially the ‘Cut Rule’, with $\psi$ as the Cut-formula):

(iii) $\varphi, \psi \rightarrow \bot \vdash_{st} \bot$.

Here is where the Deduction Theorem fails, stopping us from moving $\varphi$ across to the right as an antecedent: $\psi \rightarrow \bot \vdash_{st} \varphi \rightarrow \bot \left(= \neg \psi \vdash_{st} \neg \varphi \right)$.

To return to Shoemaker’s discussion, then, we have the following summary of the situation. Shoemaker suggests that self-knowledge in respect of one’s beliefs is best thought of as part and parcel of what it is for someone who is fully rational to have those beliefs, rather than as an ‘add-on option’ in the form of some special faculty of ‘inner sense’. While this suggests that a logical articulation along the lines of Section 1 would be appropriate, we saw also that the drift of Shoemaker’s own discussion—trying to justify the Question-Linking Principle in terms of the desideratum of avoiding Moore-paradoxicality—does not really work, and we should probably make any justificatory moves in the reverse direction. There is no difficulty in seeing what autoepistemic logic (of the monotonic form envisaged in Section 1) would have to say about Moore’s Paradox: no consistent stable set containing $\varphi$ can contain $\neg B \varphi$. This is no doubt not an explanation of what is wrong with Moore-paradoxicality, so much as a way of putting the claim that something is wrong with it into a systematic setting; but a satisfactory explanation was in any case given long ago in the framework of traditional epistemic logic, and by the founder of traditional epistemic (‘alloepistemic’) logic; transplanting it into the setting of autoepistemic logic should be well suited to the first-person point of view in which Shoemaker has taken such an interest.\(^{15}\)

---

\(^{14}\) Hintikka [1962], 64–82. For an excellent discussion, defending essentially Hintikka’s treatment but with a coverage of much in the literature since (and prior to) its appearance, see Chapter 1 of Sorensen [1988]. One point which is clear from Hintikka’s and Sorensen’s discussions is that Moore’s Paradox has nothing to do with assertion: the anomaly arises over the mere prospect of believing that (e.g.) it is raining even though I do not believe that it is. In Chapter 4 of Shoemaker [1996], ‘Moore’s paradox and self-knowledge’, Shoemaker has come round to recognising this; the discussion in Shoemaker [1988] is very much in the older assertion-obsessed style – witness his use, for example, of the term ‘pragmatic paradox’.

\(^{15}\) Shoemaker’s declared intention in his [1988] is to defend “a moderate Cartesianism—a version of the view that it is of the essence of mind that each mind has a special access to its own contents”, which presumably covers not only belief but the absence of belief. Nothing in his discussion—Moore’s paradox, the Question-Linking Principle, etc.—seems to bear on the negative autoepistemic condition (–), however, and the same goes for Gallois [1996]. (The paper does touch on the extension to other propositional attitudes, such as desire, as does Gallois [1996], and several items in Shoemaker [1996] deal with phenomenal states as opposed to propositional attitudes collectively. These do not connect with our autoepistemic logic theme, however.)
3. Khamara on Descartes

Several aspects of Descartes’ discussion of scepticism in the *Meditations*, and some subsequent secondary literature on the subject, direct our attention to matters autoepistemological. We shall illustrate this with reference to one contribution to that literature here: Khamara [1984]. By way of background to the discussion in Section 5 of that paper, we note that its first two sections concern respectively the notions of assertion and of assertive (and other forms of) self-refutation. The third discusses so-called ‘non-inferential’ interpretations of the *Cogito*, and in the fourth, the Dream Argument (for scepticism) is presented after the fashion of G. E. Moore by Khamara, with a key premiss being “I do not know (occurrently) that I am not now dreaming.” Khamara faults the argument by claiming that this premiss is assertively self-refuting in that “the truth of the proposition expressed by this sentence precludes its assertion by the proponent of the argument”. For present purposes, we need not discuss the merits of this charge, beyond noting that it does seem something of a distraction to lay so much stress on the notion of assertion. (Compare note 14 above.) Amongst the things that make for felicitous assertion are those features which serve the purposes we may think of the practice of assertion as ‘designed to’ fulfil, such as the communication of truths. It seems that a less social concept – such as that of assent to a proposition – ought to replace the role played by assertion here if we are evaluating the position of a thinker explicitly non-committal with respect to the very existence of the social context (other people, that is) into which the concept of assertion fits. If assent is something like (full) belief, then ‘assentively’ self-verifying propositions would presumably be those that are incorrigible. Self-refutation would arise when the truth of a proposition is incompatible with its being believed. (Such a proposition is a belief blindspot, in the sense of Sorensen [1988].)

From Descartes’ difficulties in mounting a sceptical challenge, Khamara turns to difficulties he finds in Descartes’ attempt to meet this challenge, attending in his fifth and final section, ‘The *Cogito* as an inference’, to Descartes’ *Cogito* as this has been traditionally construed: as an argument to extricate its user from a scepticism so total as to cover the matter of the user’s own existence. Khamara holds that Descartes may indeed succeed in proving, on the basis of his current mental activity, that he exists, but that any such success counts for nothing in the battle against scepticism, as represented by the possibility that he is dreaming or being deceived by an Evil Demon, since ‘what he is after is not the proposition “I now exist” but the stronger epistemic proposition “I know occurrently that I now exist”.’ With regard to the dream hypothesis, Khamara writes (p.115):

Suppose now that I am dreaming. Then it follows that I am having an experience (or am “thinking” in the relevant Cartesian sense). It follows that, as the owner of that

16 The other premiss is “If I do not know (occurrently) that I am not now dreaming, then I do not know occurrently that I am now sitting by the fire” and the conclusion is “I do not know (occurrently) that I am not now dreaming.”

