International Journal of Theoretical Physics, Vol. 31, No. 4, 1992

Memory Systems, Computation, and the Second Law of Thermodynamics

David H. Wolpert[1]

Received March 11, 1991

A memory is a physical system for transferring information from one moment in time to another, where that information concerns something external to the system itself. This paper argues on information-theoretic and statistical mechanical grounds that useful memories must be of one of two types, exemplified by memory in abstract computer programs and by memory in photographs. Photograph-type memories work by exploiting a collapse of state space flow to an attractor state. (This attractor state is the "initialized" state of the memory.) The central assumption of the theory of reversible computation tells us that in *any* such collapsing, regardless of whether the collapsing proceeds from the past to the future or vice versa, the collapsing must increase the entropy of the system. In concert with the second law, this establishes the logical necessity of the empirical observation that photograph-type memories are temporally asymmetric (they can tell us about the past but not about the future). Under the assumption that human memory is a photograph-type memory, this result also explains why we humans can remember only our past and not our future. In contrast to photograph-type memories, computer-type memories do not require any initialization, and therefore are not directly affected by the second law. As a result, computer memories can be of the future as easily as of the past, even if the program running on the computer is logically irreversible. This is entirely in accord with the well-known temporal reversibility of the process of computation. This paper ends by arguing that the asymmetry of the psychological arrow of time is a direct consequence of the asymmetry of human memory. With the rest of this paper, this explains, explicitly and rigorously, why the psychological and thermodynamic arrows of time are correlated with one another.

INTRODUCTION

Many studies have investigated the relations between various aspects of the different arrows of time (Gold, 1967; Davies, 1974; Layzer, 1976; Bitbol, 1988; Wolpert, 1988; Hawking, 1988; Wheeler and Zurek, 1983). Most of

[1]Theoretical Division and Center for Non-linear Studies, MS B213, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

0020-7748/92/0400-0743$06.50/0 © 1992 Plenum Publishing Corporation

these studies take it as given that the psychological arrow of time derives from the thermodynamic arrow of time (i.e., from the second law of thermodynamics). Yet until now no mathematical proof of this connection has been offered. This has allowed some to even go so far as to make the claim that the two arrows of time are not related at all (Popper, 1965).

Without reducing the psychological arrow of time to a mathematically well defined phenomenon, there is no way to rigorously prove a relation between the psychological arrow of time and the thermodynamic one. In this paper, the "mathematically well-defined phenomenon" is taken to be the human ability to remember the past but not the future.

This paper is primarily an analysis of memory systems and their relationship with the second law. This analysis shows that the asymmetry of human memory is a direct reflection of the asymmetry of the second law. The implication is that if the second law "went the other way," then we would remember the future, not the past, and the psychological arrow would point toward the past, not the future.

Memory

Before presenting an outline of this paper, it will help to present an informal summary of what is meant by "memory". For the purposes of this paper, a memory system is any physical system whose state at the present time t_0 can be proven to provide information concerning the state of the world external to the memory system at a time $t_1 \neq t_0$. Intuitively, a memory system is a means of ferrying information from one moment in time to another. Perhaps the simplest example of a memory system is a photograph on a piece of film. The film is the memory system. Its current state at t_0 (i.e., the image on the film) provides "information concerning the state of the world external to the memory system at a time $t_1 \neq t_0$". Although this particular example of a photograph is asymmetric in time ($t_0 > t_1$), in general no *a priori* temporal asymmetry is assumed: t_1 can either precede t_0 or come after it.

More formally, let S be the space of a memory system and let W be the state space of the world external to S. We are given some information about the current state of the memory. For example, this information could be an exact specification of the current value of S: $S(t_0) = s_0$, where t_0 is the present. In general, we might also be given some extra information J which does not directly concern $S(t_0)$. A memory system is used by taking the value of s_0 together with J and then inferring something about the state of W at some time $t_1 \neq t_0$. The only inference tools at our disposal are the laws of physics and, if needed, the information theory principle of entropy maximization

(MaxEnt) (Jaynes, 1957*a,b*, 1982; Smith and Erickson, 1989; Skilling, 1989*a,b*). In terms of probability distributions, a memory system is a system designed so that J, the fact that $P_{t_0}(s \in S) = 0$ for $s \neq s_0$, and Hamiltonian dynamics, used together (along with MaxEnt, if need be), result in a non-uniform distribution $P_{t_1}(W)$.

