Definition A knight-knave island is an island in which each native is classified as either a knight or a knave. Knights make only true statements and knaves make only false ones.

Our two main characters are a logician L who visits an island and meets a native N. Let \( k \) be the proposition that N is a knight. Whenever N asserts a proposition \( p \), the reality of the situation is that \( k \equiv p \) is true (N is a knight if and only if \( p \)). For any proposition \( p \), let \( Bp \) be the proposition that L does or will believe \( p \) — i.e., L does or will decide that \( p \) is true — the alternate notation \( B : p \) will sometimes be used to denote \( Bp \).

Definition A logician L is (always) accurate iff he never believes any false proposition. Accuracy is expressed by the belief scheme

\[
\frac{-p}{-Bp}
\]

Problem 1 An accurate logician L visits a knight-knave island and meets a native N who makes a certain statement. Once the native has made this statement, it becomes logically impossible for L to ever decide whether N is a knight or a knave (if L should ever decide either way, he will lose his accuracy). What statement could N make to ensure this?

HINT:
Suppose \( p \) is a proposition such that \( p \equiv -Bp \) is true. If L is accurate, then \( p \) is true, but L will never believe \( p \), nor will he believe the false proposition \( -p \).

Undecidability

Let \( \langle c \rangle \) denote the source code for computation \( c \), and let \( \text{halt}(c, i) \) assert that \( c \) halts on input \( i \). Let \( s \) be a computation based on input \( \langle c \rangle \) such that

\[
\text{halt}(s, \langle c \rangle) \equiv B(\neg \text{halt}(c, \langle c \rangle))
\]

(1)

If \( p = \neg \text{halt}(s, \langle s \rangle) \), then

\[
p \equiv \neg \text{halt}(s, \langle s \rangle) \equiv \neg B(\neg \text{halt}(s, \langle s \rangle)) \equiv -Bp
\]

Hence neither halting nor nontermination of \( s(\langle s \rangle) \) is believed, provided L is accurate.

Extra Credit: Interpret \( Bq \) to mean the truth value of \( q \) is both computable and true. Show that the truth value of \( \text{halt}(c, i) \) is not computable (hint: if it were, then a computation \( s \) exists which satisfies (1) above).
Definition A logician is of type 1 if:

1. He believes every tautology.
2. If he ever believes both \( p \) and \( p \Rightarrow q \), then he will believe \( q \).

NOTE: The following belief schemes are true of type 1 logicians:

\[
\begin{align*}
&\text{\( p \) is a tautology} & \Rightarrow & \text{\( B : p \Rightarrow q \)} \\
&\text{\( B : p \Rightarrow p \)} & \Rightarrow & \text{\( B : q \)}
\end{align*}
\]

Whenever \( L \) believes a collection of propositions, he will believe their consequences. For instance, the following additional belief schemes can be derived based on the indicated tautologies:

i. \((p \Rightarrow q) \Rightarrow ((p \Rightarrow r) \Rightarrow (p \Rightarrow (q \land r)))\)

\[
\begin{align*}
&B : p \Rightarrow q \\
&B : p \Rightarrow r \\
&B : p \Rightarrow (q \land r)
\end{align*}
\]

ii. \((p \Rightarrow q) \Rightarrow ((p \land r) \Rightarrow (q \land r))\)

\[
\begin{align*}
&B : p \Rightarrow q \\
&B : (p \land r) \Rightarrow (q \land r)
\end{align*}
\]

iii. \((p \Rightarrow q) \Rightarrow ((q \Rightarrow r) \Rightarrow (p \Rightarrow r))\)

\[
\begin{align*}
&B : p \Rightarrow q \\
&B : q \Rightarrow r \\
&B : p \Rightarrow r
\end{align*}
\]

iv. \((p \Rightarrow q) \Rightarrow ((q \Rightarrow p) \Rightarrow (p \equiv q))\)

\[
\begin{align*}
&B : p \Rightarrow q \\
&B : q \Rightarrow p \\
&B : p \equiv q
\end{align*}
\]

We henceforth assume that \( L \) is of type 1 and when he visits the knight-knave island and hears \( N \) assert any proposition \( p \), he believes \( k \equiv p \) where \( k \) is the assertion that \( N \) is a knight.

\[
\begin{align*}
&k = \text{\( N \) is a knight} \\
&B : \text{\( N \) asserts} p \\
&B : k \equiv p
\end{align*}
\]

Definition A logician is conceited if he believes he is always accurate (i.e., he believes \( Bp \Rightarrow p \)).

