Let $\Delta$ be an unspecified affirmative sentential operator (where an affirmative sentential operator is informally characterized as one for which a principle of distribution over conjunction is not obviously implausible). Examples of such operators include the familiar modal operators $\Box$ and $\Diamond$, and also various epistemic and doxastic operators. I want to consider certain principles that may be held to govern certain of these operators. One of these, just alluded to, is

1) Dist'&': $\Delta(\emptyset \& \psi) \rightarrow \Delta\emptyset \& \Delta\psi$.

$\Box$ and $\Diamond$ conform to this principle in all 'normal' systems of modal logic, and it seems reasonable to suppose that a knowledge operator and a belief operator should as well, at least when it is 'ideally rational' belief that we attempt to characterize by means of the belief operator.

The distribution principle above in fact follows from two other principles that together determine the class of normal systems of modal logic, and find frequent application to epistemic and doxastic operators. They are

2) $\Delta$-nec: $\vdash \emptyset \Rightarrow \vdash \Delta\emptyset$;

and

3) Dist'→': $\Delta(\emptyset \rightarrow \psi) \rightarrow \Delta\emptyset \rightarrow \Delta\psi$.

Together they express the deductive closure of $\Delta$. And again, the beliefs and knowledge of ideally rational agents can plausibly be held to meet this standard.

Also commonly invoked, for such agents, is a consistency condition,

4) Consis: $\Delta \neg \emptyset \rightarrow \neg \Delta\emptyset$. 
This condition fails for some unusual, but normal, systems of modal logic. And where one is concerned to represent the beliefs of actual, non-ideal, agents, the consistency condition is often felt to be inappropriate; although its failure, even in such cases, is not uncontroversial.

Let 'Ka' represent the knowledge operator, indexed to an agent a. Then 'Ka ø' stands for the statement 'a believes that ø'. If knowledge implies truth (as is commonly held); i.e., if Ka ø → ø, then the consistency condition holds for 'Ka', regardless of the ideality of a's epistemic state. That is, the veridicality condition

5) Ver: Δ ø → ø

implies Consis (assuming that Non-contradiction is among the laws of the propositional logic to which the theory of Δ is appended; for definiteness I will henceforth assume that our basic propositional logic is classical).

Let 'Ba' be the belief operator, construed analogously to the knowledge operator 'Ka'. Veridicality is clearly an inappropriate condition on 'Ba'—even ideally rational agents will typically harbour some false beliefs.

We should observe that the veridicality principle is quite different in nature from the first four principles mentioned above. They [(1) - (4)] jointly assert that Δ reflects the logical character of, and relations among, propositions in its scope; that Δ generates filters on the algebra of propositions. Veridicality, on the other hand, has it that Δ preserves not just logic, but truth—Δ accurately reflects non-logical facts about the world.

The desire to represent the natural reasoning of more realistically conceived agents leads to various attempts to modify the properly logical principles. But I am concerned here to argue for limits on the inferential obligations of ideally rational agents. Accordingly, I want to focus attention on the veridicality principle, and various of its consequences and congeners, since these extend the closure of Δ beyond what is logically required.
Before turning to these, I should mention that the ideally rational agent, as I conceive it, need not be what Sobel has termed an 'ideal intellect'.\footnote{See J. H. Sobel, 'Self-Doubts and Dutch Strategies', \textit{Australasian Journal of Philosophy} 65 no.1, March 1987.} Lapses of memory, for example, do not necessarily impugn one's \textit{rationality}, although they do constitute a kind of intellectual shortcoming. Another element of Sobel's ideal intellect, namely, perfect self-awareness, also seems too much to demand of a rational agent. More on that shortly.

The 'non-logical' conditions on $\Delta$ we shall be considering are located in a kind of hierarchy under entailment, when the logical conditions, 2), 3), and 4) are assumed. The strongest is Redundancy,

\begin{align*}
6) \text{Red:} & \quad \emptyset \leftrightarrow \Delta \emptyset,
\end{align*}

which, of course, entails both 5) and its converse. Redundancy also entails Strong Reflection,

\begin{align*}
7) \text{St.Refl:} & \quad \Delta (\emptyset \leftrightarrow \Delta \emptyset),
\end{align*}

which in turn yields both 'ordinary' reflection, and a 'converse' form of it:

\begin{align*}
8) \text{Refl:} & \quad \Delta (\Delta \emptyset \rightarrow \emptyset),
\end{align*}

and

\begin{align*}
9) \text{ConRefl:} & \quad \Delta (\emptyset \rightarrow \Delta \emptyset).
\end{align*}

Reflection also follows from Veridicality, and likewise Converse Reflection from the converse to Veridicality.

