CHAPTER 4

On Circular Concepts

I argued in the previous chapter that circular definitions and circular concepts are legitimate. My aim in the present chapter is to clarify the method for discovering whether a concept is circular (§§4.2 & 4.4) and to apply this method to test the logical character of two ordinary concepts, namely, “rational choice” (§4.3) and “belief” (§4.4). I begin with a sketch of a semantics and logic for a special class of definitions—finite circular definitions—defined below.

4.1. Semantics and Logic for Finite Circular Definitions

Let $L$ be a classical first-order language and let $M$ be an interpretation of $L$. Let us extend $L$ to $L^+$ by adding to it the definition $D$,

$$G(x) = \text{df} \varphi(x, G),$$

where $G$ is a new predicate and $\varphi(x, G)$ is a formula that may contain occurrences of $G$ but that contains no free occurrences of variables other than $x$.¹ We saw in chapter 3 that the circular definition equips the definiendum with a rule of revision, a rule that enables us to improve a proposed extension of the definiendum. This rule can be defined as follows. Let us call possible interpretations of definiendum $G$ in $M$ hypotheses; so a hypothesis, $h$, is a subset of the domain of $M$. Let $M + h$ be the interpretation of $L^+$ that is just like $M$ except that it interprets $G$ via $h$. We can now define the rule of revision $\delta_{D,M}$ for definition $D$ in the interpretation $M$: $\delta_{D,M}$ is an operation on the power set of the domain, and for all objects $d$ in the domain,

$$d \in \delta_{D,M}(h) \leftrightarrow d \text{ satisfies } \varphi(x, G) \text{ in } M + h.$$
So the revised hypothesis, $\delta_{D,M}(h)$, consists of those objects in the domain that satisfy the definiens of $D$ under the interpretation $M + h$.

For example, consider the following definition of a one-place predicate $F$:

$$(1) \quad F(x) =_{\text{Df}} [x = \text{Socrates} \lor (x = \text{Plato} \land \neg F(x))].$$

Suppose that ‘Socrates’ and ‘Plato’ are assigned their actual denotations. Then, Socrates and Plato are the only entities that satisfy the definiens of (1) when $F$ is assigned the extension $\emptyset$. Hence the rule of revision for $F$ yields $\{\text{Socrates, Plato}\}$ for the input $\emptyset$. Similar calculations yield the following table for the revision rule of (1).

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>$\Rightarrow$</td>
</tr>
<tr>
<td>${\text{Socrates}}$</td>
<td>$\Rightarrow$ ${\text{Socrates, Plato}}$</td>
</tr>
<tr>
<td>${\text{Plato}}$</td>
<td>$\Rightarrow$ ${\text{Socrates}}$</td>
</tr>
<tr>
<td>${\text{Socrates, Plato}}$</td>
<td>$\Rightarrow$ ${\text{Socrates}}$</td>
</tr>
</tbody>
</table>
| ...            | $\Rightarrow$ ...

Even though the revision rule has a hypothetical character, it provides a basis for categorical judgments. The intuitive idea for making the transition to the categorical is this. One considers all possible hypotheses for the extension of the predicate and tries to improve them through repeated applications of the revision rule. Hypotheses that survive this process—i.e., those that are found to occur over and over again—are the ones that are deemed best by the revision rule. If a claim holds under all the best hypotheses, then it is true categorically; if it fails under all these hypotheses, then it is false categorically.

This intuitive idea can be made precise as follows. Let $\delta^n(h)$ be the result of $n$ applications of a revision rule $\delta$ to a hypothesis $h$. So $\delta^0(h) = h$; and $\delta^{n+1}(h) = \delta(\delta^n(h))$. Let us call a hypothesis $h$ reflexive iff, for some $n > 0$, $\delta^n(h) = h$. (Example: The revision rule of definition (1) yields two reflexive hypotheses, $\{\text{Socrates}\}$
and \{Socrates, Plato\}. Observe that if \(\delta^n(h) = h\), then \(\delta^{n\cdot m}(h) = h\), for all natural numbers \(m\). Hence, once a reflexive hypothesis is reached in the course of revision, it is bound to occur over and over again in subsequent stages. On the other hand, nonreflexive hypotheses, never recur in the finite stages of revision. So only the reflexive hypotheses survive the \textit{finitary} revision process; the nonreflexive ones are eliminated.

The theory of definitions would be much simpler than it is if one could always restrict attention to finitary processes. One could then declare the reflexive hypotheses to be the best ones. Unfortunately, certain definitions require transfinite applications of the revision rule, and a general theory of definitions—one that aims to be applicable to \textit{all} definitions—has to provide a treatment of them. (A significant part of Belnap’s and my \textit{Revision Theory} book is devoted to this task.) Nevertheless, for a range of circular definitions the restriction to finitary processes \textit{is} plausible. This range includes definitions that meet the following \textit{finiteness requirement}:

\textbf{Finiteness requirement on a definition} \(D\). For all interpretations \(M\) of \(L\) there is a natural number \(n\) such that, for all hypotheses \(h\) for interpreting the definiendum in \(M\), \(\delta^n_{D,M}(h)\) is reflexive.\(^2\)

Definitions that meet this requirement will be said to be \textit{finite}. Note that definitions of the following two forms are finite.

\begin{enumerate}
  \item \(G(x) =_{Df} (x = a_1 \lor x = a_2 \lor \ldots \lor x = a_m) \land \varphi(x, G)\).
  \item \(G(x) =_{Df} \psi(x) \lor [(x = a_1 \lor x = a_2 \lor \ldots \lor x = a_m) \land \varphi(x, G)]\), where \(\psi(x)\) contains no occurrences of \(G\).
\end{enumerate}

Note also that definitions whose definientia are logically equivalent, or logically contradictory, to those of (2) and (3) are also finite.\(^3\) The rules of revision for such definitions have ranges with cardinality less than or equal to \(2^m\). Hence a
hypothesis at the $2^m$-th stage of revision is bound to be reflexive. (Argument: Suppose that a hypothesis occurring at the $2^m$-th stage is not reflexive. Then no hypothesis occurring at an earlier stage can be reflexive, for the result of applying the revision rule to a reflexive hypothesis is invariably reflexive. Consequently, the hypotheses occurring up to the $2^m$-th stage must all be distinct and must exhaust the range of the revision rule. One of these hypotheses will occur at the $2^{m+1}$-th stage and thus be reflexive. We have a contradiction.)

We can now state the semantics for definitions that meet the finiteness requirement. Fix a definition $D$. And let us say that a hypothesis $h$ is reflexive in an interpretation $M$ (and $D^\dagger$) iff $h$ is reflexive for the revision rule $\delta_{D,M}$ generated by $D$ in $M$. The important notions of validity can now be defined as follows:

A sentence $\psi$ of $L^+$ is valid (on $D$) in $M$ iff $\psi$ is true in $M + h$, for all hypotheses $h$ that are reflexive in $M$.

We shall say that $\psi$ is valid (absolutely, i.e., without relativization to $M$) iff $\psi$ is valid in every interpretation $M$ of the language $L$. An argument $\varphi_1, \varphi_2, \ldots, \varphi_n \therefore \psi$ will be called valid iff the conditional formula

$$(\varphi_1 \land \varphi_2 \land \ldots \land \varphi_n) \rightarrow \psi$$

is valid. Examples: On definition (1), the following are valid in the actual state of affairs,

$$F(Socrates), \neg F(Aristotle), \text{ and } \forall x(F(x) \rightarrow x = \text{ Socrates } \lor x = \text{ Plato});$$

but these items are not:

$$F(Plato), \neg F(Plato), \text{ and } \neg F(Plato) \therefore F(Plato).$$

Let us call a sentence $\psi$ categorical in $M$ iff either $\psi$ or $\neg \psi$ is valid in $M$, and let us call $\psi$ pathological iff $\psi$ is not categorical. Different types of pathological
sentences can be distinguished. Let us say that $\psi$ is stable relative to a reflexive hypothesis $h$ iff the following condition obtains:

For all natural numbers $n$, $\psi$ is true in $M + h$ iff $\psi$ is true in $M + \delta^n(h)$.

Then, a pathological sentence is paradoxical iff it is not stable relative to any reflexive hypothesis; it is quasi-categorical iff it is stable relative to every reflexive hypothesis; and it is quasi-paradoxical iff it is neither paradoxical nor quasi-categorical (i.e., it is stable relative to some but not all reflexive hypotheses). If ‘Socrates’ and ‘Plato’ are given their actual denotations, then $F(\text{Plato})$ is pathological; in fact, it is paradoxical. If the two names are given the same denotation, then $F(\text{Plato})$ turns out to be valid. Examples of quasi-categorical and of quasi-paradoxical sentences will be found below.

The semantic notion of validity of arguments, defined above, is captured by a simple calculus—the calculus $C_0$ of §3.3. Let us say that a sentence $\psi$ is deducible in $C_0$ from the premisses $\phi_1, \phi_2, \ldots, \phi_n$, relative to definition $D$ iff a derivation of $\psi^0$ can be constructed in $C_0$ from the indexed formulas $\phi_1^0, \phi_2^0, \ldots, \phi_n^0$ on the basis of $D$. For finite definitions, $C_0$ can be shown to be both sound and complete with respect to the semantics given above. That is, it can be shown that an argument $\phi_1, \ldots, \phi_n / \therefore \psi$ is valid iff $\psi$ is deducible in $C_0$ from $\phi_1, \ldots, \phi_n$. (The proof of Theorem 5B.1 in Revision Theory can easily be adapted to establish this.)