17 Khamara [1984], p.113. The idea of looking at Cartesian scepticism and the *Cogito* argument with the concepts of self-verification and self-refutation to hand goes back to earlier work by A. J. Ayer, E. J. Lemmon, and J. L. Mackie in the 1950s and ’60s, reference to some of which work may be found in Khamara [1984].

18 We use the term ‘proposition’ here with a warning that it should not be understood in the familiar ‘set of worlds’ sense. In particular, the proposition that Khamara is standing up, for instance, needs to be distinguished from the proposition expressed by an utterance by Khamara of “I am standing up”. Khamara’s saying “It is raining though Edward Khamara does not believe that it is”, is not, for example, Moore-paradoxical.
experience, I now exist. So far (let us say) so good. But it does not follow that I know occurrently that I now and indeed the latter epistemic proposition is inconsistent with our initial premiss, namely the supposition that I am now dreaming.

The final sentence here invokes the principle that if one is dreaming one does not know anything, whose plausibility appears to depend entirely on a certain anti-Cartesian view of dreams. It is, we should also note, no objection to an argument to say, as Khamara does here, that its conclusion is inconsistent with one of its premisses, if the argument occurs as part of a course of reasoning designed to impugn that premiss. But before putting this observation into perspective, we should see Khamara’s corresponding treatment of the Evil Demon hypothesis (also p.115); he has distinguished an interpretation of the relevant passage in Descartes as literally envisaging a powerful deceiver from a second more metaphorical interpretation according to which this is just a way of supposing that all one’s beliefs are false\textsuperscript{19}:

\textsuperscript{19} Though it is somewhat tangential to our autoepistemological theme, it deserves to be said that this second interpretation, with its hypothesis that all one’s beliefs—or indeed all of anyone’s beliefs—are false is highly problematic. To see this, we should first notice a problem in a subject’s taking take seriously the threat of error in making even the simplest inferences. (An alternative reply to Khamara would be to deny that Descartes needs this confidence in his inferential abilities in order for knowledge to be gainable by their use, in the same way that it is often denied a perceiver needs to know or even believe anything about his/her perceptual apparatus in order for the normal functioning of this apparatus to give rise to perceptual knowledge.) The difficulty rises over the difficulty of making sense of having a belief without having—or at least been disposed to form—other logically related beliefs. Could I really believe, for example, that I was warm and comfortable, without believing that I was warm? Could I believe that I was in a quiet room without believing that someone was in a quiet room? Or at least, being prepared to concede that this did follow? What if I said to someone that I was sure I had a coin in my hand but not quite sure, because I wasn’t too confident about inference that day, that I had something in my hand, even though there being something in my hand admittedly appeared to me to follow from there being a coin in my hand? Assuming that reluctance to commit oneself to the immediate consequences of one’s declared beliefs rather undermines the declaration, consider the possibility that Oliver is being deceived about everything concerning the material contents of the room he’s in. (So we set aside any special questions about deceivability as to one’s own occurrent mental activity, etc.) Oliver thinks that object \textit{a} is in the room but that \textit{b} isn’t. So what must really be going is that \textit{a} isn’t in the room and \textit{b} is. But wait a minute. Does Oliver believe that there is something in the room? If not, then for the sort of reason just gestured toward, I begin to lose my grip on the claim that he believes the object \textit{a} is in the room. If so, then this is a belief of his that is going to be—not false but—true, since as we’ve just seen \textit{b} is indeed in the room. Conclusion: it’s harder than you might first think to have all your beliefs be false, even those not about the current state of your mind. (Here’s a more abstruse example: Oliver believes the something is in the room and from this, or perhaps just on the basis of pure reflection, he concludes that either something is in the room or nothing is. It doesn’t matter whether the process whereby he came by this belief is prone to error, thanks to an Evil Demon perhaps: it’s not a belief that could be false.) Slightly varying the example, suppose that Oliver believes that \textit{a} is in the room but—not in this case specifically \textit{b}—something or other is not in the room. The beliefs can’t both be false, so, here wheeling back in introspective knowledge of one’s own beliefs, if Oliver has the two beliefs just mentioned he can know he is not the victim even of a small-scale Evil Demon intent on trying to bring it about that all his beliefs about what’s in the room are false. The same goes for those about what’s in the whole of the external world. But, Khamara would stress, we aren’t dealing with these low-budget deceptions, or we might as well have stuck with dreams (conceived as Descartes conceives them). We are supposed to be worried into fear of deception even as to what follows from what. We have seen that the hypothesis that all one’s beliefs about the external world are false may well be incoherent. Yet if we are taking seriously the possibility of a rationality-disabling Evil Demon, can we really trust any reasoning we might employ to demonstrate its incoherence? We return to this point at the end of the present section.
Here then is the first interpretation: “Assume that I am being literally deceived at this moment. It follows that I now exist (...). For deceiving is a relation between a deceiver and a deceived; and given that I am now being deceived it necessarily follows that as the deceived, I exist at this moment.” Clearly this does not yield the desired result, that I know occurrently that I now exist, but the weaker conclusion that I now exist.