For a knowledge operator, as already remarked, Veridicality is appropriate. Its converse is an omniscience condition, which even ideally rational knowers cannot be expected to satisfy. Strong Reflection must also fail for 'Ka', since it, together with 5), entails Redundancy and omniscience. The analogues, for 'Ba', of none of these conditions is reasonable.

It is with Reflection that one first encounters a principle which might with some plausibility be argued to hold of ideally rational belief. For conformity to it guarantees that agents won't be susceptible to a peculiar sort of Dutch book argument.
(or 'Dutch strategy', as it is sometimes termed, to distinguish it from the classical Dutch book arguments). 2

The basic idea is this. Suppose an agent, a, were to hold beliefs of a 'Moore-paradoxical' sort; say, $Ba(\neg \varnothing \& Ba\varnothing)$. Then the agent should be willing, other things equal, to bet on the proposition that $\neg \varnothing \& Ba\varnothing$. If either conjunct is false, the agent loses the bet. But suppose both are true. Then in virtue of believing that $\varnothing$ (and Consis), the agent does not believe that $\neg \varnothing$, and so does not recognize that he or she has won the bet, and so will not claim his or her winnings. An unscrupulous bookie can thus be assured of a profit in such a bet; which is to say, the agent who holds such beliefs is assured a loss. And to hold beliefs which expose one to such assured losses is surely (it is claimed) a departure from ideal rationality.

If Reflection is added to the logical conditions on belief (1)-4) above), the situation just described cannot arise, for $Ba(\neg \varnothing \& Ba\varnothing)$ then implies a contradiction. That is, agents who satisfy Reflection do not hold such beliefs.

The foregoing considerations do not seem to me sufficient to warrant adopting Reflection as a rationality condition on belief. However, before going into the reasons for this view, I would like to extend a bit the hierarchy of non-logical principles begun above, and take a closer look at the Moore-type beliefs that seem to cause trouble for non-Reflective believers.

Reflection has as a consequence the reduction principle

\[ \Delta \Delta \neg \varnothing \rightarrow \Delta \neg \varnothing; \] (10)

and Converse Reflection the iteration principle

\[ \Delta \varnothing \rightarrow \Delta \Delta \varnothing; \] (11)

(still assuming the logical principles 2)-4)). From (11) follows the reduction principle

\[ \Delta \neg \Delta \neg \varnothing \rightarrow \neg \Delta \neg \varnothing; \] (12)

(10) and (12) are of course both instances of Veridicality, 5) above.

---

Applied to belief, reduction and iteration principles like 10)-12) have a distinctly Cartesian flavour, together expressing a doctrine of reliable self-awareness. And they too can be employed to derive contradictions from Moore-paradoxical statements.

Let us say that a statement of the form $\Delta\psi$ is self-refuting if $\vdash \Delta\psi \rightarrow \neg \psi$. An interesting feature of the Moore-paradoxical statements, in the absence of 'non-logical' principles on $\Delta$, is that they are consistent, but self-refuting. That is, $\Delta(\neg \emptyset \& \Delta\emptyset)$, for example, is consistent, and hence, given the logical conditions on $\Delta$, so is $\neg \emptyset \& \Delta\emptyset$. But

$$\vdash \Delta(\neg \emptyset \& \Delta\emptyset) \rightarrow \neg(\neg \emptyset \& \Delta\emptyset).$$

Beliefs of this form (and related ones) would thus appear to occupy a middle ground between logical incoherence, and mere falsity. Adopting principles like Reflection, as conditions on rational belief, amounts to classing them with the logically incoherent beliefs.