Definitions that meet the finiteness condition have, then, both a simple semantics and an attractive calculus for working with them. Moreover, the semantics is a special case of a general theory that applies to all definitions—namely, the theory $S^\#$ presented in Revision Theory. That is, over finite definitions that meet the finiteness requirement, the semantics given is equivalent to the $S^\#$ semantics.\footnote{5}
4.2. Methods for Establishing Circularity

Let us consider critically several methods for establishing circularity. This will lead us to a method that is satisfactory.

Method #1. To show that a predicate $G$ is circular, provide a circular definition of $G$ that is *sense-adequate*, that is, provide a definition that captures the sense of $G$.

This method is perhaps the first that comes to mind. The difficulty with it is that it is virtually impossible to apply: sense-adequate definitions are unavailable for nearly all predicates of English, circular as well as non-circular.

Method #2. To show that a predicate $G$ is circular, provide a circular definition that is *intensionally adequate*, that is, provide a definition whose definiendum has, under all circumstances, the same extension (more generally, the same signification) as that of $G$.

Method #2, unlike Method #1, has some known applications. It can be used to show the circularity of truth and other semantic concepts. The Tarski biconditionals for a language $L$, that is, sentences of the form,

$\text{‘}p\text{‘} \text{ is true in } L =_{\text{Df}} p$,  

Suppose a community introduces a circular predicate $H$ into its language through definition (4).

\begin{equation}
(4) \quad H(x) \equiv_{\text{Df}} [x \text{ is gold} \lor (x \text{ is silver} \land \neg H(x))].
\end{equation}

Then (5) is not a sense-adequate definition of $H$.

\begin{equation}
(5) \quad H(x) \equiv_{\text{Df}} [x \text{ is gold} \lor (x \text{ is Ag} \land \neg H(x))].
\end{equation}

For the sense of ‘silver’ is not the same as that of ‘Ag’. But (5) is an intensionally adequate definition of $H$: ‘Silver’ and ‘Ag’ pick out the same things in all worlds and, hence, the intension of the concept defined by (5) is identical to that defined by (4).

Method #2, unlike Method #1, has some known applications. It can be used to show the circularity of truth and other semantic concepts. The Tarski biconditionals for a language $L$, that is, sentences of the form,

$\text{‘}p\text{‘} \text{ is true in } L =_{\text{Df}} p$,  

\begin{equation}
(4) \quad H(x) \equiv_{\text{Df}} [x \text{ is gold} \lor (x \text{ is silver} \land \neg H(x))].
\end{equation}

Then (5) is not a sense-adequate definition of $H$.

\begin{equation}
(5) \quad H(x) \equiv_{\text{Df}} [x \text{ is gold} \lor (x \text{ is Ag} \land \neg H(x))].
\end{equation}

For the sense of ‘silver’ is not the same as that of ‘Ag’. But (5) is an intensionally adequate definition of $H$: ‘Silver’ and ‘Ag’ pick out the same things in all worlds and, hence, the intension of the concept defined by (5) is identical to that defined by (4).

Method #2, unlike Method #1, has some known applications. It can be used to show the circularity of truth and other semantic concepts. The Tarski biconditionals for a language $L$, that is, sentences of the form,

$\text{‘}p\text{‘} \text{ is true in } L =_{\text{Df}} p$,  

\begin{equation}
(4) \quad H(x) \equiv_{\text{Df}} [x \text{ is gold} \lor (x \text{ is silver} \land \neg H(x))].
\end{equation}

Then (5) is not a sense-adequate definition of $H$.

\begin{equation}
(5) \quad H(x) \equiv_{\text{Df}} [x \text{ is gold} \lor (x \text{ is Ag} \land \neg H(x))].
\end{equation}

For the sense of ‘silver’ is not the same as that of ‘Ag’. But (5) is an intensionally adequate definition of $H$: ‘Silver’ and ‘Ag’ pick out the same things in all worlds and, hence, the intension of the concept defined by (5) is identical to that defined by (4).

Method #2, unlike Method #1, has some known applications. It can be used to show the circularity of truth and other semantic concepts. The Tarski biconditionals for a language $L$, that is, sentences of the form,
yield an intensionally adequate definition of truth. And the definition is circular
when $L$ contains the predicate ‘true in $L$’. It should be stressed that this argument for
circularity does not require the strong claim that the Tarski biconditionals yield a
sense-adequate definition of ‘true’; it requires only the weaker claim of intensional
adequacy. The stronger claim, which is characteristic of the philosophical position
known as deflationism, is in my view false.\footnote{Method #2 suffers from the same basic problem, however, as the earlier method,
namely, limited applicability. Intensionally adequate definitions are practically im-
possible to discover for most ordinary predicates.\footnote{Fortunately, however, better
methods are available.}

**Method #3.** To show that a predicate $G$ is circular, provide a
circular definition that is **extensionally adequate**, that is, provide
a circular definition whose definiendum has the same extension
(more generally, the same signification) as that of $G$.

This method is more widely applicable than the earlier ones. But we still have
a problem: the method is unreliable. Extensionally adequate circular definitions
exist for noncircular predicates also. For the purposes of illustration, assume that
‘taught Aristotle’ is true only of Plato, and consider the following definition.

\[(6) \ J(x) =_{\text{Df}} [x = \text{Plato} \land (J(\text{Plato}) \lor J(\text{Alexander}))] \lor
\quad [x = \text{Alexander} \land (\neg J(\text{Plato}) \land \neg J(\text{Alexander}))].\]

The revision rule for $J$ has exactly one reflexive hypothesis, \{Plato\}. Hence (6) is
an extensionally adequate circular definition of ‘taught Aristotle’. Observe that the
circularity in (6) cannot be eliminated by the substitution of a logically equivalent
formula for the definiens—the way it can be in (7),

\[(7) \ J(x) =_{\text{Df}} [x = \text{Plato} \land (J(\text{Plato}) \lor \neg J(\text{Plato}))].\]

The definiens of (6), unlike that of (7), is not logically equivalent to any $J$-free
formula.
This problem can be overcome if, in Method #3, we require that the definition be essentially circular. Call a definition \( D \) extensionally equivalent to a definition \( D' \) iff the definiendum of \( D \) is extensionally equivalent to that of \( D' \); that is, iff the two definienda have the same signification. Then an essentially circular definition is one that is not extensionally equivalent to any noncircular definition. Definitions (6) and (7) are not, but (1) and (4) are, essentially circular. Essentially circular definitions display, at least to some degree, the pathological behavior that is so distinctive of circularity.

The idea of essentially circular definitions needs modification before it will serve its intended function. The difficulty is that a concept may be circular and yet it may not be possible to give an essentially circular definition for it. Contingent facts may make the concept extensionally equivalent to a noncircular one. For example, the definiendum of

\[
(8) \quad K(x) =_{\text{Df}} [x = \text{Plato} \lor (x \text{ is a king of France} \land \neg K(x))]
\]

expresses a concept that is circular. But because of contingent facts (the nonexistence of kings of France), definition (8) is not essentially circular.

We can overcome this difficulty by moving to a relational notion of essential circularity, “essential circularity in a possible situation (or world or interpretation).” The method for establishing circularity can now be formulated thus:

**Method #4.** To establish the circularity of a predicate \( G \), it suffices to provide a definition \( D \) and a possible situation \( s \) such that

(i) \( D \) is essentially circular in \( s \) and (ii) in \( s \), the definiendum of \( D \) is extensionally equivalent to \( G \).

This method brings with it the advantage that it allows us to work with simple idealized situations. The advantage is important because concepts of philosophical interest—such as truth, rationality, and belief—are so rich and complex that it is
practically impossible to provide definitions that do justice to their actual signifi-
cations.

So, in order to establish that a concept is circular, we are not required to re-
veal the essence of the concept (whatever essence may be). Nor are we required to
exhibit the procedures underlying the application of the concept. All that we are
required to do is show that in one situation the extensional behavior of the concept
fits that of an essentially circular definition. We should recognize, nevertheless,
that there is advantage if the proposed definition is not list-like—if the definition
reveals principles or procedures underlying the concept. For this can increase our
confidence in the intuited behavior of the concept (especially when the behavior is
pathological), and this, in turn, can strengthen the argument for extensional ade-
quacy.

A final note: to establish the circularity of a relation $R(x, y)$ it suffices to estab-
lish, for some $x$, the circularity of $R_x(y)$—where, for all objects $y$, $R_x(y)$ iff $R(x, y)$.

4.3. The Concept of Rational Choice

I shall argue that the concept of rational choice is circular. I was persuaded of
this thesis by André Chapuis, and my argument is built on his ideas. In particular,
Chapuis has stressed the parallels between the behavior of the concept of rational
choice and circular concepts such as truth.