Problem 2 A type 1 logician visits the knight-knave island and \( N \) says to him: “You will never believe that I’m a knight”. The interesting thing now is that if he is conceited (i.e., he believes he is always accurate), then he will become inaccurate. Why is this?
**Definition** A logician is *peculiar* if there is some proposition $p$ such that he believes $p$ and also believes that he doesn’t believe $p$ (this strange condition doesn’t necessarily involve an inconsistency, but it is certainly a psychological peculiarity).

**Problem 3** Show that under the hypotheses of Problem 2, $L$ will become not only inaccurate, but peculiar as well!

**Definition** A *type 2* logician is one who is type 1 and believes $(Bp \land B(p \Rightarrow q)) \Rightarrow Bq$.

NOTE: The following additional belief scheme is true of type 2 logicians:

\[
\begin{align*}
\text{true} & \quad B : (Bp \land B(p \Rightarrow q)) \Rightarrow Bq
\end{align*}
\]

Being type 1, type 2 logicians also believe $B(p \Rightarrow q) \Rightarrow (Bp \Rightarrow Bq)$; to illustrate deriving additional belief schemes, we sketch a proof:

\[
\begin{align*}
\text{true} & \quad B : (B(p \Rightarrow q) \Rightarrow (Bp \Rightarrow Bq))
\end{align*}
\]

1. $((Bp \land B(p \Rightarrow q)) \Rightarrow Bq) \Rightarrow (B(p \Rightarrow q) \Rightarrow (Bp \Rightarrow Bq))$ tautology
2. $B : ((Bp \land B(p \Rightarrow q)) \Rightarrow Bq) \Rightarrow (B(p \Rightarrow q) \Rightarrow (Bp \Rightarrow Bq))$ by type 1 belief scheme
3. $B : (Bp \land B(p \Rightarrow q)) \Rightarrow Bq$ type 2 belief scheme
4. $B : B(p \Rightarrow q) \Rightarrow (Bp \Rightarrow Bq)$ by type 1 belief scheme

**Definition** A logician is *consistent* if he never believes any proposition and its negation.

NOTE: An inconsistent type 1 logician will eventually believe every proposition $q$, since $p \Rightarrow (\neg p \Rightarrow q)$ is a tautology. Also, if we take a contradictory proposition $f$ to represent false (like $p \land \neg p$), a type 1 logician is consistent if and only if he never believes $f$ (the proposition $f \Rightarrow q$ is a tautology for every $q$, hence if the logician believes $f$, he will be inconsistent).

**Definition** A logician *believes he is consistent* if he believes $\neg((Bp) \land (B\neg p))$; for every $p$ he believes: “I will never believe both $p$ and $\neg p$”.

**Definition** A logician is *normal* if whenever he believes $p$, he also believes that he believes $p$.

**Definition** A *type 3* logician is one who is normal and also type 2.

NOTE: The following additional belief scheme is true of type 3 logicians:

\[
\begin{align*}
B : & \quad p \\
& \quad \frac{B : p}{B : Bp}
\end{align*}
\]
A type 3 logician who believes \( p \Rightarrow q \) will also believe \( B(p \Rightarrow q) \) (by normality), hence will believe \( Bp \Rightarrow Bq \) (since he is type 2 he believes \( B(p \Rightarrow q) \Rightarrow (Bp \Rightarrow Bq) \)). The corresponding scheme is

\[
\begin{array}{c}
\text{true} \\
B : (Bp \land Bq) \equiv B(p \land q)
\end{array}
\]

Hence if a type 3 logician hears a native N assert a proposition \( p \), then he will not only believe \( k \Rightarrow p \) (which he will, since he will believe \( k \equiv p \)), but will also believe \( Bk \Rightarrow Bp \) (he will believe: “If I should ever believe he’s a knight, then I will believe what he said”). He will also believe \( (B\neg k) \Rightarrow (B\neg p) \).