Both the passages quoted from Khamara make the same complaint, then, and at first sight this complaint may seem strange. One would think that proving something to be the case is pretty good as a way of coming to know that it is so, whereas Khamara has somehow managed to ‘up the ante’: to join the ranks of those who know that \( p \) it is apparently no longer enough to prove conclusively that \( p \) – now you have to prove that you know that \( p \), where this is being conceived of as a further task not automatically discharged by your having proved that \( p \). The complaint raises a special difficulty for a specifically epistemic (as opposed to doxastic) version of autoepistemic logic – essentially as in Section 1 above but with everything said there about ‘\( B \)’ (‘I believe that’) said instead about ‘\( K \)’ (‘I know that’). Having concluded that \( p \), one should be able to prefix a ‘\( K \)’ without further ado. In this transition, the interim conclusion (that \( p \)) has to be understood as having a particular status: roughly, in terms of the distinction between two types of premisses drawn in the following paragraph, it must not depend on anything which has the status of a mere supposition. This is precisely what we take the Evil Demon hypothesis, construed literally, to be: the supposition that some being is deliberately deceiving me. Descartes points out that this requires my existence, so that from the supposition that I am being deceived it follows that one thing I can’t be being deceived about is the belief, should I have such a belief, that I do indeed exist. Khamara in effect objects that while I must indeed exist to be deceived, I do not need to know I exist in order to be deceived. We could make the point in terms of the principle that a valid argument whose premisses are known can be used to give a conclusion which is known, in which terms the trouble is of course that here the premiss (“There is an Evil Demon deceiving me”) is not known. We shall see that there is a way of construing what Descartes is doing which is neither exegetically far-fetched nor vulnerable to any objection along these lines.

To introduce this construal, we must first distinguish two ways in which premisses can function in arguments, let’s say (i) as expressing (what the proponent of the argument takes to be) given data, by contrast with (ii) as expressing suppositions. The typical use of premisses of the second sort may occur in the course of an argument (in the typical case) only later to be discharged, so that the eventual conclusion does not actually depend on them. The word ‘assumption’ doesn’t help much to mark out either category since it covers both: there are one’s fundamental assumptions, taken to be true, and then there are assumptions ‘for the sake of argument’. The latter assumptions, the ones naturally enough I am calling suppositions, do not themselves have to be known to be true for the argument to yield, given knowledge of its premisses of the former sort—data-premisses—knowledge of the conclusion. It is here that we make a connection with the ideas of autoepistemic logic, alluded to lately in connection with the propriety of prefixing a ‘\( K \)’ operator. After illustrating the difference between the two kinds of premiss, we will consider the use of the distinctively autoepistemic rule legitimating such to statements derived exclusively from data-premisses. (The question will arise as to whether the correct autoepistemic principles for prefixing “I know that” differ from those which were our focus in Section 1 for “I believe that”, but we defer this until Section 4.)

By way of example, take as a premiss expressing (part of) my data “Kim’s sister owns a Fiat”. I can then argue to the conditional conclusion “If Kim’s sister comes to the party then someone who owns a Fiat will be at the party” by first making the supposition that Kim’s sister will come, combining this with my initial premiss to get that Kim’s sister both owns a Fiat and will come, and hence later be at the party, so that someone who owns a Fiat will be at the party”, so finally, discharging the supposition “Kim’s sister
comes to the party” by the rule that natural deduction systems – exercises in which this little example is meant to recall – often call conditional proof, we get “If Kim’s sister comes to the party, someone with a Fiat will be at the party”. If I do indeed know what I described as my initial premiss to be true, it seems I can use this argument to gain knowledge of—or to justify my claim already to have knowledge of—its eventual conclusion. It would be no objection to this procedure to say “But you supposed that Kim’s sister would come to the party, not that you knew she would come, and in fact perhaps if you had even so much as suspected she might come, she would have stayed away”.

Returning to Descartes, we offer the following (hardly novel) reconstruction. “Either I am deceived about my existence or I am not. Suppose I am. Then I must exist to be deceived. Suppose I’m not. Then I can take my pre-sceptical impression that I exist as unendangered, and I do exist. But either I am being deceived or I am not. Since either way, it follows that I exist, it must be that I exist.” For this argument to yield knowledge of its conclusion by the ‘knowledge in, knowledge out’ principle we’ve been discussing, I have to know the truth of the relevant case of the law of excluded middle. But I do not have to know the truth of either of its disjuncts, let alone both, even though both function essentially as premisses during the course of the argument, because their premiss role is as temporary suppositions later discharged (by ‘or-elimination’: see Section 4 for the general logical background). Now you may doubt that Descartes is in a position to have the relevant logical knowledge (the law of excluded middle), and Khamara certainly seems to doubt this, and there is – in the above presentation at least, another bit of knowledge needed, as to the nature of one’s pre-sceptical impressions– of which knowledge also seems needed. But what you cannot complain about is that at the point at which it is made, the supposition that I am being deceived is something I would have to have knowledge of in order to get epistemic mileage in this way out of the argument. Yet this seems to be precisely Khamara’s complaint. We can do the reconstruction even more economically if all we are interested in is the survival of Cogito considerations in the face of the Evil Demon hypothesis. Having just put yourself with the aid of the former into a position to say ‘I know that I exist’, the question is raised whether you have not overlooked here certain epistemic possibilities which are such that if they obtained it would be false that you existed. That is the way to question whether you really do know. Well, the possibility of deception by an Evil Demon is one such epistemic possibility: if it obtained, what then? Well, if it obtained, you reason – along Descartes’ lines – then you would still have to be in existence, and conclude that your claim to know you exist survives a consideration of the possibility envisaged. In reasoning “along Descartes’ lines”, you make the supposition that you are being deceived and see what follows about your existence. You do not make the gratuitously epistemic supposition that you know you are being deceived, and nor does the falsity of the latter supposition in any way prevent the reasoning from giving you – or in bolstering up your – knowledge of your own existence.