4.3.1 Preliminary Matters. The concept of rational choice is ternary in character:
“action $x$ is rational for an agent $y$ in a setting $z$.” (Example: “Voting for Smith
is rational for Jones in the 1998 Congressional elections.”) Let us, for simplic-
ity, individuate actions so that acts of distinct agents count as distinct and, more
specifically, the identity of the agent is recoverable from the identity of the act.
This will allow us to work with the simpler, binary, concept “action $x$ is rational
in setting $z$.”) In order to show that the binary concept is circular, it suffices to show that there are settings $z$ in which the unary concept “action $x$ is rational (in $z$)”—symbolized as $\text{Rat}_z(x)$—is circular.

The settings we shall work with are games in normal form. These are settings in which there are finitely many agents, also known as players, each of whom has finitely many choices—choices that are mutually exclusive and exhaustive. Players choose simultaneously and independently of one another, and each player receives a reward that depends on choices made by all. Players have the resources and the motivation to arrive at their choices by reasoning out the consequences of their and others’ actions, and all relevant information about the game and the players is common knowledge.

We shall consider only strict games—games in which a player is never indifferent between her alternatives: for all possible combinations of choices of the other players, there is a unique best choice for the player.

The important information about games in normal form can be displayed in a tabular form, as in the following example.

**Example and notation.** The game $\Gamma_1$. $\Gamma_1$ is a two-player binary game in which the alternatives for one player, $A$, are those of doing an act (represented by $a$) and of not doing it (represented by $\overline{a}$). Similarly, the alternatives for the other player, $B$, are $b$ and $\overline{b}$.

\[
\begin{array}{c|cc}
 & b & \overline{b} \\
\hline
a & 2,2 & 0,1 \\
\overline{a} & 1,3 & 2,2 \\
\end{array}
\]
4.3 Concept of Rational Choice

Let us represent outcomes in a game by unordered pairs of actions. For example, \(\{a, b\}\)—abbreviated to \(\bar{a}b\)—is the outcome in which A does not opt for \(a\) but B opts for \(b\). Outcomes in a three-person game are represented by unordered triples of actions, one action for each player; and similarly for the other \(n\)-person games.

The payoffs in the four possible outcomes in \(\Gamma_1\) are indicated in the table; the first number gives the payoff for A and the second number the payoff for B. Let us use the expression 
\[u_1(x, y)\]
to denote the payoff, or utility value,\(^{11}\) for a player \(x\) in an outcome \(y\) relative to a game \(\Gamma\). Here and elsewhere, the subscript \(\Gamma\) may sometimes be dropped when the context allows it. In particular, the utility value in a game \(\Gamma_n\) will be indicated by the expression \(u_n(x, y)\). So in \(\Gamma_1\), we have 
\[u_1(A, \bar{a}b) = 1 \land u_1(B, \bar{a}b) = 3.\]
Observe that \(\Gamma_1\) is a strict game. Player A has a best play, namely \(\bar{a}\), for the alternative \(\bar{b}\) of B; and player B has a best play, namely, \(b\), for the alternative \(\bar{a}\) of player A; and similarly for the other two possibilities. The game would fail to be strict if, for example, \(u_1(B, \bar{ab})\) were identical to \(u_1(B, \bar{ab})\).

The utility values in a game need not indicate, as far as our purposes are concerned, anything more than the preferential orderings of the players. They need not provide any basis for comparisons of preferences across players (comparisons of the sort needed when considering fairness and justice). And they need not provide any basis for calculations of expected utility (calculations of the sort needed when probabilities need to be taken into account).
Any attempt to apply game theory to large-scale social institutions has to exercise great caution. The assumptions built into the theory (e.g., in games in normal form) are strong and may well not constitute a useful idealization for understanding most (perhaps all) interesting social phenomena. (I myself do not hold such a skeptical view, however.) The argument below rests, I want to stress, not on any large claim about the usefulness of game theory but on a very weak claim, namely, that strict games in normal form are possible. Agents can be in settings in which the assumptions underlying these games hold. It is immaterial to the argument whether these settings are natural and common, or contrived and rare.

4.3.2 The Definition. Consider a variant of \( \Gamma_1 \) in which A “plays” against Nature, not against a calculating and thinking thing. In order to decide his best course of action in this new game, A has to consider the probabilities of \( b \) and \( \overline{b} \), and has to calculate the expected utilities of his two acts. The act with the greater expected utility is his best course of action. If, given A’s beliefs, the probabilities of \( b \) and \( \overline{b} \) are equal, then it is rational for A to choose \( \overline{a} \) (with the expected utility of 1.5) over \( a \) (with the expected utility 1.0). This sort of decision-theoretic argument is reasonable when the other “player” is Nature, but not when the other player is a reasoning and thinking being. In a game such as \( \Gamma_1 \), A has to figure out not probabilities but the rational course of action for the other player B. And, because B’s situation is similar to A’s, what is rational for B depends, in turn, on what is rational for A. This sort of circularity is ever-present in game-theoretic settings and is absent from decision-theoretic ones. It is this circularity that makes game theory more complex than decision theory. Decision theory is trivial when probabilities can be left out of account; game theory is not.

The definition of rationality presented below—actually it is a collection of definitions, one for each game—is built on the following type of intuitive reasoning.
Player B argues thus in $\Gamma_1$: “Either act $a$ is rational or it is not rational. If the former, then, in view of the fact that
\[ u_1(B, ab) > u_1(B, a\overline{b}), \]
I gain more by choosing $b$ than by choosing $\overline{b}$, and so $b$ is the rational act for me. If the latter, then, in view of
\[ u_1(B, \overline{ab}) > u_1(B, \overline{a}b), \]
again $b$ is the rational act for me. In either case, $b$ is the rational act for me.” There is a transition in the first part of the argument from the premiss
\[ \text{Rat}_1(a) \land (u_1(B, ab) > u_1(B, a\overline{b})) \]
to the conclusion
\[ \text{Rat}_1(b). \]
And in the second part of the argument there is a parallel transition to the same conclusion from the premiss
\[ [\neg\text{Rat}_1(a) \land (u_1(B, \overline{ab}) > u_1(B, \overline{a}b))]. \]
The following definition of rationality is built on transitions such as these.

Let $\Gamma$ be a two-person binary strict game and let $\varphi_a$ and $\varphi_b$ abbreviate, respectively, the formulas (9) and (10).

\begin{align*}
(9) \quad & [\text{Rat}_\Gamma(b) \land (u_\Gamma(A, ab) > u_\Gamma(A, a\overline{b}))] \lor [\neg\text{Rat}_\Gamma(b) \land \\
& (u_\Gamma(A, a\overline{b}) > u_\Gamma(A, a\overline{b}))].
\end{align*}

\begin{align*}
(10) \quad & [\text{Rat}_\Gamma(a) \land (u_\Gamma(B, ab) > u_\Gamma(B, a\overline{b}))] \lor [\neg\text{Rat}_\Gamma(a) \land \\
& (u_\Gamma(B, a\overline{b}) > u_\Gamma(B, a\overline{b}))].
\end{align*}

Note that $\varphi_b$ captures, from $b$’s point of view, the two transitions present in the intuitive argument given above, and $\varphi_a$ captures two parallel transitions from $a$’s point of view. Now the definition of rationality in game $\Gamma$ is this:
\[ \text{Rat}_\Gamma(x) =_{df} (x = a \land \varphi_a) \lor (x = \overline{a} \land \neg \varphi_a) \lor (x = b \land \varphi_b) \lor \\
(x = \overline{b} \land \neg \varphi_b). \]
The definition says that if the condition $\varphi_a$ obtains then $a$ is rational, otherwise $\overline{a}$ is rational; and similarly for $\varphi_b$. The definition is easily generalized to apply uniformly to all two-person strict binary games. Definitions for $n$-person strict games (with possibly multiple choices for players) can be built in a parallel way. I give the definition for a three-person binary game, and for a two-person tertiary game, in the next footnote. All these definitions are built on a common intuitive basis and on a common pattern. Let us call them by a common name, ‘$D$’.

4.3.3. Rules of Revision The rules of revision generated by definitions $D$ have an attractive and useful visual representation. Observe, first, that each outcome corresponds to a hypothesis. For example, the outcome $ab\overline{c}$ corresponds to the hypothesis that the extension of the definiendum is $\{a, \overline{b}, \overline{c}\}$. Second, in view of the particular character of $D$, the action of the revision rule is significant only on these special hypotheses. The remaining hypotheses quickly drop out of the revision process. For instance, in a binary game, none is to be found after the initial stage of revision (because the definiens is invariably satisfied by exactly one act of each player).

Let us use the same notation for hypotheses and outcomes. This will allow us to speak of the hypothesis $ab\overline{c}$ as well as the outcome $ab\overline{c}$. We can now turn to the visual representation of revision rules.