Intuitively, a memory system will work if we can conclude from J and s_0 that there is an interaction between S and W at some time between t_0 and t_1 which serves to correlate $P_{t_0}(S)$ and $P_{t_1}(W)$. Colloquially, s_0 is the "memory" of {the state of W at t_1}, and the memory was "stored" in S during the interaction between S and W. The sharper the constraints imposed by s_0 on $W(t_1)$, the sharper the memory.

Note that it is *not* enough that $P_{t_0}(S)$ and $P_{t_1}(W)$ happen to be correlated to say that "$P_{t_0}(S)$ serves as a memory of the state of W at t_1". In addition, we require that correlation itself is deducible, from $P_{t_0}(S)$ and J. The reason for this is that even if $P_{t_1}(W)$ and $P_{t_0}(S)$ are perfectly correlated, if the user of the memory cannot deduce this correlation, then the memory system S does not convey any useful (to the user of the memory) information from another time to the present.

Also note the implied existence of an observer in all of this. *Someone* is doing the remembering, i.e., someone is observing $P_{t_0}(S)$ and J and thereby inferring something concerning W at t_1. This observer need not be all-knowing. For example, in general the observer will only have access to some of the degrees of freedom of S. Of course, only those degrees of freedom which *are* observed can be used to remember something concerning W at t_1.

We have four distributions involved in memory: $P_{t_0}(S)$, $P_{t_0}(W)$, $P_{t_1}(S)$, and $P_{t_1}(W)$. The distribution $P_{t_1}(W)$ is what we (i.e., the observer) wish to deduce. Therefore two of the distributions are possible sources of J: $P_{t_0}(W)$ and $P_{t_1}(S)$. As it turns out, both of these distributions can help constrain $W(t_1)$.

Useful memory systems which exploit information about $P_{t_0}(W)$ are invariably of the form which will be designated as *c-type* memory systems. These systems work by evolving the joint system $S \times W$ through time from t_0 to t_1. For example, if $P_{t_0}(W)$ specified exactly the phase space position at t_0 of the outside world, and if $P_{t_0}(S)$ does the same for the memory system, then we can deterministically evolve the joint system through time to determine the exact state of W at time t_1. To agree with the colloquial use of the term "memory," for these types of systems we usually require that the deduced state $W(t_1)$ varies with the value of s_0. (Otherwise, $J = W(t_0)$ fully specifies $W(t_1)$, and the memory system S is superfluous.)

As will be seen, "memory" as it occurs in computer programs is a c-type memory system. Furthermore, c-type memory is symmetric in time; it can be used to "remember" the future as well as the past. The time symmetry

of such a memory is in accord with the well-known fact (Bennett, 1982, 1988; Landauer, 1961, 1985; Fredkin and Toffoli, 1982) that computers, unlike the human brain, can be run in a completely time-reversible fashion. The primary drawback of c-type memory systems is that they require W to be small enough and well-ordered enough so that it is feasible both to have $P_{t_0}(W)$ sharply peaked and to evolve the joint system $S \times W$. This drawback becomes particularly pronounced when W is the entire physical universe; few (if any) naturally occurring memory systems are c-type.

Besides $P_{t_0}(W)$, the other possible source of J is $P_{t_1}(S)$. Useful memory systems which exploit $P_{t_1}(S)$ are invariably of the form which will be designated as *p-type* memory systems. Because they are privy only to the states of S, p-type memory systems cannot work like c-type memory systems by deterministic evolution of the joint system $S \times W$. To see how a p-type memory system can work, imagine that $P_{t_1}(S)$ specifies exactly the phase space position at t_1 of the memory system, and that $P_{t_0}(S)$ does the same for the time t_0. For such a situation we can start with $P_{t_1}(S)$ and calculate the state s_1 which the memory system should be in at time t_0 if there is no interaction between S and W during the interval between t_0 and t_1. If, however, $s_1 \neq s_0$, the provided value of $S(t_0)$, then we know that there must be an interaction between S and W at some time between t_0 and t_1. For the cases where there is such an interaction, the resultant perturbation from s_1 to s_0 is strongly dependent on the state of W at t_1. In fact, we can explicitly calculate what states of W at time t_1 are consistent with the information that $S(t_0) = s_0$ and $S(t_1) = s_1$. In this way, knowledge of $P_{t_1}(S)$ together with $P_{t_0}(S)$ can set constraints on $W(t_1)$.