Perhaps it would be well to consider an example in some detail. The one I would like to take is the Surprise Exam paradox. Everyone knows this one, in one form or another, but I'll run through it quickly anyway, if only to fix terminology and other conventions. On Friday the teacher informs the class that one day next week (Monday - Friday) an exam will take place, and that it will be unexpected until announced, on the day it is to take place. The students reason that it can't take place on Friday, since, were it not to occur by Thursday, they would be expecting it Friday, contrary to the teacher's specification. If it can't take place Friday, then it can't take place Thursday either, for similar reasons. And so forth. The students conclude that no such exam can take place. But of course it can. So, what went wrong?

Informally, I take the real structure of the paradox to be something like this:

1) The teacher's assertions (concerning the existence of an unexpected exam) are seemingly inconsistent (by the familiar line of reasoning);
So, no exam can satisfy those assertions;
But, because those assertions are 'inconsistent', they form an unsound basis for rational expectation;
So, an exam can, after all, satisfy the assertions.

Familiar presentations of the paradox obscure this structure, by having the students reason that the exam can't occur Friday, hence, it can't occur Thursday either, and so forth. But even the first step is not obvious. Imagine Thursday evening arrives, and there has been no exam. The students reason that it must take place Friday. But then it wouldn't be unexpected, and so it can't occur Friday. But if it can't occur Friday, the teacher's assertion that there will be an exam that week is false. Exam or no exam, one of the teacher's assertions is in doubt. So, there is no sound basis for expecting the exam to occur or not to occur; and so it can occur, unexpectedly, even on Friday.

More formally, we can distinguish in the teacher's assertions an existence claim, and an unexpectedness claim. They will be denoted $A_E$ and $A_U$ respectively. Let us think of the days Monday through Friday as numbered 1 - 5. For $1 \leq i \leq 5$, let $E_i$ stand for the proposition that the exam occurs on or before the $i$th day. (So, for example, the claim that it occurs precisely on the third day, Wednesday, would be written '$E_3 \& \neg E_2'$.)

We may formalize the teacher's assertions as follows:

$A_E$: $E_5$
$A_U$: $(1 \leq i \leq 5)[p(E_i|\neg E_{i-1}) < 1].$

These probabilities are, of course, rational subjective probabilities of the students. It may be thought that, to reflect true unexpectedness, the value in $A_U$ would have to be much less than 1; perhaps less than .5. However, it will suffice for our purposes that it simply not be equal to 1.

The students also, provisionally at least, take their teacher to be reliable with respect to the above assertions. (Otherwise, the paradox doesn't get off the ground.) That is, they assume

$R$: $p(A_E \& A_U) = 1.$
The inconsistency of these three statements is almost immediate.

1) \( p(E_5 | \neg E_4) < 1 \) \\
2) \( \frac{p(E_5 & \neg E_4)}{p(\neg E_4)} < 1 \) \hspace{2cm} \text{def. } p(A | B) \\
3) \( p(E_5) = 1 \) \hspace{2cm} \text{R, probability calculus} \\
4) \( \frac{p(\neg E_4)}{p(\neg E_4)} < 1 \) \hspace{2cm} 2, 3, \text{ probability calculus} \\

Thus, 'A_E & A_U & R' is inconsistent. But A_E and A_U can jointly be satisfied. And 'A_E & A_U' is all that the teacher asserts. But once the assumption of reliability is withdrawn, no expectations about the exam are warranted. What the teacher says can be true, but the students, it appears, are somehow not justified in believing it. \(^3\)

Now, the joint inconsistency of R, A_E, and A_U can be expressed

\[ R \Rightarrow \neg (A_E \& A_U); \]

that is,

\[ 13) \ p(A_E \& A_U) = 1 \Rightarrow \neg (A_E \& A_U). \]

Come Thursday evening and no exam, \( p(\neg E_4) = 1 \), and so \( p(E_5 | \neg E_4) = p(E_5) \). Thus the instantiation of A_U reduces to \( p(E_5) < 1 \). And so

\[ 14) \ p(E_5 \& p(E_5) < 1) = 1 \Rightarrow \neg (E_5 \& p(E_5) < 1). \]