Example continued. The game $\Gamma_1$. It can be verified that the definition of $\text{Rat}_1$ simplifies to (11) when the utility values are taken into account. (That is, in light of these values, the definiens of the original and the simplified definitions are equivalent. In particular, if one definiens can be derived in $C_0$ at an index, then so also can the other at the same index.)

\[
(11) \quad \text{Rat}_1(x) =_{\text{Df}} (x = a \land \text{Rat}_1(b)) \lor (x = \overline{a} \land \neg \text{Rat}_1(b)) \lor \\
(x = b \land (\text{Rat}_1(a) \lor \neg \text{Rat}_1(a))).
\]
The content of (11) is plainer if it is transformed into two partial definitions. (This is possible because, for each player, exactly one act satisfies the definiens.)

\[
Rat_1(a) =_{Df} Rat_1(b).
\]

\[
Rat_1(b) =_{Df} Rat_1(a) \lor \neg Rat_1(a).
\]

Now consider the hypothesis \(a \overline{b}\). Under this hypothesis, we have \(Rat_1(a)\) and \(\neg Rat_1(b)\), and so the definiens of the first partial definition is evaluated false and that of the second true. The partial definitions dictate that \(b\), but not \(a\), falls under \(Rat_1\). Hence the revision rule yields the output \(\overline{a}b\) for the input \(a\overline{b}\).

We can represent this action of the revision rule by an arrow—drawn in bold below—from the \(a\overline{b}\) box to the \(\overline{a}b\) box. The action on the other hypotheses is calculated in a similar manner and is indicated by broken arrows in the diagram. In later examples of games, I shall always give the partial definitions and the revision rule—the latter diagrammatically.

![Diagram](image)

Note that the reflexive hypotheses can be read visually off the diagrams: they lie on looped paths. There is only one reflexive hypothesis, \(a\overline{b}\), in the present example.
The action of the revision rule can intuitively be viewed thus. Given a hypothesis, we ask, from each agent’s point of view, what the best move would be for the agent on the assumption that the others act as indicated in the hypothesis. Strictness implies that there will always be a unique answer to our query. These answers, one for each player, constitute the revised hypothesis.

The ranges of the revision rules for our definitions $D$ are invariably finite. Hence the definitions meet the finiteness requirement of §4.1, and the semantics sketched there applies: the status of a sentence is fixed by the pattern of its truth values in the reflexive hypotheses. To review: A sentence is valid iff it is true in all reflexive hypotheses. It is categorical iff it has the same truth value in all reflexive hypotheses; it is pathological otherwise. A pathological sentence is quasi-categorical iff its truth value stays the same along every loop; it is paradoxical iff its truth-value fluctuates along every loop. Finally, a pathological sentence is quasi-paradoxical iff it is neither paradoxical nor quasi-categorical.

Let us call games in which there is exactly one reflexive hypothesis regular; the rest let us call irregular. In regular games (and only in regular games), the revision process converges to, and yields, a definite verdict about what is rational for each player to do. And this verdict is a Nash equilibrium for the game: it satisfies the condition that no player can improve her payoff by unilaterally playing differently. That is, if the others’ play conforms to a Nash equilibrium, one is best off conforming to the Nash equilibrium also. Note that a Nash equilibrium $h$ for a game $\Gamma$ is a fixed point of the revision rule $\delta_\Gamma$—that is, $h$ is a Nash equilibrium iff $\delta_\Gamma(h) = h$.

**Example concluded.** The game $\Gamma_1$. $\Gamma_1$ is regular and the revision process converges to the Nash equilibrium $ab$. The sentences $Rat_1(a)$ and $Rat_1(b)$ are, therefore, valid—a result that agrees with our sense of how A and B should play the game.
Definitions $D$ and their revision rules display very attractive behavior in some games. I point out some of this behavior in §§3.4 and 3.5. There are games (alas!) in which their behavior is not so attractive. I draw attention to this behavior in §3.6.

4.3.4 Iterated Dominance and Regular Games. A common and intuitive way of reasoning about a game is *argument by iterated dominance*. In this form of reasoning, one eliminates successively poor or *dominated* plays, where a play $c$ of a player $C$ is *dominated* iff $c$ is not the best play for $C$ in any possible combination of plays by the others. (Example: In $\Gamma_1$, the play $b$ is dominated.) One begins an iterated-dominance argument by eliminating one dominated play, say $c$, obtaining thereby a reduced game in which $c$ is no longer an option. The reduced game may, in turn, have some dominated acts, and one proceeds by eliminating one of these and thus reducing the game further. This kind of argument can be carried out sufficiently many times in some games to yield a solution for the game. The following relation holds between revision processes and arguments by iterated dominance.

**Fact.** If an iterated-dominance argument yields a solution in a strict game $\Gamma$, then the revision process converges to the same solution (and hence $\Gamma$ is bound to be regular).

**Outline of proof.** Let $\Gamma$ be a strict game and let $P$ be a set of plays that have been eliminated at a stage of the iterated-dominance argument. Observe that an act $c$ is dominated in the game that results when $\Gamma$ is reduced by $P$ iff, for all hypotheses $h$ such that $h \cap P = \emptyset$,

$$c \not\in \delta_{\Gamma}(h).$$

Hence, if an act $c$ is eliminated by the $n$th stage in an iterated-dominance argument, then $c$ will not belong to any hypothesis at the $n$th revision stage. This implies the desired conclusion.
The converse relationship does not hold, however. There are games in which iterated-dominance does not yield a solution, but the revision process does so.

**Example. The Game** $\Gamma_2$. $\Gamma_2$ is a strict game with three players, A, B, and C. The first table below gives the payoffs when C plays $c$, and the second when C plays $\overline{c}$. The three numbers in each outcome indicate the payoffs for A, B, and C respectively.

\[
\begin{array}{ccc}
 & b & \overline{b} \\
 a & 2, 2, 2 & 1, 1, 1 \\
 C \text{ plays } c & & \\
\overline{a} & 1, 1, 1 & 0, 0, 1 \\
 a & -1, -1, 0 & 0, 0, 0 \\
 C \text{ plays } \overline{c} & & \\
\overline{a} & 0, 0, 2 & 2, 2, 0
\end{array}
\]

\[
\text{Rat}_2(a) \equiv_{DF} \text{Rat}_2(c);
\]
\[
\text{Rat}_2(b) \equiv_{DF} \text{Rat}_2(c);
\]
\[
\text{Rat}_2(c) \equiv_{DF} \text{Rat}_2(a) \lor \neg \text{Rat}_2(b).
\]

Observe that dominance considerations do not eliminate any of the six acts in $\Gamma_2$, but the revision process converges to the intuitively correct solution $abc$. Furthermore, there is a simple and intuitive argument—one that can be formalized in $C_0$—that
leads to this solution. The formalized version of the argument is as follows. To establish (12),

\[(12) \ [\text{Rat}_2(a) \land \text{Rat}_2(b) \land \text{Rat}_2(c)],\]

it suffices to derive each of the conjuncts at index 0. Now it is a fact about $C_0$ that if an indexed formula $\varphi^i$ can be derived from hypotheses that do not contain any occurrences of the definiendum, then, for any index $j$, $\varphi^j$ can also be derived from the same premises. To establish (12), therefore, it suffices to derive each of its conjuncts at some index or the other. This can be done as follows. Begin by assuming, for reductio, that $\neg\text{Rat}_2(c)^0$. Facts about $\Gamma_2$ and Definiendum Introduction (DfI) yield

$\neg(\text{Rat}_2(a) \lor \text{Rat}_2(b))^{-1}$.

[For the supposition $(\text{Rat}_2(a) \lor \text{Rat}_2(b))^{-1}$ and an application of DfI to the third partial definition above yield a contradiction.] Rules of Propositional Logic now allow us to deduce $\neg\text{Rat}_2(a)^{-1}$ and $\text{Rat}_2(b)^{-1}$. The latter yields by an application of Definiendum Elimination (DfE) that $\text{Rat}_2(c)^{-2}$ [using the second partial definition] and this, in turn, yields $\text{Rat}_2(a)^{-1}$ by DfI [using the first partial definition]. We have deduced a contradiction from $\neg\text{Rat}_2(c)^0$. So $\text{Rat}_2(c)^0$ can be derived. Hence $\text{Rat}_2(a)^1$ and $\text{Rat}_2(b)^1$ are also derivable by DfI using the first two partial definitions. This establishes that (12) can be derived in $C_0$.

The use of more than two players in $\Gamma_2$ is essential:

**Fact.** In every two-person strict game that is regular, iterated dominance yields a Nash equilibrium.
Proof. Suppose, for reductio, that \( \Gamma \) is a two-person, strict, regular game in which an iterated-dominance argument does not yield a solution. Let the set of plays that cannot be eliminated for the players, say A and B, be \( P_A \) and \( P_B \). Define functions \( f \) and \( g \) on, respectively, \( P_A \) and \( P_B \) as follows:

\[
\begin{align*}
  f(a) &= \text{the best play for } B \text{ if } A \text{ plays } a, \\
  g(b) &= \text{the best play for } A \text{ if } B \text{ plays } b.
\end{align*}
\]

Note that \( f(a) \in P_B \), because \( f(a) \), being the best play when \( A \) plays \( a \), cannot be eliminated by a dominance argument. Note further that if \( b \in P_B \) then \( b \in \text{the range of } f \), for otherwise \( b \) would be a dominated act and would thereby be eliminable. It follows that \( P_B = \text{the range of } f \). Hence the cardinality of \( P_B \) is less than or equal to the cardinality of \( P_A \). A parallel argument establishes that the cardinality of \( P_A \) is less than or equal to the cardinality of \( P_B \). Hence \( P_A \) and \( P_B \) must be equinumerous. It follows that the functions \( f \) and \( g \) are one-one.