As will be seen, "memory" as it occurs in photographs is a p-type memory system. Moreover, as is discussed later, all current evidence indicates that human memory is p-type as well.

The c-type memory systems can acquire the additional information they require (beyond that contained in $P_{t_0}(S)$) simply by expanding the scope of what is observed at the present, t_0. (Instead of just looking at the present state of S, it suffices if they also look at the present state of W.) This is not true for p-type memory systems, because the extra information p-type memory systems require exists at a different time, t_1. How can one acquire this extra information from the time t_1 when one is (by definition) stuck in the present, t_0? The details of the answer to this question are quite complex, but one rather obvious point can be made: the acquiring of the needed extra information is made extraordinarily easier if a state-space collapsing process operates in S, taking a multitude of states at the time t_2 to a single state at time t_1. (t_2 is shortly before (after) t_1 if t_1 comes before (after) t_0. For example, if $t_1 < t_0$, the collapsing process is a many-to-one mapping starting with a multitude of possible S states before t_1 and ending at t_1, with $S = s_1$.)

Given the existence of such a collapsing process, we only have to *directly* infer that the memory system is in one of the multitude of states at t_2, rather than that it is exactly in the single state at t_1.

The preceding argument is only intended to indicate the *reasonableness* of requiring that p-type memory systems involve state-space collapsing processes. a more detailed exposition indicating the *necessity* of such collapsing processes in p-type memory systems is presented below. Given a p-type memory system with such a collapsing process, it is illuminating to invoke the central assumption of the theory of reversible computation; any state-space collapsing process, whether it goes from the past to the future or vice versa, must have higher entropy on the collapsed side of the mapping. (This assumption, hereafter abbreviated as "the central assumption", is taken for granted by Bennett (1982, 1988). Landauer (1961, 1985), and Fredkin and Toffoli (1982), and is rigorously analyzed by Wolpert (1990).) Together with the second law, this assumption tells us that t_2 must be shortly before t_1, not shortly after it. This in turn means that t_1 precedes t_0, and therefore that all p-type memory systems can only be of the past.

In this temporal asymmetry of theirs, p-type memory systems contrast markedly with c-type memory systems, which do not require state-space collapsing processes. This asymmetry of p-type memory systems is caused directly by the second law—without this law, p-type systems could remember the future as easily as the past. Under the working assumption that human memory is p-type, this constitutes (the outline of) a proof of why the psychological and thermodynamic arrows of time are correlated.

In addition to memory systems having $J = P_{t_1}(S)$ and memory systems having $J = P_{t_0}(W)$, it is theoretically possible to have hybrid systems, in which J gives information about both $S(t_1)$ and $W(t_0)$. It appears that in practice, however, there exist few (if any) memory systems which exploit such a hybrid J. Accordingly, such hybrid memory systems are not considered in this paper.

As another variation, it is possible that J is independent of both $P_{t_1}(S)$ and $P_{t_0}(W)$. This is the case, for example, if J is empty, or if it contains prior knowledge not contained in $P_{t_1}(S)$ or $P_{t_0}(W)$. (An example of such prior knowledge is the following information: In N randomly chosen experiments where $S(t_0)$ equalled the value s_0, the value of $W(t_1)$ was always w_1.) Such memory systems are called *b-type* memory systems. They are usually not very useful, for reasons that are discussed in Appendix A. Moreover, in the case where J involves prior knowledge, analysis of b-type memory systems implicitly extends far beyond analysis of the physics of Hamiltonian evolution in $S \times W$ combined with the principle of MaxEnt. An analysis of such b-type memory systems would also necessitate investigating how the prior knowledge can be acquired, how reliable it is,

how applicable it is, etc. For these reasons, b-type memory systems will not be considered in this (already lengthy) paper in any depth.

Finally, there might be circumstances in which one is interested in using $P_{t_0}(S)$ and J to infer something other than $P_{t_1}(W)$. For example, one might have some reason to want to infer $P_{t_0}(W)$ rather than observe it directly. Some such cases are discussed in passing in this paper. The primary goal of this paper, however, is to investigate systems for inferring $P_{t_1}(W)$.