We sketch a derivation of the following belief scheme for type 3 logicians:

\[
\begin{array}{c}
\text{true} \\
B : (Bp \land Bq) \equiv B(p \land q)
\end{array}
\]

By scheme iv, it suffices to show

\[
\begin{array}{c}
B : B(p \land q) \Rightarrow (Bp \land Bq) \quad (2) \\
B : (Bp \land Bq) \Rightarrow B(p \land q) \quad (3)
\end{array}
\]

Appealing to scheme i — and symmetry of \( p \) and \( q \) — (2) would follow from

\[
B : B(p \land q) \Rightarrow Bp
\]

But \( (p \land q) \Rightarrow p \) is a tautology, so the above is true. To show (3), note that scheme iii implies it would suffice that

\[
\begin{array}{c}
B : (Bp \land Bq) \Rightarrow (Bq \land B(q \Rightarrow (p \land q))) \quad (4) \\
B : (Bq \land B(q \Rightarrow (p \land q))) \Rightarrow B(p \land q) \quad (5)
\end{array}
\]

The latter belief (5) is an instance of the belief scheme for type 2 logicians. The former belief (4) would follow from scheme ii provided

\[
B : Bp \Rightarrow B(q \Rightarrow (p \land q))
\]

But \( p \Rightarrow (q \Rightarrow (p \land q)) \) is a tautology, so the above is true.

**Problem 4** Show that a type 3 logician believes he is consistent if and only if he believes he will never believe \( f \).

**HINT:** Let \( f = (p \land \neg p) \), and let \( c_p = \neg(Bp \land B\neg p) \) represent consistency (w.r.t \( p \)). Show that

\[
B : c_p \equiv \neg Bf
\]

by making use of the scheme

\[
\begin{array}{c}
\text{true} \\
B : (Bp \land Bq) \equiv B(p \land q)
\end{array}
\]

Note that \( B(p \land q) \) is equivalent to \( Bp \land Bq \) for type 1 logicians, and use belief (6) to finish the problem.
**Definition** A type 4 logician is one who is type 3 and also believes he is normal.

NOTE: The following additional belief scheme is true of type 4 logicians:

\[
    B : \quad Bp \Rightarrow BBp
\]

**Problem 5** A type 4 logician visits a knight-knave island (or at least he believes it to be one) and meets N who says: “You will never believe that I’m a knight”. Prove that if L is consistent, he can never know that he is (if L ever believes that he is consistent, he will become inconsistent).

HINT: Since L believes \( k \equiv \neg Bk \), it follows that \( B(Bk \Rightarrow \neg k) \), and therefore

\[
    B : \quad BBk \Rightarrow B\neg k
\]

is true (L is type 3). Use this together with \( B : Bk \Rightarrow BBk \) to conclude (via scheme iii) that

\[
    B : \quad Bk \Rightarrow B\neg k
\]

Let \( c_k = \neg(Bk \land \neg k) \) represent consistency (w.r.t k) and use scheme ii with the belief above to obtain

\[
    B : \quad Bk \Rightarrow \neg c_k
\]

If L believes he is consistent, conclude that \( B\neg Bk \). Use this together with \( B(k \equiv \neg Bk) \) to conclude \( Bk \), then use normality to show L is inconsistent.

A logician L of type 4 is thinking of visiting the island of knights and knaves because he has heard a rumor that the sulphur baths and mineral waters there might cure his rheumatism. He is home discussing this with his family physician and asks whether the cure really works. The doctor replies: “The cure is largely psychological; the belief that it works is self-fulfilling. If you believe that the cure will work, then it will”.

The logician fully trusts his doctor and so he goes to the island with the prior belief that if he should believe that the cure will work, then it will (i.e., L assumes the axiom \( B(Bw \Rightarrow w) \), where \( w \) is the proposition that the cure will work). He takes the cure, which lasts a day, and which is supposed to work in a few weeks (if it works at all). But the next day, he starts worrying: “I know that if I should believe that the cure will work, then it will, but what evidence do I have that I will ever believe that the cure works? And so how do I know that it will?”

**Problem 6** A native N passes by and asks L why he looks so disconsolate. The type 4 logician explains the situation (assume the context of the previous two paragraphs) and concludes: “...and so how do I know that the cure will work?” The native then draws himself up in a dignified manner and makes a certain statement. Amazingly enough, L will consequently believe that the cure will work, and, if his doctor was right, it will. What could N have said?

HINT: Show that if a type 4 logician believes \( Bp \Rightarrow p \), and if there exists a proposition \( q \) such that he believes \( q \equiv (Bq \Rightarrow p) \), then he will believe \( p \).
A first step is to show the following are both true and believed

\[ Bq \Rightarrow (BBq \land B(Bq \Rightarrow p)) \]

\[ (BBq \land B(Bq \Rightarrow p)) \Rightarrow Bp \]

The first follows from normality, belief in normality, \( B(q \equiv (Bq \Rightarrow p)) \), and scheme i. The second follows from the belief schemes for type 1 and type 2 logicians. Conclude that it is true and believed (via scheme iii) that \( Bq \Rightarrow Bp \).