As \( \Gamma \) is regular, there is a unique reflexive hypothesis, say \( a'b' \), for it. Define a set of hypotheses \( H \) thus:

\[
ab \in H \iff [a \neq a' \land b \neq b' \land a \in P_A \land b \in P_B].
\]

Because the iterated dominance argument does not yield the solution \( a'b' \) for \( \Gamma \), there must be plays \( a \in P_A \) and \( b \in P_B \) such that \( a \neq a' \) and \( b \neq b' \). That is, \( H \) must be nonempty. Now consider the action of the revision rule \( \delta \) of \( \Gamma \) on an arbitrary member \( ab \) of \( H \). Observe that

\[
\delta(ab) = g(b)f(a).
\]

As \( f \) and \( g \) are one-one and \( g(b') = a' \) and \( f(a') = b' \), we can conclude that
That is, \( \delta(ab) \in H \). So the image of \( H \) under \( \delta \) is a subset of \( H \). As \( H \) is finite, some members of \( H \) must be reflexive. This contradicts the regularity of \( \Gamma \).

### 4.3.5 Extensional Adequacy and Essential Circularity

The type of intuitive reasoning that works so well in \( \Gamma_1 \) and \( \Gamma_2 \) can sometimes lead to seemingly strange conclusions, as the next example shows.

**Example. The Game \( \Gamma_3 \).** \( \Gamma_3 \) is a strict game with payoffs as indicated.

\[
\begin{array}{cc|cc}
 & b & \bar{b} \\
\hline
a & 2, 3 & 3, 2 \\
\bar{a} & 3, 2 & 2, 3 \\
\end{array}
\]

\[
\text{Rat}_3(a) =_{DF} \neg \text{Rat}_3(b); \text{Rat}_3(b) =_{DF} \text{Rat}_3(a).
\]

One can argue as follows in \( \Gamma_3 \). Suppose that \( b \) is rational. Then, since \( A \) receives a greater benefit in the outcome \( \bar{a}b \) than in the outcome \( ab \), \( a \) is not a rational act for \( A \). But if \( a \) is not rational then, since \( B \) receives a greater benefit in the outcome \( a\bar{b} \) than in the outcome \( \bar{a}b \), \( b \) is not a rational act for \( B \). So from the premiss that \( b \) is rational, we can deduce the conclusion that \( b \) is not rational. A parallel argument takes us the other way, from the premiss that \( b \) is not rational to the conclusion that \( b \) is rational. We seem to have arrived at a contradiction.
This argument, minus the contradiction, can be mimicked in $C_0$. We begin with the supposition $Rat_3(b)^0$. The rules for definitions allow us to conclude $\neg Rat_3(a)^1$. [The hypothesis $Rat_3(a)^1$ and an application of Definiendum Elimination pave the way to a contradiction.] And $\neg Rat_3(a)^1$ similarly leads to $\neg Rat_3(b)^2$. So from $Rat_3(b)^0$ we can derive $\neg Rat_3(b)^2$. A parallel argument takes the other way, from $\neg Rat_3(b)^2$ to $Rat_3(b)^0$.

We can thus go back and forth between $Rat_3(b)$ and its negation—as in the intuitive reasoning—but only when the two formulas have different indices. From this no contradiction can be derived. Calculus $C_0$ enables us to harness the power of many intuitive arguments while keeping at bay their seemingly contradictory conclusions.

In $\Gamma_3$, there is just one looped path and the truth values of the sentences $Rat_3(a)$ and $Rat_3(b)$ fluctuate along it. Hence the semantics given above deems the sentences to be paradoxical. And this seems to accord with our intuitive sense of rationality in $\Gamma_3$: judgments about the rationality of $a$ and of $b$ are all unstable. This intuition can perhaps be made more vivid if we modify the game a little. Let us have the players make their moves (not secretly but out in the open), and let us allow them to change their moves as many times as they wish. Once no player wishes to change his or her move, the game ends and the players receive the payoffs for the selected outcome. Notice that in $\Gamma_3$ the players will keep changing their moves; they will not settle on an outcome. At each stage, one player will have good reason for preferring a different move. If at one stage the proposed play is, for example, $ab$, then A will have reason to change his move to
π. But with this change, B has reason to change her move to \( b \). And this change, in turn, gives A a reason to change his move back to \( a \); and so on. This kind of flip-flopping is reminiscent of the Liar sentence under semantic evaluation,\(^{17}\) and it confirms the classification of \( \text{Rat}_3(a) \) and \( \text{Rat}_3(b) \) as paradoxical. Definitions \( D \) are plainly essentially circular in \( \Gamma_3 \), and they appear to be extensionally adequate.\(^{18}\)

Let us call a game in which there is no Nash equilibrium \textit{unstable}. An unstable game, then, is one in which the revision rule has no fixed points; \( \Gamma_3 \) being an example. Let us call a game \( \Gamma \) \textit{quasi-regular} iff \( \Gamma \) is not regular and every reflexive hypothesis in \( \Gamma \) is a Nash equilibrium. A quasi-regular game is like a regular game in that repeated applications of the revision rule invariably end up in a Nash equilibrium.\(^{19}\) It is unlike a regular game, however, in that there is no convergence to one equilibrium point; revisions starting at different initial hypotheses can end up in different equilibria. Here is an example.

\textbf{Example. The Game} \( \Gamma_4 \). The utility values in \( \Gamma_4 \) are as follows:
On Circular Concepts

\[ \text{Rat}_4(a) = \text{Df} \ \text{Rat}_4(b) \land \text{Rat}_4(c); \]
\[ \text{Rat}_4(b) = \text{Df} \ \text{Rat}_4(a) \land \text{Rat}_4(c); \]
\[ \text{Rat}_4(c) = \text{Df} \ \text{Rat}_4(a) \land \text{Rat}_4(b). \]

The two Nash equilibria, \( abc \) and \( \overline{abc} \), are the only reflexive hypotheses. \( \Gamma_4 \) is, therefore, quasi-regular. Note that all three statements \( \text{Rat}_4(a) \), \( \text{Rat}_4(b) \), and \( \text{Rat}_4(c) \) are assessed as pathological and, in particular, as quasi-categorical.

Intuitive reflection, too, does not provide guidance on how players should play \( \Gamma_4 \). Indeed, claims of rationality in \( \Gamma_4 \) seem intuitively to display pathologicality of the sort found in the Truth-Teller ("This very statement is true"). If we suppose that \( a \) is rational, we are led to conclude that \( b \) and \( c \) are rational, and this confirms that \( a \) is rational. Similarly, if we suppose that
4.3 Concept of Rational Choice

If a is not rational, we are led to conclude that neither b nor c is rational, and this confirms that a is *not* rational. Another way of bringing out the similarity with the Truth-Teller is through a variant game in which players are allowed to change their moves (as in the variant of \( \Gamma_3 \) above). Note that irrespective of how the players choose initially, they will end up at an equilibrium point. But different starting points may result in different equilibria.

Conclusion: The definition of rationality in \( \Gamma_4 \) is essentially circular and appears to be extensionally adequate.\(^{20}\)

It is a recurring idea in game theory that the “solution” for a game—that is, a specification of the best play for each player—has to be the unique Nash equilibrium for the game. And our definitions are in accord with this idea. But the converse idea—that every unique Nash equilibrium is a solution—is false. And, fortunately, our definitions do not imply it: Irregular games with unique Nash equilibrium points exist. Here is an example.

**Example. The Game \( \Gamma_5 \).** \( \Gamma_5 \) is a strict game in which each player has three choices.

\[
\begin{array}{ccc}
  & b_1 & b_2 & \bar{b} \\
 a_1 & 1, 1 & 0, 0 & 0, 0 \\
 a_2 & 0, 0 & 2, 3 & 3, 2 \\
\bar{a} & 0, 0 & 3, 2 & 2, 3
\end{array}
\]

\[
\begin{align*}
\text{Rat}_5(a_1) & = \text{Df} \, \text{Rat}_5(b_1); \quad \text{Rat}_5(a_2) = \text{Df} \, \neg \text{Rat}_5(b_1) \land \neg \text{Rat}_5(b_2); \\
\text{Rat}_5(b_1) & = \text{Df} \, \text{Rat}_5(a_1); \quad \text{Rat}_5(b_2) = \text{Df} \, \text{Rat}_5(a_2).^{21}\end{align*}
\]
There is a unique Nash equilibrium \( a_1b_1 \) in \( \Gamma_5 \), but the players are better off avoiding the outcome it represents. So \( a_1b_1 \) is not intuitively the correct solution for the game. Note that the revision process does not converge to the Nash equilibrium and does not provide a specific recommendation for the players. This seems to be in accord with our intuitive sense of the game.\(^{22}\)

It is important to be clear on the meaning of Nash equilibrium. To affirm of an outcome, say \( ab \), that it is a Nash equilibrium is to affirm the rationality of the acts \( a \) and \( b \) conditionally, not absolutely. It is to affirm that \( a \) is rational for A if \( b \) is rational for B, and further that \( b \) is rational for B if \( a \) is rational for A. Nothing follows from this about the absolute rationality of \( a \) and \( b \). Even if \( ab \) is a unique Nash equilibrium, it does not follow that \( a \) and \( b \) are rational absolutely, for the players might be better off doing something different from \( a \) and \( b \). The idea that a solution has to be a unique Nash equilibrium should be sharply separated from the idea that every unique Nash equilibrium is a solution. The first idea is plausible; the second is not.\(^{23}\)

There is another reason for not identifying the concept of solution with that of Nash equilibrium. There are unstable games (i.e., games without Nash equilibria) in which definite recommendations can be made about how some of the players should play. A good theory of rationality should yield proper recommendations in such cases. Here is a particularly simple example that illustrates the point.