With all of this, Khamara would presumably find a problem: how does the logic manage to get in? If the Evil Demon is doing the job properly, the appearance of having one’s hands on a logical truth or of reasoning in accordance with reliable logical principles might be part of the grand deception and so cannot be trusted. So – the objection would proceed – no such argument as I have ascribed to Descartes in the period of doubt is after all available to him. The Cogito argument is an argument (this is the ‘inferential’ interpretation, after all) and no argument can defeat the hypothesis that one is being deceived about the truth of its conclusion by a being with deceptive powers so pervasive as to include the ability to deceive one about the validity of arguments to such conclusions. But this—whatever one might think about it—seems a different problem from the one raised by the passages quoted from Khamara above. is a different problem from the one we have been trying to solve: the question of how Descartes can be interpreted as arguing successfully to an epistemic conclusion in spite of ignorance in respect of such matters as whether he was dreaming, or being deceived by an Evil Demon.
4. Premises as Suppositions vs. Premises as Given Data

The most straightforward formal setting in which to analyze arguments such as that about Kim’s sister and the Fiat, or Descartes and his own existence, given in the previous section, is a natural deduction system in the style of Lemmon [1965], but with the following refinement. Initial assumptions are marked explicitly either as data-premisses or as suppositions. Some familiarity with such systems is assumed here. Where the following would count as a Lemmon proof of $q \rightarrow (p \land q)$ from the assumption $p$:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1) $p$</td>
<td>Assumption</td>
</tr>
<tr>
<td>2</td>
<td>(2) $q$</td>
<td>Assumption</td>
</tr>
<tr>
<td>1,2</td>
<td>(3) $p \land q$</td>
<td>1, 2 $\land$I</td>
</tr>
<tr>
<td>1</td>
<td>(4) $q \rightarrow (p \land q)$</td>
<td>2–3 $\rightarrow$I</td>
</tr>
</tbody>
</table>

we now envisage a version of the proof looking like this:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1d</td>
<td>(1) $p$</td>
<td>Data-premiss</td>
</tr>
<tr>
<td>2s</td>
<td>(2) $q$</td>
<td>Supposition</td>
</tr>
<tr>
<td>1d,2s</td>
<td>(3) $p \land q$</td>
<td>1, 2 $\land$I</td>
</tr>
<tr>
<td>1d</td>
<td>(4) $q \rightarrow (p \land q)$</td>
<td>2–3 $\rightarrow$I</td>
</tr>
</tbody>
</table>

We also make room for various autoepistemic moves, such as the prefixing of a ‘K’, for ‘I know that’ to any formula at a line on which it occurs as depending solely on data-premisses, and not on any mere suppositions. (Thus all line numbers on the far left should be accompanied by a ‘d’ marking.) Call this ‘KI’ (“I” for “introduction”). This is because we think of such premisses as representing part of the reasoner’s information, taken as given. Such premisses cannot be discharged by the assumption-discharging rules $\rightarrow$I (alias Conditional Proof) and $\lor$E.\(^{20}\) (We can think of the Reductio ad Absurdum rule of Lemmon [1965] as replaced by the first of these rules by treating the negation of a formula $\varphi$ as being the formula $\varphi \rightarrow \bot$, as mentioned in Section 2.) The following, for example, would not be a correct proof of $p \rightarrow Kp$ as depending on no assumptions:

(Incorrect Proof)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1d</td>
<td>(1) $p$</td>
<td>Data-premiss</td>
</tr>
<tr>
<td>1d</td>
<td>(2) $Kp$</td>
<td>1, KI</td>
</tr>
<tr>
<td></td>
<td>(3) $p \rightarrow Kp$</td>
<td>1–2, $\rightarrow$I</td>
</tr>
</tbody>
</table>

The application of $\rightarrow$I at line (3) purports to discharge an assumption which is a data-premiss, and so is incorrect. In Section 1, we noted that $p \rightarrow Bp$ was not an element of every stable set, which is the reason that a would-be proof along the above lines does need to be blocked somewhere. (We shall address the suspicion presently that the material in that section needs to be reconsidered if we move from ‘B’ to ‘K’.)

\(^{20}\) Forbes [1990] presents a descendant of the natural deduction system of Lemmon [1965], in which those formulas Lemmon would call assumptions but which do not get discharged during the course of the proof ‘premisses’ (actually, ‘premises’), reserving the label ‘assumption’ for what Lemmon would call (eventually) discharged assumptions. Thus Forbes’ assumption/premiss distinction is close to our supposition/data-premiss distinction (though he is not concerned with non-assumption-discharging rules such as KI here which cannot apply to formulas depending in part on suppositions). Lemmon’s objection—reasonable enough—to the use of the term ‘premiss’ to apply to (what he calls) assumptions had been that we need to distinguish the latter from the premisses of any given application of one of the rules during the course of a proof.
On the other hand, if to circumvent that restriction on discharging data-premisses, we make the first line a supposition instead, we get the following:

(Incorrect Proof)

1s  (1) $p$  Supposition
1s  (2) $Kp$  1, $KI$

(3) $p \rightarrow Kp$  1–2, $\rightarrow I$

This time the $\rightarrow I$ at line (3) is fine, but the $KI$ at line (2) violates the restriction on that rule that it not be applied to any formula depending on one or more suppositions (whether or not it also depends on data-premisses).

There are several different types of things which have legitimately been called suppositions in discussions of contrasts between suppositional status and something else.\(^{21}\) Supposition-based accounts of natural language conditionals may represent the antecedent of either an indicative or a subjunctive conditional as a supposition, in the latter case often emphasizing the idea of a belief-contravening supposition. For applications of the type encountered in Section 3, it is especially belief-compatible suppositions that are important, as is reflected in the above derivation of $q \rightarrow (p \land q)$ from $p$. The $\land$-Introduction step at line 3 pools premisses of the two types. The same phenomenon occurred in the illustrative argument of the previous section when we pooled the data-premiss “Kim’s sister owns a Fiat” with the suppositional premiss “Kim’s sister comes to the party”, in order to draw the interim conclusion, depending on both, that someone who owns a Fiat will be at the party (from which we proceeded to the final conclusion that if Kim’s sister comes to the party then someone who owns a Fiat will be at the party). This pooling would not make much sense if suppositions inconsistent with one’s information were made in the interests of seeing what would be the case if things were as thus supposed, since one would need to suspend those parts of what one actually took to be the case which are such that if the supposed course of events were to occur, then those things would not be the case (things not ‘cotenable’ with the given suppositions, in Nelson Goodman’s terminology). We are not concerned with suppositions of the latter sort here.