Next show that \( B(Bq \Rightarrow p) \) follows (via scheme iii) from believing both \( Bq \Rightarrow Bp \) and \( Bp \Rightarrow p \). Conclude using \( B : q \equiv (Bq \Rightarrow p) \) that \( Bq \), and therefore \( Bp \).

A logician is unstable if there is some proposition \( p \) such that he believes that he believes \( p \), but he doesn’t really believe \( p \). A logician is stable if he is not unstable (i.e., for every \( p \), if he believes \( Bp \) then he believes \( p \)).

A certain country is ruled by a tyrant who has a brain-reading machine with which he can read the thoughts of all inhabitants. There is one particular proposition \( e \) which all inhabitants are forbidden to believe — any inhabitant who believes \( e \) gets executed! Given any proposition \( p \), we say that it is dangerous for an inhabitant to believe \( p \) if his believing \( p \) will lead him to believing \( e \).

**Problem 7** Prove that if an inhabitant is a normal and stable type 1 logician who believes it is dangerous for him to believe \( p \), then it really is dangerous for him to believe \( p \)!

**HINT:** Make use of normality and stability to show that if he believes \( Bp \Rightarrow Be \), then \( Bp \Rightarrow Be \) is true.

**Extra Credit:** A type 4 logician visits the knight-knave island because of a rumor that there is gold buried there. He meets a native and asks: “Is there really gold here?” The native then makes two statements:

1. “If you ever believe I’m a knight, then you will believe that there is gold here.”
2. “If you ever believe I’m a knight, then there is gold here.”

Is there gold on the island? Why?

**HINT:** Let \( q \) be the proposition that there is gold on the island. Note that the following are true and believed,

\[ k \equiv (Bk \Rightarrow Bg) \]

\[ k \equiv (Bk \Rightarrow g) \]

According to the hint for problem 6, the belief scheme below is true of type 4 logicians,

\[ B : Bp \Rightarrow p \]

\[ B : q \equiv (Bq \Rightarrow p) \]

\[ Bq \land Bp \]

Let \( q = k \) and \( p = g \).
1. First, $k \equiv p$ since $k$ implies that when N says $p$ then $p$ must be true. Conversely, $\neg k$ implies that when N says $p$ then $p$ must be false.

Second, let $a_p = (Bp \Rightarrow p)$ represent accuracy (w.r.t $p$), and note that if $p \equiv \neg Bp$ then

$$\neg p \Rightarrow (\neg p \land Bp)$$

because both $\neg p \Rightarrow \neg p$ and $\neg p \Rightarrow Bp$. It follows that if $p \equiv \neg Bp$ then

$$\neg p \Rightarrow \neg a_p$$

so if L is accurate, then $p$ is true, and consequently $\neg Bp$ is true so L cannot believe $p$. Moreover, he will (due to accuracy) never believe the false proposition $\neg p$.

It remains to choose $p$... since $k \equiv p$, a natural choice is $p = \neg Bk$ so that $k \equiv \neg Bk$.

2. Since $(k \equiv \neg Bk) \Rightarrow (a_k \Rightarrow k)$ is a tautology (see the previous question), L believes it. Since L believes $k \equiv \neg Bk$ (N made the assertion $\neg Bk$), it follows that he believes $a_k \Rightarrow k$. If he is conceited, then $Ba_k$ is also true, and therefore he believes $k$. However, since $B(k \Rightarrow \neg Bk)$ is true — $(k \equiv \neg Bk) \Rightarrow (k \Rightarrow \neg Bk)$ is a tautology — he must believe $\neg Bk$. Let $q = \neg Bk$, and observe that $q$ is false (L believes $k$), but $Bq$ is true.

3. By the answer to the previous problem, if L is conceited then both $Bk$ and $B\neg Bk$ are true.

4. Let $f = p \land \neg p$ and let $c_p = \neg(Bp \land B\neg p)$ represent consistency (w.r.t $p$). If L believed $c_p \equiv \neg Bf$, then $B(c_p \Rightarrow \neg Bf)$ and $B(\neg Bf \Rightarrow c_p)$ — L is type 1 — hence $Bc_p \equiv B\neg Bf$.