Outline

This paper is organized as follows. Memory systems which exploit $P_{t_0}(W)$ in addition to $P_{t_0}(S)$ are investigated in Section 1. In particular, it is argued there that all useful memory systems of this sort must obey (1.1). the formal definition of c-type memory systems presented in Section 1.

Memory systems which exploit $P_{t_1}(S)$ in addition to $P_{t_0}(S)$ are investigated in Section 2. In particular, it is argued there that all useful memory systems of this sort must obey (2.1), the formal definition of p-type memory systems presented in Section 2. Section 2 goes on to discuss in detail the important fact that p-type memory systems (unlike c-type memory systems) explicitly rely on the existence of a process directing state-space flow to an attractor state, the state of the memory system when it is "initialized."

Like (1.1), the definition (2.1) is temporally symmetric. (In particular, the state-space flow is allowed to collapse in going from the future into the past as well as vice versa.) The overt temporal symmetry of (2.1) ensures that any explanation of the asymmetry of real-world p-type memory (and hence of the psychological arrow of time) does not arise through asymmetric definitions.

It is only with symmetric definitions that we can rigorously resolve the following paradox: The future is both the temporal direction into which "information is dissipated" (due to the second law) and the direction into which "information can be preserved" (via p-type memory).

Section 3 of this paper is intended to provide the reader with an intuitive understanding of the material in Section 2. This section is made up of a series of examples of how the p-type memory systems found in nature (e.g., craters on the moon, a footprint on a beach, a photograph) meet the definition given in (2.1). In particular, Section 3 argues on commonsense grounds (as opposed to on the formal grounds of Section 2 and Wolpert (1990)) that the asymmetry of the second law induces an asymmetry in the allowed initialization processes of p-type memory systems. In fact, initialization in such systems seems to always *exploit* the second law.

Section 4 concludes this paper by arguing in detail that the asymmetry of human memory is the sole cause of the psychological arrow. Since the

psychological arrow is an ill-defined phenomenon (at least until it is reduced to something mathematically precise, like memory), this section is necessarily somewhat philosophical in nature.[2]

The work presented in this paper can be viewed in two ways. First and most naturally, it can be viewed as an extension of the work of Jaynes and many others (Jaynes, 1957*a*,*b*, 1982; Smith and Erickson, 1989; Skilling, 1989*a*,*b*) on applying the ideas of information and probability theory to physical systems.[3] Alternatively, this paper can be viewed as related to the work of Bennett and others (Bennett, 1982, 1988; Landauer, 1961, 1985; Fredkin and Toffoli, 1982) investigating Maxwell's demon and the thermodynamic behavior of computing devices. The main differences are (1) the work presented in this paper is completely formal in its treatment of memory and has a correspondingly wider scope than the preceding work, extending the concept of memory beyond computing devices, (2) this paper shows that running an irreversible computer program does not necessitate an asymmetric computer program memory, (3) the analysis of this paper shows explicitly that computer program memory and human memory are of a fundamentally different nature, and shows that these are the only two kinds of useful memory, and (4) this paper investigates the implications of the analysis of p-type memory systems for the psychological arrow of time.

All of the arguments of this paper are presented in a classical context. Similar arguments hold for quantum systems. There is considerable variation in the degree of formal rigor in the arguments presented in this paper; the less rigorous arguments should be viewed more as an initial foray into a subtle subject than as a conclusive treatment of a fully understood issue.

1. c-TYPE MEMORY SYSTEMS

This section investigates memory systems which rely on $P_{t_0}(W)$. First, this section examines a real-world example of such a memory system. Then this example is generalized to give (1.1), the definition of c-type memory systems. After this, the question of how memory systems which rely on

[2]For those who are not convinced by the arguments of this section, this paper does not explain the psychological arrow *in its entirety* as a consequence of the thermodynamic arrow. Rather, such people should view this paper as establishing the relation between the thermodynamic arrow of time and a subset of the psychological arrow of time, namely that part of the psychological arrow bound up in the asymmetry of human memory.