**Example. The Game \( \Gamma_6 \).** \( \Gamma_6 \) is a binary game with three players, A, B, and C. The payoffs for A and B are independent of how C plays, and are as stated in \( \Gamma_3 \). The payoffs for C are also independent of how A and B play. Let us say that the payoff is 2 for play \( c \) and 0 for play \( \bar{c} \). There are no Nash equilibria in this game, but plainly \( \text{Rat}_6(c) \) is true. Our definition deems \( \text{Rat}_6(c) \)
to be valid and $\text{Rat}_6(a)$ and $\text{Rat}_6(b)$ to be paradoxical. Again the
deinition is essentially circular and appears to be extensionally
adequate.

This is all the evidence that I wish to present for now that the concept of rational
choice is circular. Let me stress that I have not claimed—and for the purposes
of establishing the desired conclusion I do not need to claim—that definitions $D$
capture the meaning, or part of the meaning, of ‘rational choice’, nor that they fix
the intension of this expression, nor even that they are extensionally adequate in
all games. I have claimed only that the definitions are both extensionally adequate
and essentially circular in some games in normal form. This, if true, is sufficient to
establish the circularity of the concept of rational choice.

4.3.6 Limitations of the Definition. I turn now to some games in which definitions
$D$ are not, in my opinion, extensionally adequate.

Example. The Game $\Gamma_7$.

\[
\begin{array}{cc|c}
\text{a} & \text{b} & \neg \text{b} \\
\hline
1, 1 & 3, 0 \\
0, 3 & 2, 2 \\
\end{array}
\]

$\text{Rat}_7(a) = \text{Df } \text{Rat}_7(b) \lor \neg \text{Rat}_7(b)$;

$\text{Rat}_7(b) = \text{Df } \text{Rat}_7(a) \lor \neg \text{Rat}_7(a)$;

This game is a version of the Prisoner’s Dilemma. Acts $\text{a}$ and $\neg \text{b}$
(representing non-confession) are both dominated. Hence there
is only one Nash equilibrium, \( ab \) (representing confession for both prisoners), and the revision process converges to it.

There is debate in the philosophical literature on whether the Nash equilibrium is the correct solution in the Prisoner’s Dilemma (and other similar examples). André Chapuis has strongly urged to me the viewpoint that it is not, and I am inclined to agree with him. As Chapuis observes, there is something odd about affirming the rationality of \( a \) and \( b \) when it is plain—and it is plain to the players—that the outcome \( \overline{ab} \) is more profitable for both players.

Extensional inadequacy of \( D \) in \( \Gamma_7 \) is perhaps debatable. Less debatable examples do exist, and the following fact points to one.

**Fact.** There are no two-person strict games that are quasi-regular. (Three-person quasi-regular strict games do exist, as \( \Gamma_4 \) shows.)

**Proof.** Suppose, for reductio, that there is a two-person strict game \( \Gamma \)—with the revision rule \( \delta \) and players A and B—that is quasi-regular. There must, then, be at least two Nash equilibria for \( \Gamma \), say \( xy \) and \( uv \). Because \( \Gamma \) is strict, \( x \neq u \) and \( y \neq v \). Because \( uv \) is a Nash equilibrium, \( u \) is A’s best play if B plays \( v \). Similarly, because \( xy \) is a Nash equilibrium, \( y \) is B’s best play if A plays \( x \). It follows that

\[
\delta(xv) = uy.
\]

A similar argument shows that \( \delta(uy) = xv \). Hence \( xv \) is a reflexive hypothesis that is not a Nash equilibrium. The reductio is complete.

Now what about this game?

**Example. The Game \( \Gamma_8 \).**
4.3 Concept of Rational Choice

There are two Nash equilibria here, \(ab\) and \(\overline{ab}\), but these are not the only reflexive hypotheses; \(\overline{ab}\) and \(\overline{a\overline{b}}\) are reflexive also. Consequently, according to \(D\), \(\Gamma_8\) is not quasi-regular, and the sentences \(\text{Rat}_8(a)\) and \(\text{Rat}_8(b)\) are quasi-paradoxical. But what \(D\) yields here is not, I think, in agreement with intuition. Intuitively, we do not count \(\overline{ab}\) and \(\overline{a\overline{b}}\) as options at all, and hence we do not perceive any instability in \(\text{Rat}_8(a)\) and \(\text{Rat}_8(b)\). These sentences seem intuitively to behave like the Truth-Teller.

Here is another example where extensional inadequacy of \(D\) is beyond debate.

**Example. The Game \(\Gamma_9\).**

\[
\begin{array}{c|c}
\text{a} & \text{b} \\
\hline
2, 2 & 0, 0 \\
\hline
\overline{a} & \overline{b} \\
\hline
0, 0 & 2, 2 \\
\end{array}
\]

\(\text{Rat}_9(a) = \text{Df} \text{Rat}_9(b); \text{Rat}_9(b) = \text{Df} \text{Rat}_9(a)\).
The behavior of the revision rule is the same in $\Gamma_8$ and $\Gamma_9$. Hence the definitions yield the same assessments for the two games: $\text{Rat}_9(a)$ and $\text{Rat}_9(b)$ are both deemed quasi-paradoxical. Intuitively, however, it is plain that in $\Gamma_9$ player A should play $a$ and B should play $b$. The sentences $\text{Rat}_8(a)$ and $\text{Rat}_8(b)$ ought therefore to be assessed valid.

An outcome in a game $\Gamma$ is said to be Pareto optimal iff no outcome in $\Gamma$ improves the payoffs for one or more players without at the same time reducing the payoffs for some others. (Examples: In $\Gamma_3$, all outcomes are Pareto optimal; in $\Gamma_7$, the Prisoner’s Dilemma, only the outcome $ab$ fails to be Pareto optimal; and in $\Gamma_9$, only $ab$ is Pareto optimal.) I conjecture that definitions $D$ are extensionally adequate in all games in which only Pareto-optimal hypotheses are reflexive. In particular, they are adequate in games in which all hypotheses are Pareto-optimal (e.g., $\Gamma_3$). Note that in all the examples above in which $D$’s extensional adequacy is in doubt, the problem is invariably due to the presence of non-optimal points among the reflexive hypotheses.

I hope it is plain that the limitations of definitions $D$ do not cast any doubt on the soundness of the argument of §4.3.5 for the circularity of the concept of rational choice. I hope it is plain also that definitions $D$ are worthy of study. I believe they provide a good beginning and a good foundation for the search for a fully adequate theory of rational choice.

### 4.4. The Concept of Belief

The logical character of the concept of belief, unlike that of rational choice, is elusive. A brief discussion of it will be useful, nonetheless, for it will highlight some points about the method presented in §2 and will lead to a refinement of it.

Byeong Deok Lee has argued that the concept of rational belief is circular. Lee has shown that the traditional, self-referential, paradoxes about belief arise
only under certain strong assumptions about rationality, and under these assumptions, the concept of rational belief is circular.\textsuperscript{25}

But what about the concept of belief \textit{simpliciter}? Is it circular? There are in the philosophical literature general views about belief that, if true, could help provide an answer. Thus, there is the view that beliefs are sentences in the head. That is, to believe that $p$ is to have inscribed in one’s head—more specifically in one’s “belief box”—a sentence that has the same content as $p$ (or one whose content places it in the logical neighborhood of $p$). If this view were correct, then there would be reason (not necessarily conclusive) to think that belief is noncircular. For whether certain tokens are inscribed in a certain place depends simply on brute facts and is not subject to the pathologicality that is a mark of circular concepts. (Compare ‘this very sentence is not inscribed on this page’, which is plainly false, with ‘this very sentence is not true’, which is pathological.) A competing and popular view, \textit{functionalism}, characterizes beliefs and other mental states by their role in the production of behavior. A mental state does not produce behavior in isolation, however, but only in tandem with other mental states. Hence a functionalist characterization of one mental state has to refer to other mental states. And the characterization of these latter states can, in turn, refer to the former state. In short, functionalism lends some support to the idea that belief is circular.\textsuperscript{26}

I will not enter into the debate between these two philosophical conceptions of belief. I want to stress instead a methodological point from \S 4.2: it is not necessary to settle on a full analysis of a concept, or on an account of its nature, in order to resolve the logical issue of circularity. In fact, when one pursues the logical issue, it is often \textit{not} a good strategy to seek a definition that captures the nature, or the intension, of a concept. A better strategy is to seek extensionally adequate definitions; these are often sufficient to settle the issue of circularity. The philosophical conceptions of belief mentioned above, though important, are far too dubious to serve as a basis for an argument for or against the circularity of belief.
Belief is a relational notion. It relates a person $x$ and a moment of time $m$ to a proposition $z$. Let us fix on a particular moment of time—say, noon, January 1, 1998—and see if the resulting binary notion, $Bel(x, z)$, is circular. To show that it is circular, it suffices to establish that the unary notion $Bel_x(z)$ is circular for at least one $x$.