Note that $B(c_p \equiv \neg Bf)$ follows from $B((Bp \land B\neg p) \equiv B(p \land \neg p))$ since

$$(c_p \equiv \neg Bf) \equiv ((Bp \land B\neg p) \equiv B(p \land \neg p))$$

and L is type 1. Moreover, $B((Bp \land B\neg p) \equiv B(p \land \neg p))$ follows from

$$B : (Bp \land Bq) \equiv B(p \land q)$$

5. Since L believes to be on a knight-knave island, he believes $k \equiv \neg Bk$. Therefore $B(Bk \Rightarrow \neg k)$ is true — $(k \equiv \neg Bk) \Rightarrow (Bk \Rightarrow \neg k)$ is a tautology — and therefore (since L is type 3) so is

$$B : \quad B \neg Bk \Rightarrow B \neg k$$

Since L is type 4, $B : Bk \Rightarrow BBk$ is also true, so by scheme iii we can conclude

$$B : \quad Bk \Rightarrow B \neg k$$

Let $c_k = \neg(Bk \land B\neg k)$ represent consistency (w.r.t $k$) and use scheme ii with the belief above to obtain

$$B : \quad Bk \Rightarrow \neg c_k$$

It follows from the contrapositive of what is believed above that if L believes he is consistent, then $B\neg Bk$. Combining that with $B(k \equiv \neg Bk)$ implies $Bk$, and so (by normality) $BBk$. Therefore, $\neg c_q$ where $q = Bk$. 

7
6. We will show that if a type 4 logician believes $Bp \Rightarrow p$, and if there exists a proposition $q$ such that he believes $q \equiv (Bq \Rightarrow p)$, then he will believe $p$.

A first step is to show the following are both true and believed

$$Bq \Rightarrow (BBq \land B(Bq \Rightarrow p))$$

$$(BBq \land B(Bq \Rightarrow p)) \Rightarrow Bp$$

The first follows from normality ($Bq \Rightarrow BBq$), belief in normality ($B : Bq \Rightarrow BBq$), belief in $q \equiv (Bq \Rightarrow p)$ — thus $B : q \Rightarrow (Bq \Rightarrow p)$ since L is type 1, $B : Bq \Rightarrow B(Bq \Rightarrow p)$ since L is type 3, and $Bq \Rightarrow B(Bq \Rightarrow p)$ — and scheme i. The second follows from the belief schemes for type 1 and type 2 logicians. We can conclude that $Bq \Rightarrow Bp$ is true and believed (via scheme iii).

A second step is to show $B(Bq \Rightarrow p)$. This follows (via scheme iii) from belief in both $Bq \Rightarrow Bp$ and $Bp \Rightarrow p$.

Since $q \equiv (Bq \Rightarrow p)$ and $Bq \Rightarrow p$ are believed, it follows that $Bq$, and therefore (by what has been established above) $Bp$.

It remains to choose $p$ and $q$. Let $w$ be the proposition that the cure works, and choose $p = w$.

A hypothesis (of problem 6) is that L believes $Bw \Rightarrow w$. Choosing $q = k$ turns the other required belief ($q \equiv (Bq \Rightarrow p)$) into $k \equiv (Bk \Rightarrow w)$; hence N might say “If you ever believe I am a knight then the cure will work”.

7. Since L is normal, $Bp$ implies $BBp$ and therefore L believes $Be$ because $B : Bp \Rightarrow Be$. Since L is stable, it follows that he believes $e$.

Extra Credit:

There is gold on the island! If it can be established that $B(Bg \Rightarrow g)$, then it follows (via the hint) that $Bk$ and $Bg$, which implies there is gold on the island since the two true (and believed) statements displayed in the hint would reduce to

$$k \equiv Bg$$

$$k \equiv g$$

Tautology $Bg \Rightarrow (Bk \Rightarrow Bg)$ is believed, and both $(Bk \Rightarrow Bg) \Rightarrow k$ and $k \Rightarrow (Bk \Rightarrow g)$ are believed. It follows (via scheme iii) that $Bg \Rightarrow (Bk \Rightarrow g)$ is believed. Since $k \equiv (Bk \Rightarrow g)$ is believed and $(k \equiv (Bk \Rightarrow g)) \Rightarrow (g \Rightarrow k)$ is a tautology, $g \Rightarrow k$ is believed (the logician is type 1) and therefore $Bq \Rightarrow Bk$ is believed (the logician is type 3). The desired conclusion $B(Bg \Rightarrow g)$ follows from the scheme

$$B : p \Rightarrow (q \Rightarrow r)$$

$$B : p \Rightarrow q$$

$$B : p \Rightarrow r$$

by choosing $p = Bg$, $q = Bk$, and $r = g$. 