[3]There are a number of interesting philosophical aspects of Jaynes' principle "choose the ρ which maximizes the entropy, subject to the known constraints." For example, this principle says that what constraints are known to the researcher modify the physics (e.g., the entropy) of the system, even if the researcher never physically contacts the system. Just as in quantum mechanics, in statistical mechanics the observer affects the physics of the system *of necessity*.

$P_{t_0}(W)$ can work is investigated from an information-theoretic and statistical mechanical point of view. This investigation indicates that *all* such memory systems must obey (1.1). Finally, this section ends by discussing the time-symmetry characteristics of c-type memory systems as well as the limitations which exist on using such systems in the real world. (It is these limitations which preclude c-type memory systems from being candidates for how human memory works.)

1.1. Memory in a Computer

In this subsection we investigate an example of a memory system which exploits $P_{t_0}(W)$: abstract memory in abstract computers. We deal with "abstract" memory and computers so as to keep the conversation general, i.e., so as to avoid issues concerning particular hardware implementations and so as to avoid issues concerning the physical world external to the computer. This means in particular that here, in this example, the dynamical laws with which we are *directly* concerned are those of the computer's programming language, not those of Hamiltonian physics. (It is implicitly assumed, however, that the programming language can be implemented in the physical world, so that its dynamics can be expressed in terms of Hamiltonian dynamics if so desired.)

The discussion in this subsection is intended to be pedagogical. Accordingly, there is no reason to be so formal as to consider (for example) Turing machines; anyone who has ever written a computer program should be able to follow the analysis presented here. Although in the next subsection this analysis will be generalized into a time-symmetric form, the discussion in the current subsection largely concerns computer memory as we humans usually think of it, i.e., computer memory of the past.

In that it is pedagogical, this subsection implicitly assumes one particular colloquial meaning for the word "memory" when it is applied to (abstract) computer programs. Sociological questions of whether or not most researchers agree with this meaning are not relevant; the discussion is only intended as an illustration of memory systems relying on $P_{t_0}(W)$.

Consider an abstract piece of Random Access Memory (RAM) being used in an abstract computer program. For simplicity, label the address of that piece of RAM as 0001. Assume that at time t_0 there is a particular pattern s_0 at address 0001. We say that "s_0 is the state of 0001 at time t_0." In what sense is s_0 a memory of the state the outside world was in at some time before t_0?

The "memory system" in this case is the RAM at 0001. The "outside world" is the rest of the abstract computer, including both the entire program code running on the computer, all of the computer's data-storing RAM

(outside of that at 0001, of course), all registers, the program counter, etc.

Knowledge of the value s_0 alone (!) tells us nothing about the state of this outside world, at any time. Therefore, by itself, 0001 does not constitute a useful memory system. To transfer data from one moment in time to another (i.e., to have the value of s_0 tell us something concerning the state of the world external to 0001 at a time $t \neq t_0$), abstract computer program memories rely on the fact that it is possible to deterministically evolve the entire computer system, program, RAM, and all, through time. The memories of such systems work if it is provable from the computer program that s_0, the current contents of 0001, is a reflection of some data which existed in a certain place in the computer in the past. In such a case s_0 is a "memory" of that data from the past.

As an example, consider the following code segment embedded in some larger program:

```
. . .
SAVE_LOOP=LOOP;
LOOP=0;
. . .
```

We say that {the state of SAVE_LOOP after this piece of code has executed} is a memory of {the state loop was in just before this code was executed}. The abstract piece of RAM containing SAVE_LOOP is the memory system. Everything else in the abstract computer is the outside world. In particular, the RAM containing LOOP is part of SAVE_LOOP's outside world.

Define t_0 to be the moment this code segment has finished. Then SAVE_LOOP is a "memory of the state of LOOP at a time previous to t_0" in the sense that we can conclude that the value of the memory system containing SAVE_LOOP at time t_0 tells us something about the state of that system's outside world (namely, the value of LOOP) at an earlier time. How do we reach this conclusion that before this code was executed LOOP had the value currently found in SAVE_LOOP? We reach it by logically backtracking the *entire* computer program in conjunction with the memory containing SAVE_LOOP.[4] In other words, along with SAVE_LOOP we deterministically evolve all of SAVE_LOOP's outside world, including both the coding sections and the data sections of the program.