We have not been concerned to provide a complete set of rules for the natural deduction system entertained here – just to sketch enough to give the idea. The characterization of stability in terms of Kripke models with universal accessibility relations mentioned in Section 1 can be turned into a precisification of what completeness amounts to in the present context. A formula $\psi$ should be derivable from data-premisses $\varphi_1, \ldots, \varphi_n$ (with all suppositions having been discharged by the end of the proof) just in case whenever for any such model we have all of $\varphi_1, \ldots, \varphi_n$ true in (sic. throughout) the model, $\psi$ is true in the model. Conspicuously, we are here taking over ideas developed in connection with autoepistemic logic with ‘B’ for self-attributions of belief and applying them intact to ‘K’ for self-attributions of knowledge, but we defer comment on this until we have filled out the semantic suggestion just made. The requirement that all suppositions be discharged is somewhat onerous, and it has the technical disadvantage that not every initial segment of a proof counts as a proof. What is the semantic status of a non-terminal line such as line (3) of the second proof of this section, in which a formula is recorded as depending on a mixture of data-premisses and suppositions? We should like to record the two kinds of assumption in two distinct ‘slots’ in

---

\(^{21}\) Distinctions somewhat similar to that employed here may be found in Skyrms [1987] and Collins [1989]. While the term ‘updating’ is often used – as in these sources – for acquisition of new information, and thus for what provides a reasoner with new data-premisses, it also has numerous other uses and to avoid possible confusion, we avoid it altogether.
a considerably generalized version of the notion of a consequence relation. If we separate the two kinds of sets of assumptions by a semi-colon, we could write something like \( \Gamma_1; \Gamma_2 \) on the left and the formula in question on the right. Here \( \Gamma_1 \) houses the data-premises and \( \Gamma_2 \) the suppositions in force (i.e. undischarged) at the stage in question; thus at the line just mentioned the former is the set \( \{ p \} \) and the latter is \( \{ q \} \). The sense in which the formula concerned (namely, \( p \land q \)) ‘follows from’ these two sets semantically is that at any point in a Kripke model throughout which \( p \) is true (or more generally, all formulas in \( \Gamma_1 \) are true), and at which \( q \) is true (all formulas in \( \Gamma_2 \) are true), is a point at which \( p \land q \) is true.

In the interests of symmetry, we could allow for two slots on the left as well—suppositional conclusions and ‘data’ conclusions—and in these same interests, we might as well allow each such slot to be multiply occupied. (We subsume ‘unoccupied’ under ‘multiply occupied’ here, as with the multiple conclusion consequence relation \( \models \) in Section 1; readers who omitted that discussion may like to skip the rest of this – and the whole of the next – paragraph.) This gives us a ‘four-place’ consequence relation we shall call \( \models \), whose holding between four sets \( \Gamma_1, \Gamma_2, \Delta_1, \Delta_2 \), we notate by writing: \( \Gamma_1; \Gamma_2 \models \Delta_1; \Delta_2 \). Semantically we interpret this as saying that for any model of the kind described, and any point \( w \) in that model, if all formulas in \( \Gamma_1 \) are true throughout the model and all formulas in \( \Gamma_2 \) are true at \( w \) in the model, then either some formula in \( \Delta_1 \) is true throughout the model or some formula in \( \Delta_2 \) is true at \( w \) in the model.

We can use this framework to put some of the material in Section 1 into context. The consequence relation \( \vdash_{st} \) of our discussion there can be extracted from \( \models \) thus: \( \Gamma \vdash_{st} \psi \) just in case \( \Gamma; \emptyset \models \{ \psi \}; \emptyset \).

The sense in which the ‘Deduction Theorem’ fails is that from its being the case that \( \Gamma_1 \cup \{ \varphi \}; \Gamma_2 \models \Delta_1 \cup \{ \psi \}; \Delta_2 \), it does not follow that

\[
\Gamma_1; \Gamma_2 \models \Delta_1 \cup \{ \varphi \rightarrow \psi \}; \Delta_2
\]

Suppositionally, by contrast, the analogous transition is acceptable. From its being the case that

\[
\Gamma_1; \Gamma_2 \cup \{ \varphi \} \models \Delta_1; \Delta_2 \cup \{ \psi \}
\]

it does follow that

\[
\Gamma_1; \Gamma_2 \models \Delta_1; \Delta_2 \cup \{ \varphi \rightarrow \psi \}
\]

Suitable structural rules and principles governing the behaviour of the operator ‘\( K \)’ (or ‘\( B \)’) for this four-place setting may be recovered from §5 of Blamey and Humberstone [1991], where the interpretation of (what we are here calling) \( \models \) initially appears somewhat different. With a particular class of models in mind, the interpretation was that any point \( w \) in that model, if all formulas in \( \Gamma_1 \) are true throughout \( R(w) \) and all formulas in \( \Gamma_2 \) are true at \( w \) in the model, then either some formula in \( \Delta_1 \) is true throughout \( R(w) \) or some formula in \( \Delta_2 \) is true at \( w \) in the model. Here \( R(w) \) denotes the set of points accessible to \( w \). (And in the paper referred to, the sets \( \Gamma_1 \) and \( \Delta_1 \) are written, respectively, in subscript and superscript position to the turnstile.) However, this coincides with our current ‘throughout the model’ reading of \( \models \) when the models in question are those in which the accessibility relation is universal.