---

3 It may be objected that the reliability condition R is not appropriately imputed to the students, who, after all, have non-zero probabilities for various events which would prevent the teacher's intentions from being fulfilled. If we think of N as standing for the proposition that none of these disruptive events takes place (N is a 'normality' condition), we can conditionalize R and A_U on N. The inconsistency of A_E, A_U, and R (the last two thus conditionalized), still holds, although the proof is not quite so quick. I thank Dave Sharp for an assist with that proof.
In words: if the students expect (on Thursday) the exam to occur unexpectedly on Friday, then it won't. But their expectations, although consistent, are thereby demonstrably unreliable. 'p(...) = 1', it will be remarked, distributes over '&'. That is, it can be regarded as an affirmative sentential operator, somewhat like □ or 'Ba'. Then 13) shows that the Reliability condition R is self-refuting, and 14) that it is so in virtue of its Moore-paradoxical form— it has, in effect, the form Δ(Ø & ¬ΔØ).

A probabilistic analogue of the Reduction principle 10), namely

\[ p(p(Ø)=r) = 1 \Rightarrow p(Ø)=r, \]

allows a contradiction to be derived from R. And 15) is a direct consequence of the probabilistic Reflection principle

\[ p(Ø|p_t(Ø)=r) = r. \]

So if probabilistic Reflection is adopted as a rationality condition on subjective probabilities, then the students in the Surprise Exam Paradox cannot rationally expect their teacher's assertions to come to pass, even though it is clearly a genuine possibility.

We may note in passing that probabilistic Reflection is stronger than its 'modal' counterpart, Δ(ΔØ → Ø), since expressions of the form p(...)=r may, depending on the value of r, express either affirmative or negative modalities. Thus, the reduction principle 15) corresponds to both 10) and 12) above.

The idea that the Reliability condition constitutes a departure from ideal rationality on the part of the students presupposes that they ought to know better. And this may seem plausible, since ideally rational agents will be able to reproduce the reasoning outlined above. They will know that R is self-refuting.

But knowing that R is self-refuting does not imply knowing (believing) that one holds a false belief, if R obtains. That requires knowing (believing) that R obtains; i.e., believing that one believes that what the teacher says is true. One can be expected to know (believe) that what one believes is false, in
virtue of R obtaining, only on the further assumption that one has access to one's own cognitive states.

Thus the view that one ought to know better than to hold Moore-type beliefs seems to depend on a Cartesian view of the mind, one that, as suggested above, finds expression in reduction and iteration principles on belief. That such principles guarantee agents reliable access to empirical facts renders them (in my view) counter-intuitive, as conditions on rationality. Facts about the weather, say, and facts about one's cognitive state, are logically unrelated. How can it be a requirement of rationality that beliefs about the one sort of fact yield beliefs about the other?

One might, I suppose, respond by pointing out that it is surely a departure from ideal rationality to fail to believe what one has overwhelming evidence of. For example, to fail to accept the present testimony of one's senses is (typically) irrational. And likewise, it might be claimed, with respect to the present testimony of introspection.

But this observation simply assumes that one has (or can have) overwhelming evidence of one's beliefs. And just as there are facts about the world for which sensory evidence is unavailable, so there might be facts about one's doxastic state to which one has no introspective access.

Quite apart from the philosophically suspect nature of principles that forge 'logical' links between so-called 'first-order' and 'second-order' beliefs, they are apparently contradicted by examples of clearly rational belief that fail to conform to them. van Fraassen notes that Ulysses, in anticipating that certain of his future beliefs will be false, and in taking measures to prevent himself from then acting on those false beliefs, seems eminently rational, but violates Reflection. 4 Robert Koons describes situations in which an agent

---

4 Bas van Fraassen, 'Belief and the Problem of Ulysses and the Sirens', presented to the Philosophy Department, University of Western Ontario, February 1990. It seems that van Fraassen defends Reflection not simply as a condition of rationality, but
could find him or herself, where the balance of evidence would rationally compel beliefs that violate a version of the reduction principle 12).\(^5\)

The remarks of the last paragraphs suggest that principles designed to legislate against irreflective belief are not appropriate to the characterization of ideally rational belief. But then what are we to make of the 'Dutch strategy' arguments, like the one earlier raised against the agent who believes that \(\neg \emptyset \land \text{Ba}\emptyset\)?