We are not making any initialization assumptions. As a result, unless we use deterministic evolution to exploit knowledge of the code segment, we

[4]The reasoning is as follows: At the beginning of the code segment SAVE_LOOP is set to LOOP's value. Never again in the code segment is SAVE_LOOP touched. Therefore at the end of the code segment it contains the value that LOOP had at the beginning of it.

can conclude nothing from {the state of SAVE_LOOP after the piece of code has executed}.

Consider the following code segment:

```
    . . .
    M=3;
    LOOP=0;
10: LOOP=LOOP+1;
    J=J+1;
    if (J<10)
      {goto 10;
      }
    . . .
```

We can say that the state of LOOP after the code has executed is a memory of the state of J before the code executed. Together with its external world (i.e., the code segment), the value of LOOP after the code segment has executed suffices to fix exactly the value of J just before the code segment has started.

Note, however, that we do *not* say that the state of LOOP after the code is a memory of the state of M before the code, despite the fact that the state of LOOP after the code, together with the state of LOOP's outside world after the code, constitutes sufficient information to determine the state of M before the code. (M's state before the code is given directly by M's state after the code, which in turn is a subset of LOOP's outside world after the code.) Knowing the state of LOOP after the code adds nothing to our knowledge of the state of M before the code—knowledge of LOOP's outside world is both necessary and sufficient to reach conclusions about the state of M before the code. This is why we do not view LOOP after the code as a memory of M before it. Generically, we require that knowledge of $S(t_0)$ must help us fix $W(t_1)$ in order to say that the state $S(t_0)$ is a memory of $W(s_1)$.

1.2. c-Type Memory Systems

Although the preceding example was presented as memory of the past, we can easily generalize from it to get a time-symmetric definition of memory systems which work via the same logical mechanism as memory in abstract computers. This generalization, (1.1), defines c-type memory systems. It is presented and discussed in this subsection. We can hypothesize that not only abstract computer memory systems, but in fact *all* memory system which exploit $P_{t_0}(W)$ are necessarily c-type. Arguments establishing this hypothesis are presented in the subsection following this one.

From now on the conversation moves from the abstract to the concrete, i.e., from now on we are to think of the memory system and the outside world as real physical systems evolving according to Hamiltonian dynamics. This is to allow us to apply statistical mechanics later.

(1.1) If a system S is in a state s_0 at a time t_0, that state is said to be a *c-type memory*, be it of the past or of the future, and the system is said to be a *c-type memory system* if:

1. In addition to s_0, one also has some information concerning the state of W, the world external to S, at that same time t_0.

2. Using $S(t_0) = s_0$ and the information concerning $W(t_0)$, it is possible to evolve the joint system $S \times W$ through time to a time t_1 and come to a conclusion about $W(t_1)$, the state of W at t_1.

3. With the possible exception of the precise "information concerning $W(t_0)$" provided, nothing precludes the memory system's being closed for the time period between t_0 and t_1, and nothing precludes the memory system's being open for that time period.

The rationale for requirement 3 is that *a priori* we want to allow both the possibility that a memory is stored in S and the possibility that S is "empty," containing no memory of its external world. To make (1.1) more in accord with colloquial usage, one usually adds the following requirement to (1.1):

4. For the provided information concerning $W(t_0)$, there must exist changes to s_0 which modify the conclusion about $W(t_1)$, and similarly for the provided s_0 there must exist changes to the provided information concerning $W(t_0)$ which modify the conclusion concerning $W(t_1)$.

The rationale for this fourth requirement is that if the conclusion does not vary with the information concerning $W(t_0)$, then we have a b-type memory. On the other hand, if it does not vary with s_0, then our "conclusion about $W(t_1)$" follows from simply evolving W by itself through time, and S adds nothing to our knowledge[5]. This dependence of our "conclusion about $W(t_1)$" on s_0 constitutes the memory.

[5]Despite requirement 4, there is still a sense in which the semantics of (1.1) might conflict with colloquial use of the term "memory." This arises from the requirement, implicit to all memory systems investigated in this paper, that S is to serve not as a memory of itself, but only of the world outside of itself. (We try to infer $P_{t_1}(W)$, not $P_{t_1}(S)$, from $P_{t_0}(S)$.) The reason for this requirement is that we are implicitly viewing memory systems as devices which exploit *whatever means they can* to convey information from one moment to another. That is their purpose. Now one *could* try to deduct $P_{t_1}(S)$ via $P_{t_0}(W)$ (for example). However, the easiest and most efficacious way to have a memory system remember its own past is simply to ensure it is always closed and then evolve it through time. For such a case no external world is necessary at all—we are just evolving S. This is a trivial and uninteresting case, which is why it is excluded from the definitions in this paper.