Finally, there is the matter of our apparent indifference as to which of knowledge (‘\( K \)’) and belief (‘\( B \)’) is at issue in autoepistemic reasoning. In a nutshell, the justification for this indifference is the familiar observation that from one’s own perspective, what one knows is indistinguishable from what one (firmly) believes.\(^\text{22}\) Alloepistemically, the logics of ‘\( K \)’ and ‘\( B \)’ are plausibly distinguished in view of the obvious

\(^{22}\) Compare Halpern [1996], p.488, speaking of “the intuition that the agent believes that his beliefs coincide with his knowledge”; there is something inappropriately general about the content of this supposed belief (for which reason the formulation in the text seems preferable) but the gist is clearly on the right track.
point that \( \varphi \) follows from \( K\varphi \) but not from \( B\varphi \), and the less obvious point that under suitable idealization \( B \rightarrow B\varphi \) follows from \( \neg B\varphi \), while under no suitable idealization does \( K \rightarrow K\varphi \) follow from \( \neg K\varphi \). Let us address these two points in turn. Even with the four-place consequence relation lately considered, the former difference between knowledge and belief cannot emerge, since what is ‘supposed’ is supposed as holding at some point in the model truth throughout which corresponds to belief. Ordinarily, one would not think of \( \varphi \) as following from the mere supposition that \( B\varphi \), but this is because it is belief-contravening suppositions—requiring suspension of the contradictory belief to the effect that \( \neg B\varphi \) on pain of having everything follow—that one has in mind, rather than the belief-compatible suppositions we have been dealing with. (To accommodate such suppositions, models with points outside of the cluster of mutually accessible points would need to be considered, as in Moore [1988], pp.112–116.) The second issue, sometimes called negative introspection, needs to be treated with care.

Any one of several epistemic/doxastic principles might be referred to as expressing an assumption of negative introspection; in particular any one of the following (we continue the numbering from Section 1):

\[
\begin{align*}
(9) & \quad \neg B\varphi \rightarrow B\neg B\varphi \\
(10) & \quad \neg B\varphi \rightarrow K\neg B\varphi \\
(11) & \quad \neg K\varphi \rightarrow K\neg K\varphi
\end{align*}
\]

While (9) and (10) represent aspects of idealized introspective powers on the part of the agent whose knowledge and belief are ascribed with the operators ‘\( K \)’ and ‘\( B \)’, (11) does not represent a plausible idealization in this direction, for the following well-known reason. Once belief is distinguished from knowledge in even the one crucial respect that what is believed by an individual (however rational) need not be true while what is known must be, the possibility arises that such an individual may erroneously believe themselves to know that \( p \), the error not arising from any failure to recognise the belief that \( p \) to be amongst their beliefs, but from the failure of that belief to constitute knowledge, on the grounds that it is false. They are thus in the position of not knowing that \( p \) while also—contradicting (11)—not knowing that they do not know that \( p \) since they (falsely) believe that they do know that \( p \). This argument against (11)—or against 5 (or ‘the S5 principle’) to give it its more customary name—may be found at various points in the philosophical literature, starting with Hintikka [1962], p.106; the argument is more fully spelt out in Lennsen [1978], p.79, and further appeals and references to it may be found at p.187 of Humberstone [1988], p.32 of Williamson [1990], and Halpern [1996], pp.484–485.\(^{23}\) The last paper gives a full

---

\(^{23}\) In note 5 (p.187) of Humberstone [1988], authors coming from the computer science field are accused of being “blissfully ignorant” of the objection just rehearsed to S5 as a logic of knowledge. In the course of pressing this – somewhat overstated – point, the paper Halpern and Moses [1985] was misattributed to D. Harel and A. Pnueli; though overstated, since the literature complained of does indeed make reference to work which the objection appears, something of the point does seem to be correct: Fagin and Halpern [1988], for example, refer to S5 as the “classical logic of knowledge” – suggesting that the objection cannot really have been taken to heart. Halpern’s reaction to this (Fagin et al. [1995], p.8; Halpern [1996], p.492) is that there is no single interpretation for words like ‘knows’ and ‘believes’ and accordingly no one all-purpose-correct epistemic-doxastic logic. Certainly Moore himself is not vulnerable to any accusation of ignorance on this score, since the relevant anti-S5 point is explicitly made by him in note 7 of his [1985], as it appears in the reprint in Moore [1995] (this being note 8 of the original publication). The argument described in the text would seem, however, to work on any reasonable interpretation of the key terms provided that they are distinguished at least to the extent of acknowledging that knowledge requires truth while belief does not. It seems to the present author, that when the S5 principle (our (11)) is assumed in discussion of various problems and puzzles concerning knowledge, the context is either autoepistemic (‘internal’), and one which, when externally described, will be seen to require nothing stronger than (10), or else one in which it is part of the description of
discuss decision of the logical situation as well as presenting an ingenious way round the objection by restricting the principle $K\varphi \rightarrow B\varphi$ it invokes: this principle is no longer endorsed for arbitrary $\varphi$ — only for $\varphi$ free of occurrences of ‘B’ and (especially) ‘K’. This then blocks the argument sketched above not, as the rejection of (11) would, in the transition from $\neg Kp$ (itself derived from the supposition that $\neg \varphi$) to $K\neg Kp$, but instead in the transition from this to $B\neg Kp$ (which then conflicts, given the consistency of the rational believer’s beliefs with the hypothesis that $BKp$): here $\varphi$ would be $\neg Kp$, and so not a permissible instance of the schema $K\varphi \rightarrow B\varphi$.\footnote{Halpern’s system is thus not closed under Uniform Substitution (of arbitrary formulas for propositional variables (or ‘sentence letters’)), depriving it of the status of a ‘logic’ in the eyes of some; the present author feels that the idea of requiring such closure for something to count as a sentential logic, as opposed to a mere theory (according to some such logic), is misguided: the intended distinction can be marked by imposing on logics the weaker condition—satisfied by Halpern’s system—of closure under (uniform) variable-for-variable substitution. There is, it should also be mentioned, a semantic characterization of the system in Halpern [1996], going some way toward dispelling the impression the restriction on $K\varphi \rightarrow B\varphi$ may create of its being an ad hoc syntactic filter.} We are not here in the business of evaluating Halpern’s proposal for alloepistemic logic, but of seeing what becomes of the original objection to negative introspection, without any such restriction in force, in the setting of autoepistemic logic.