There is a quick and easy argument for the circularity of belief. Consider the unary notion “God believes $z$” ($Bel_{\text{God}}(z)$). This notion is extensionally equivalent to truth. So, the definition

$$Bel_{\text{God}}(\text{that } p) =_{\text{Df}} p$$

is extensionally adequate. The definition is also essentially circular; hence, the argument concludes, belief is circular.

This argument may succeed in establishing the circularity of “God believes $z$,” but it does not succeed in establishing the circularity of “believes.” The concept “true sentence” is circular but that does not imply that “sentence” is circular. Similarly, “God believes $z$” may be circular but that does not imply that “believes” is circular. The circularity of “God believes $z$” may issue from a circularity in the concept of God, not from a circularity in belief.

There is another, still easier, argument for circularity that deserves consideration. Fix on a particular person, say Dan, demonstratively picked out (so that there is no threat of circularity in the concept “Dan”). Now, the argument proceeds, the concept $Bel_{\text{Dan}}(z)$ is extensionally equivalent to itself. Hence

$$(13) \quad Bel_{\text{Dan}}(z) =_{\text{Df}} Bel_{\text{Dan}}(z)$$

is an extensionally adequate definition of “Dan believes $z$.” But (13) is essentially circular and, therefore, belief is a circular concept.

This argument is plainly far too easy. If it were sound, one could quickly prove the circularity of any concept whatsoever. The error in the argument is instructive, however, and points to an important distinction. The error lies in the transition to
the claim that (13) is extensionally adequate. It is, of course, true that \( \text{Bel}_{\text{Dan}}(z) \) is extensionally equivalent to itself, but it is not true that (13) is extensionally adequate. The concept (13) defines is everywhere pathological—in exactly the way that the concept defined by (14) is everywhere pathological.

(14) \( J(x) =_{\text{Df}} J(x) \).

But ‘Dan believes \( z \)’ is not everywhere pathological—it is not pathological on the proposition that snow is white, for instance. So the fact that a formula \( \varphi \) (in the present example, \( \varphi \) is \( \text{Bel}_{\text{Dan}}(z) \)) is extensionally equivalent to the definiendum (in the present example, \( \text{Bel}_{\text{Dan}}(z) \)) does not imply that \( \varphi \) yields an extensionally adequate definition. In judging extensional equivalence, we take the interpretation of the definiendum as given. In judging extensional adequacy, however, we determine the interpretation solely through the definition, and we then compare it with the actual interpretation. The distinction between extensional equivalence and extensional adequacy has no parallel in the domain of noncircular definitions. It is a distinction that arises only with circular definitions, and it is a distinction that is important to mark.

The fact is that the task of constructing extensionally adequate definitions (ones in which the interpretation of the definiendum is not presupposed) for concepts such as belief is exceedingly difficult. The method of §4.2 for establishing circularity, for all its seeming liberality, is not liberal enough. Fortunately, the demands imposed by the method can be weakened; one can get by with less than full extensional adequacy. Let us say that two predicates \( F \) and \( G \) are \textit{extensionally equivalent in a possible situation \( s \) over a set of objects \( X \)} iff, in \( s \), the significations of \( F \) and \( G \) when restricted to \( X \) are the same. That is, in \( s \), \( F \) and \( G \) agree on the objects in \( X \): if \( F \) is true (false, paradoxical, etc.) in \( s \) of a member of \( X \) then so is \( G \), and conversely. Further, let us say that a definition \( D \) is \textit{extensionally adequate over \( X \) for a predicate \( F \) in a situation \( s \)} iff, in \( s \), \( F \) and the definiendum of \( D \) are extensionally equivalent over \( X \). Then:
Method #5. To establish the circularity of a predicate $G$, it suffices to provide a definition $D$, a possible situation $s$, and a set of objects $X$ in $s$ such that, in $s$: (i) $D$ is essentially circular; (ii) the definiendum of $D$ is false of all objects outside of $X$; and (iii) $D$ is extensionally adequate over $X$ for $G$.

This method may appear too liberal. It allows the use of definitions that are adequate only over a very narrow range. I myself think that this liberality is all right. There is a distinction between the soundness of an argument and its persuasiveness. An argument for circularity that uses a definition with narrow extensional adequacy may well be sound but, in the absence of other considerations, will not be persuasive. The dialectical power of an argument for circularity increases with the width of its extensional adequacy and with the principles the definition reveals. Methodologically, however, at the beginning of a logical inquiry, when the primary concern is discovery rather than persuasion, it is all right to work with narrowly adequate definitions.

There are features of belief that make it tempting to think that it is circular: Belief attribution is not atomistic. It is impossible to attribute beliefs one by one; only systems of beliefs can be attributed. Further, some beliefs (e.g., Dan’s belief that he believes that snow is white) are grounded in other beliefs. And, finally, there are paradoxes (e.g., those due to John Buridan) in which belief is an essential constituent. Nevertheless, despite the liberality of Method #5, it is not easy to construct a satisfactory argument for the circularity of belief. It is not easy to show that belief ever displays the pathological behavior that is distinctive of circular concepts. Consider, for instance, the following two propositions, respectively, $p$ and $q$,

that Dan believes this very proposition at noon on 1 January 1998,
and

that Dan does not believe this very proposition at noon on 1 January 1998.

Note that the concept “Dan believes” need not be pathological over these two propositions. If Dan has never given a thought to such propositions, it is simply true (and not pathological) that he does not believe them—and so $p$ is false and $q$ is true. It even seems possible that Dan has considered these propositions and believes them (at noon on 1 January 1998). If he does believe them, then $p$ is true and $q$ is false. And the fact that $q$ is false (and even plainly false) is no reason to deny the possibility that Dan believes $q$. We sometimes believe things that are false—sometimes even things that are plainly false. Finally, if Dan reflects on the proposition $q$ around noon on 1 January 1998, he would go through a Liar-like flip-flop. But although this might be evidence for the pathologicality of rational belief, it is not evidence for the pathologicality of belief simpliciter. The flip-flop, it can be maintained, is simply an instance of changing beliefs.

I am not suggesting that belief does not exhibit pathologicality. Perhaps it does; perhaps it exhibits pathologicality of a subtle sort that has its roots in its holistic character. I am suggesting that the pathological behavior of belief, if it exists, is not easy to establish. The logical character of belief, as I remarked earlier, is elusive.

Notes

1The theory of definitions extends naturally to systems of mutually interdependent definitions (see Revision Theory, ch. 5) and to definitions of terms in other logical categories (e.g., many-place predicates, names, and function symbols). For the purposes of this chapter, it is easiest and also sufficient to treat only circular definitions of one-place predicates.
Note the scope of the existential quantifier. The number \( n \) may vary with interpretation \( M \). But for each \( M \) there should be a finite bound by which the revision process invariably (i.e., “for all \( h \)”) yields reflexive hypotheses.

Not all definitions that meet the finiteness requirement are of this sort, however. For a study of finite definitions, see Maricarmen Martínez, “Some Closure Properties of Finite Descriptions”; see also my “Finite Circular Definitions.”

I shall often suppress relativity to definitions.

It is tempting to extend the semantics beyond the finiteness requirement. But caution is in order. The semantics cannot be applied to all definitions because that will sometimes validate contradictions. This problem can be overcome by applying the semantics only to those definitions for which reflexive hypotheses exist under all interpretations. But now the conservativeness of definitions is lost. (See §5.2 below.) It may be suggested that the semantics be applied to those definitions \( D \) that meet the following weaker version of the finiteness requirement.

\[(*) \text{ For all } M \text{ and for all hypotheses } h, \text{ there is a number } n \text{ such that } \delta_{D,M}^n(h) \text{ is reflexive.}\]

Now the problem of conservativeness is overcome. Moreover, equivalence to \( S^\# \) remains intact, and \( C_0 \) remains sound. But do we really gain anything by this new proposal? Are there definitions that meet the weaker version of the finiteness requirement but not the original, stronger version? I leave this question open.

A plausible extension is obtained if the semantics is applied to definitions \( D \) that meet the following condition:

\[(** \text{ For all } M \text{ and for all recurring hypotheses } h, \text{ there is a number } n \text{ such that } \delta_{D,M}^n(h) \text{ is reflexive.}\]

Recurring hypotheses, intuitively, are the survivors of the transfinite revision process (see Revision Theory, Definition 5C.12). Now we have a significant enlargement of the domain of the semantics. And, again, equivalence to \( S^\# \) remains intact,
4.4 Notes

$C_0$ remains sound, and the problem of conservativeness is overcome. But we lose the completeness of $C_0$.

Signification captures the complete extensional behavior of a predicate. For example, in a three-valued language, signification captures not only the extension of a predicate but also its antiextension. For circular predicates, signification captures the ordinary, as well as the pathological, behavior. (See *Revision Theory*, pp. 30–31.)