Well, we may observe first of all that the principles in dispute, Reflection and the iteration and reduction principles that it (and Converse Reflection) engender, are stronger than is needed to rule out Moore-paradoxical beliefs. Given the logical conditions on 'Ba', \(\text{Ba}(\neg \emptyset \land \text{Ba}\emptyset)\) is equivalent to \(\text{Ba}(\neg (\text{Ba}\emptyset \rightarrow \emptyset))\), to which Reflection stands as a contrary, not the contradictory. Its contradictory is \(\neg \text{Ba}(\neg (\text{Ba}\emptyset \rightarrow \emptyset))\), or equivalently, \(\neg \text{Ba}(\neg \emptyset \land \text{Ba}\emptyset)\). This may be regarded as a weak form of Reflection, specific to the task of banishing a troublesome type of belief, without committing one to a wholesale Cartesianism as a consequence.

A more direct, and in my view correct, response to the Dutch strategy argument, is to interpret it as revealing lacunae in the applicability of the device of treating betting behaviour as a straightforward measure of belief.\(^6\) Why may not the rational agent who holds a Moore-paradoxical belief, appreciating the situation, simply avail himself or herself of Ulysses' stratagem, and take measures to thwart would-be Dutch-style bookies (say, by rather as a component of some broader sort of intellectual virtue.

\(^5\) Robert Koons, 'Doxastic Paradoxes Without Self-Reference', *Australasian Journal of Philosophy* 68, June 1990. Koons sees his examples as undermining, not the reduction principle, but rather the constellation of principles that jointly constitute doxastic logic, of which the reduction principle is a member in good standing. I construe them as a *reductio* of reduction.

\(^6\) As suggested in R. Clark, 'Pragmatic Paradox and Rationality', manuscript, 1990.
declining the bets)? This suggestion, it should be noted, does not involve rejecting the notion (which the use of betting behaviour as a measure of belief idealizes) that behaviour is intimately implicated in belief, that having a belief entails a readiness to act in certain ways (cæteris paribus, of course). The agent who holds a Moore-paradoxical belief, and declines the Dutch-strategy book, may still be expected to behave in other respects in accordance with the belief—most obviously, perhaps, by assenting to sentences expressing the Moore statement.

In short, the family of 'non-logical' conditions on ∆ are of doubtful appropriateness to the characterization of rational belief, and the Dutch strategy argument is not sufficient reason to put aside those doubts.

Having argued against one possible application of Reflection and various related principles, I would like to conclude by briefly indicating another, where it seems to me to be entirely natural.7

In an intensional semantics for a language containing a truth predicate it would seem reasonable that the so-called Tarski sentences, sentences having the form T«∅» ↔ ∅, should not be valid, that, in fact, the biconditional should fail in both directions. This is simply a consequence of the observation that various non-logical items in a language might have meant differently than in fact they do. On the assumption that the logical devices of the language are not similarly variable (being in effect essential components of the language) the principles 1)-4) on ∆ hold for 'T'. That is, e.g.,

1') T«∅ & ⊤» → T«∅» & T«⊤»,

and so forth. Although the truth predicate fails to satisfy a Redundancy principle, the Strong Reflection principle is reasonable; i.e.,

---

7 The truth theory sketched below is developed in more detail in R. Clark, 'Truth and Indexicality', manuscript, 1991. In particular, a model theory is there given which so interprets the truth predicate as to lend it the features here discussed.
should be valid. If the underlying (propositional) logic is, as assumed at the outset, classical, then the truth predicate will also satisfy
\[ 4') \quad T\langle \emptyset \rangle \leftrightarrow -T\langle \lnot \emptyset \rangle, \]
which expresses both consistency and bivalence.

In virtue of these properties of 'T', Moore-type statements are contradictory. That is, \( \models -T\langle \emptyset \rangle \& -T\langle \lnot \emptyset \rangle \), even though \( \emptyset \& -T\langle \emptyset \rangle \) is consistent. This explains why propositions like \( \emptyset \& -T\langle \emptyset \rangle \), although consistent, are not reasonably believed (in contrast to the properly Moore-paradoxical beliefs). If the language contains the devices to represent indexical expressions, like the personal pronouns (and the model theory is appropriately adapted to them), sentences like 'I do not exist' will, relative to a context of tokening, have the same feature— although consistent, it will be the case that \( \models -T\langle \text{I do not exist} \rangle \).