The crucial point about that argument was our judgment that, given the falsity of $p$, the subject can believe (that they know) that $p$, but cannot know that $p$. But this possibility is not entertainable from the first-person perspective in other than in the autoepistemically irrelevant belief-contravening mode. Let us write (11) in disjunctive form:

$$(12) \quad K\varphi \lor K\neg K\varphi$$

We already know from our discussion of (5) and (6) in Section 1, that it is not just that the rational introspectively well-informed believer will believe the disjunction (12), but, assuming that the Stalnaker conditions (+) and (–) are as reasonable for “know that” as they are for “I believe that”, such a subject will actually believe one of the disjuncts. For the subject who believes that $\varphi$ will believe $K\varphi$ (“I know that $\varphi$” – recall that we are not thinking of unconfident belief here), and the subject who does not believe that $\varphi$ will be aware of this non-belief and thus know that it does not know that $\varphi$. This time there has been no fallacious passage from ignorance to knowledge of ignorance (as in the above discussion of (11)), because the ignorance arises not from the falsity of what is not known but from the (introspectively available) fact that it is not even believed. For this point, which is basically the (–)-part of the assumption that (+) and (–) remain plausible conditions when rewritten with ‘K’ in place of ‘B’, it is to be noted that we continue to regard the stable set in question as the set of a rational agent’s beliefs. It may initially seem odd to suggest that all instances of (12) belong in every rational set of beliefs (since that is how we think of the stable sets) even though not all instances of (12) need be true. But this is nothing new: we already met the phenomenon with the schema $B\varphi \rightarrow \varphi$ in Section 1. The ‘$\rightarrow$’ here may be confusing, so let us recall that what is at issue is simply $\neg B\varphi \lor \varphi$, and a stable set must contain not just the disjunction but one or other of the disjuncts.
(Cf. our discussion of (5) and (6) in that section.) The misleadingness of the \( \rightarrow \) comes from the suggestion that the rational individual is required to assent (for arbitrary \( \varphi \)) to "If I believe that \( \varphi \), then \( \varphi \)”, or perhaps to "If I believed that \( \varphi \), then it would be that \( \varphi \)”, whose antecedents invoke potentially belief-contravening suppositions which are simply beside the point here and suggest, respectively, a high subjective probability for \( \varphi \) conditional on the supposition that I believe that \( \varphi \), and that worlds like the actual world as much as the supposition that I believe that \( \varphi \) allows are all worlds in which that belief is true.\(^{25}\) But such probability-shifting and world-shifting considerations are not relevant: the autoepistemically pertinent "I" is me in the actual world with my actual beliefs. (As is often remarked à propos of Moore’s Paradox, touched on in Section 2, there is nothing in the least odd about thinking that it might have been true that \( \varphi \) even though I did not believe that \( \varphi \).) Thus the difference, prominent as it is in alloepistemic logic, between the (there) unacceptable principle \( B \varphi \rightarrow \varphi \) and its acceptable cousin \( K \varphi \rightarrow \varphi \), has no autoepistemical counterpart. So too, we have argued, for the difference between the alloepistemically acceptable (9) and its unacceptable cousin (11).

These remarks on negative introspection conclude our foray into autoepistemic logic. As was made clear in our opening section, definitive resolution of issues in epistemology has not been our goal. We have sought merely to illustrate, with the examples presented in Sections 2 and 3, that some such issues – manifesting a distinctively first-person perspective – are fruitfully approached with this formal apparatus in mind.\(^{26}\)

Appendix: Diachronic Autoepistemology?

As the second paragraph of our Introduction suggests, when this paper was being written (in 1996–7) it was thought of as introducing the term “autoepistemology”. In the meantime, Hild [1998] has appeared, using the same terminology for roughly the same area. There are two main differences between the present usage of the term and that of Hild. The first, an unimportant difference, is that Hild considers degrees of belief (subjective probabilities) rather than, as we do, belief as an all-or-nothing matter. The second, more significantly, is that Hild considers not only beliefs about one’s present beliefs but also beliefs about one’s future beliefs, as an autoepistemological matter. They are distinguished as synchronic and diachronic autoepistemology, respectively. Here we discuss the wisdom of including these diachronic considerations under the same umbrella as the synchronic considerations. (The negative tenor of the following remarks is not intended to suggest any criticism of Hild’s discussion of the synchronic case.)

A paradigm case of a topic calling for autoepistemic treatment – and a focus of our discussion in Section 2 – is Moore’s Paradox: the oddity of believing anything of the form ‘A but I do not believe that A’. (We change schematic letters from “\( \varphi \)” to “A” for alignment with Hild’s notation.) Although not itself a contradiction, principle (3) from Section 1 converts it into one by licensing the transition from the first conjunct to “I believe that A” which now does explicitly contradict the second conjunct. If Hild can convincingly show that something appropriately parallel to Moore’s paradox arises in the diachronic case,

\(^{25}\) For simplicity here I assume the correctness of something like the comparative treatment of indicative and subjunctive conditionals in Jackson [1987]; the formulation in the text is based on Jackson [1981], where there is a concerted effort to discern a common element underneath the contrast between the two types of conditionals.