More precisely, the method can be used to show the circularity of the *logical* notion of truth in certain *idealized* languages. The qualifications are needed because the Tarski biconditionals are necessarily true only for the logical notion and only for certain idealized languages. (See *Revision Theory*, ch. 1, especially pp. 20–22.) To establish more general theses about the circularity of truth, methods such as those discussed below are needed.

See ch. 1.

An important clarification of this claim will be found in §4.4. See the distinction between extensional *adequacy* and extensional *equivalence* drawn there.

The notion of noncircularity used here is a syntactic one. A definition is *non-circular* iff the definiens contains no occurrences of the definiendum.

I will identify payoff with utility value. The distinction between the two is important in general, but not for the particular issue at hand.

If $\Gamma$ is a three-person binary game in which the choices of the players A, B, and C are, respectively, $a$ and $\overline{a}$, $b$ and $\overline{b}$, and $c$ and $\overline{c}$, then the definition of rationality is as follows:

$$\text{Rat}_\Gamma(x) =_{\text{Df}} (x = a \land \psi_a) \lor (x = \overline{a} \land \neg \psi_a) \lor$$
$$\quad (x = b \land \psi_b) \lor (x = \overline{b} \land \neg \psi_b) \lor$$
$$\quad (x = c \land \psi_c) \lor (x = \overline{c} \land \neg \psi_c),$$

where $\psi_c$ abbreviates the formula,
[\text{Rat}_\Gamma(a) \land \text{Rat}_\Gamma(b) \land (u_\Gamma(C, abc) > u_\Gamma(C, ab\overline{c}])] \lor
\[\text{Rat}_\Gamma(a) \land \neg \text{Rat}_\Gamma(b) \land (u_\Gamma(C, \overline{abc}) > u_\Gamma(C, \overline{abc}))]\lor
\[\neg \text{Rat}_\Gamma(a) \land \text{Rat}_\Gamma(b) \land (u_\Gamma(C, \overline{abc}) > u_\Gamma(C, \overline{abc}))\]

and \psi_a and \psi_b are spelled out in a parallel way.

If \Gamma is a two-person ternary game in which the choices of A and B are, respectively, \(a_1, a_2\) and \(\overline{a}, \text{ and } b_1, b_2, \text{ and } \overline{b},\) then the definition of rationality is as follows. (Note that \(\overline{a}\) represents that A does neither \(a_1\) nor \(a_2\), and similarly for \(\overline{b}\).)

\[
\text{Rat}_\Gamma(x) =_{\text{Def}} \text{Cons} \land [(x = a_1 \land \chi_1) \lor (x = a_2 \land \chi_2) \lor (x = \overline{a} \land 
\neg \chi_1 \land \neg \chi_2) \lor (x = b_1 \land \theta_1) \lor (x = b_2 \land 
\theta_2) \lor (x = \overline{b} \land \neg \theta_1 \land \neg \theta_2)],
\]

where \text{Cons} stands for the formula,

\[
\neg(\text{Rat}_\Gamma(a_1) \land \text{Rat}_\Gamma(a_2)) \land \neg(\text{Rat}_\Gamma(a_2) \land \text{Rat}_\Gamma(\overline{a})) \land
\neg(\text{Rat}_\Gamma(a_1) \land \text{Rat}_\Gamma(\overline{a})) \land
\neg(\text{Rat}_\Gamma(b_1) \land \text{Rat}_\Gamma(b_2)) \land \neg(\text{Rat}_\Gamma(b_2) \land \text{Rat}_\Gamma(\overline{b})) \land
\neg(\text{Rat}_\Gamma(b_1) \land \text{Rat}_\Gamma(\overline{b}))],
\]

and \(\theta_2\) for the formula,

\[
[\text{Rat}_\Gamma(a_1) \land (u_\Gamma(B, a_1b_2) > u_\Gamma(B, a_1b_1)) \land (u_\Gamma(B, a_1b_2) > u_\Gamma(B, a_2\overline{b}))]
\lor
[\text{Rat}_\Gamma(a_2) \land (u_\Gamma(B, a_2b_2) > u_\Gamma(B, a_2b_1)) \land (u_\Gamma(B, a_2b_2) > u_\Gamma(B, a_2\overline{b}))]
\lor
[\neg \text{Rat}_\Gamma(a_1) \land \neg \text{Rat}_\Gamma(a_2) \land (u_\Gamma(B, \overline{a}b_2) > u_\Gamma(B, \overline{a}b_1)) \land (u_\Gamma(B, \overline{a}b_2) > u_\Gamma(B, \overline{a}b_2))]
\]

and with parallel expansions for the formulas \(\theta_1, \chi_1, \text{ and } \chi_2\).

\(^{13}\text{In games in which players have multiple choices, the null hypothesis can also occur at the next stage but not beyond it.}\)
Lingering doubts about intuitive correctness can be removed by making the payoffs in the outcome $abc$ very large. This change does not alter the behavior of the revision rule or the feasibility of an iterated-dominance argument.

15 Cristina Bicchieri writes in her valuable book *Rationality and Coordination*,

Rationality alone determines a player’s action only when she has a dominant strategy; common knowledge of rationality does the trick only when successive elimination of dominated strategies eliminates all but one strategy for each player. (p. 65)

$\Gamma_2$ is a counterexample to Bicchieri’s claim.

16 It is a derived rule of $C_0$ that a contradiction at an index implies a contradiction at any other index.

17 The similarity has been noted by Brian Skyrms, Robert Koons, and Chapuis. See Skyrms’s “Truth dynamics” and his book *Dynamics of Rational Deliberation*, Koons’s *Paradoxes of Belief and Strategic Rationality*, and Chapuis’s “Rationality and Circularity.”

18 By saying that a definition is extensionally adequate in a game $\Gamma$, I mean that it is extensionally adequate in a possible situation in which $\Gamma$ is played; similarly for the claim that a definition is essentially circular in $\Gamma$. Below, I retain the briefer expressions used in the text.

19 Consequently, in a quasi-regular game (as in a regular one) definitional equivalences imply material equivalences, and one can dispense with revision indices when reasoning in such games.

20 There are many attempts in game theory to find principles for selecting one equilibrium point over others as the optimal one. For example, there is an important proposal of R. Selten in which one considers the effects of slight mistakes by players (“the trembling hand”). Equilibria in which such mistakes have the least cost are deemed preferable to the others. This kind of consideration does serve to isolate a unique equilibrium point in some quasi-regular games. There are other
attempts (e.g., those due to Myerson, and to Kohlberg and Mertens) that can go further in this direction (see Skyrms’s *Dynamics of Rational Deliberation* and Bicchieri’s *Rationality and Coordination*). But these attempts do not affect the basic point of the present example: the existence of Truth-Teller-like pathologicality in games. First, the existence of symmetrical equilibria (see $\Gamma_8$ below) implies that elimination of multiple equilibrium points can be achieved only through ad hoc maneuvers. Second, the relevance of factors such as Selten’s to idealized games of the sort considered here is doubtful. Rational drunks may have an easier time reaching a decision in some games, such as $\Gamma_4$, but why should that provide comfort to the sober?

21 These partial definitions can be used only for hypotheses that satisfy the consistency condition. See note 12.

22 I do not think, however, that $D$ is extensionally adequate in $\Gamma_5$. For, according to $D$, the outcome $a_1 b_1$ is reflexive, and hence $\neg Rat_5(a_1)$ is not valid. But players A and B are both better off if they avoid the options $a_1$ and $b_1$. So it seems to me that A ought not to do $a_1$, that $\neg Rat_5(a_1)$ ought to be valid. Nevertheless, I think that $D$ is superior to the theory that invariably designates the unique Nash equilibrium as the solution. $D$’s consequence that $\Gamma_5$ is irregular seems to me correct.

23 This sort of phenomenon is also found for the concept of truth. There are examples in which the revision rule for truth (“the Tarski jump”) has a unique fixed point, but it is not the correct interpretation of the truth predicate.

24 See Lee’s “Burge on Epistemic Paradox,” “Paradox of Belief Instability and a Revision Theory of Belief,” and “The Knower Paradox Revisited.”

25 Nicholas Asher and Hans Kamp have studied the behavior of the revision processes for knowledge and belief when certain strong conditions of rationality hold. See their “Self-Reference, Attitudes, and Paradox.”
26Functionalists, it is true, have tried to remove the circularity from their characterizations of mental states. Lee has argued that a better functionalist theory results if the circularity is left in place.

27Illustration: Suppose $G$ is defined as follows.

\[
G(x) = \text{Df} (x = \text{‘snow is white’} \land \text{snow is white}) \lor \\
(x = \text{‘snow is white’ is true’} \land G(\text{‘snow is white’}) \lor \\
(x = \text{‘what Epimenides says is true’} \land \neg G(\text{‘what Epimenides says is true’})).
\]

Then, in the actual situation, this is an extensionally adequate definition of ‘true in English’ over the set \{‘snow is white’, “snow is white’ is true’, ‘what Epimenides says is true’\}—assuming that in the actual situation, the Epimenides sentence is paradoxical. Method # 5 allows an argument for the circularity of ‘true in English’ to be based on this narrowly adequate definition.