\(^{26}\) For helpful comments on this material I am grateful to Steve Gardner, Brian Weatherson, Joe Halpern and Wlodek Rabinowicz.
that would show that his broader use of the term ‘autoepistemology’ to cover both cases was justified. Let us briefly consider his attempt to do this. Hild opens his section on diachronic autoepistemology with the following words (p.346):

Imagine that you are certain that would will believe A tomorrow. Then you are should already believe A today since you would otherwise adopt an opinion of which you are certain that it will be outdated tomorrow.

This does not seem very promising. You could be certain that you will believe A tomorrow because you are certain that attempts by another party to get you to hold this – perhaps false – belief will be successful. You are not now certain that it will be “outdated” tomorrow if that means it will no longer be the appropriate state of belief to be with regard to A. Hild edges toward the crucial test – Moore’s Paradox – of the continuity of the diachronic with the synchronic case thus (p.347):

On the other hand, if you believe A, then you must not be certain that you will disbelieve it tomorrow no matter what. If your beliefs are to have any value for handling future events, you should not content yourself with such ephemeral opinions. (…) For similar reasons, you should only adopt an opinion of which you are certain that you will still entertain it tomorrow. This would be a diachronic version of a synchronic Moore incoherence (believing, firstly, that you do not believe that A, but also, secondly, that A).

But is this a diachronic version of Moore’s Paradox? There is surely nothing in the least paradoxical about thinking: A (as I realise now), though tomorrow I will (unfortunately) not believe that A. (Note that “A” should be thought of as representing a tenseless rather than a tensed sentence, to keep extraneous matters out of the picture.) The change to a non-present time is as good as a change from first to third person: A, though Tom does not believe that A. The fact that the non-present time is a later time is not in this respect importantly different from making it an earlier time, and not even Hild is suggesting that there is also a ‘diachronic version’ of Moore’s Paradox in the reverse direction, i.e., of the form: “A, though yesterday I did not believe that A”. There is no real difference here. One can believe that know that one will forget something just as easily as one can believe that one has learnt something. Any special connection the one direction may have which the other lacks in respect of the matter alluded to in the above quotation as “value for handling future events” does not mean that the two directions are on a par in respect of (the failure of) Moore-paradoxicality.

In the passage ellipsed from the quotation given, just after this reference to handling future events, Hild refers back to an earlier example. The reference is presumably intended to be to that discussed on p.327 in the following words. (There is a reference to ‘p.6’ and this is indeed the sixth page of the paper.) The discussion follows mention of van Fraassen’s principle of Reflection, which we do not need to state for present purposes:

Reflection is the probabilistic version of an idea that has first been proposed in the context of full belief: We should already fully believe today what we are convinced we will fully believe tomorrow. What is more, it also generalizes the converse principle suggested by Binkley [1968] in connection with the Surprise Examination Paradox: Ideally, we should only fully believe today what we are convinced we will believe tomorrow. Binkley argues that, although only ideally rational agents could always meet these demands, they are necessary for successfully planning the future. Suppose that I am about to leave my house in the morning. I am also certain that in the evening I will be justified to believe in dry weather. Binkley’s Principle implies that I should already in the
morning believe that it will not rain in the evening. Otherwise, I might end up carrying an umbrella with the firm belief that it will not rain.

The tricky word in this passage is “justified”, as it occurs in the reference to my being certain now that I will be justified in believing this evening that the weather will be dry. One might wonder why it is there at all, since the principle in question speaks only of what we are convinced we will believe at a later time, not of what we are convinced we will justifiably believe at that time. But the emphasis on ideal rationality presumably excludes not only unjustified beliefs now, but unjustified beliefs later, and if we also build in, under this heading, present knowledge that any later beliefs will be justified beliefs, then we can see that the agent’s beliefs about its own later beliefs are equally well described as beliefs about its own later justified beliefs. But in any sense in which justified beliefs are the only beliefs a rational agent forms, justified beliefs need not be true beliefs. So I can perfectly well this morning justifiably believe both that it will rain tonight and that I will this evening justifiably – though falsely – believe that it will not rain. (Say, because I know I will encounter misleading though compelling evidence and also know that I will not then remember my current realisation of its misleadingness.) Whether, if this rather farfetched but certainly conceivable morning situation arises, I should take my umbrella, depends on various factors not specified. It should not be assumed that a decision to take the umbrella under these circumstances is bad planning simply because I will end up carrying an umbrella at some time this evening while at that time thinking it will not rain. If I correctly predict (in the morning) that while there will a cost in utility in being in that state, it will not cause me to discard the umbrella and it will be outweighed by the benefit of being able to use the umbrella when the rain finally starts, then taking the umbrella is good planning. Recall that I am convinced at the time of the planning, that the rain will eventually start to fall, despite my later conviction that it won’t.

The reader may not agree with these objections to the claim that rational planners need what Hild calls Binkley’s Principle. The point remains that whether violations of it constitute irrationality or, as we have been suggesting, mere inconvenience of circumstances, these considerations are not at all continuous with those raised by Moore’s Paradox. The only adverbs – such as “rationally” or “justifiably” taken in somewhat eccentric senses – that can fill the “X-ly” slot in the claim that there is something odd about a belief of the form “A, though I will later believe X-ly that not-A”, are those for which believing X-ly that something is so entails that it is indeed so, and for these cases (1) the oddity is not Moore-paradoxicality but straightforward inconsistency, and (2) the same oddity would be present if the first-person pronoun were replaced by a reference to any other thinker. There is, in the end, nothing particularly “auto” about Hild’s diachronic autoepistemology.

References