Note that whether or not a given system is a c-type memory depends not only on that system itself, but also on the knowledge one is given concerning that system's outside world. For example, for some states $W(t_0)$ is might follow from $S(t_0) = s_0$ that an interaction between S and W occurs at some time between t_0 and t_1. In such a case, $S(t_0)$ might tell us something about the state of W at times other than t_0, i.e., $S(t_0)$ might be a memory of such a state. This will be true for, example, if S is the abstract RAM containing SAVE_LOOP and if the present is immediately after completion of the following code segment:

```
. . .
SAVE_LOOP = LOOP;
LOOP = 0;
. . .
```

As before, for the external world of {the RAM containing SAVE_LOOP}, that RAM serves as a memory system.

For other states $W(t_0)$, however, even for the same value s_0 of $S(t_0)$, we might be able to conclude that there is *not* an interaction between S and W at some time between t_0 and t_1. In such a situation, $S(t_0)$ will tell us nothing about the state of W at times other than t_0. In other words, for such a $W(t_0)$, varying s_0 will not change our conclusion concerning $W(t_1)$, and $S(t_0)$ is not a memory of anything to do with W. This would be the case, for example, if S were still the abstract RAM containing SAVE_LOOP, but the entire abstract program is different and never touches that RAM containing SAVE_LOOP. For this external world, there is no interaction between S and W ever, *we know this*, and therefore S provides no memory of the state of W at anytime whatsoever. This relation between $W(t_0)$ and the question of whether or not S is a c-type memory system is an important difference between such c-type memory systems and p-type memory systems.

Note that in their hardware implementation real-world computers might make use of processes which can be viewed as non-c-type memory systems, especially in their input–output devices. (In particular, they might exploit p-type memory systems—see Section 2 and the EEPROM example in Section 3.) However, the means by which a typical program running on such a computer "remembers" from one moment in time to another is via c-type memory. This distinction, between the characteristics of the actual hardware implementation on a real-world computer and the theoretical requirements of that implementation, is analogous to the distinction between real-world computers which are all logically irreversible and the completely reversible manner in which one can design theoretical computers (Bennett, 1982, 1988; Landauer, 1961, 1985; Fredkin and Toffoli, 1982).

1.3. Why Memory Systems Which Exploit $P_{t_0}(W)$ Are c-Type

Using information theory and statistical mechanics, this subsection starts by presenting a rigorous and general formalism for investigating memory systems. This subsection then presents an argument that all physical memory systems relying on $P_{t_0}(W)$ must meet the requirements of (1.1).

Let the space S be the state space of the memory system. Denote the phase space associated with the system by Γ_S. Often (especially ion the analysis of p-type memory systems) S will be a partition on this phase space, reflecting the fact that any real-world measuring device has finite precision. Let W be the state space of the universe external to the memory system. It, too, may or may not be a partition on its phase space. Denote the phase space associated with W by Γ_W.[6] We are provided with the probability distribution over S at t_0, $P_{t_0}(S)$. We are trying to use $P_{t_0}(S)$ to gain information concerning $W(t_1)$, where $t_1 \neq t_0$. We make no *a priori* assumptions concerning whether $t_1 < t_0$ or vise versa.

Consider the space $D \equiv S \times W \times S \times W$, representing S at t_0, W at t_0, S at t_1, and W at t_1. If $P(D)$ is the probability distribution over D, then, for example, $\int P(D)\, dW(t_0)\, dW(t_1)\, dS(t_0)$ gives the probability distribution over S for the time t_1. In the usual way, the principle of maximum uncertainty induces a probability distribution over D, namely the distribution of maximum entropy subject to the dynamical laws relating $(S \times W)$ at t_0 and $(S \times W)$ at t_1, and subject to whatever external constraints apply (Jaynes, 1957*a*,*b*, 1982; Smith and Erickson, 1989; Skilling, 1989*a*,*b*).

We assume that other than (possible) interactions with each other and a means for imposing whatever external constraints apply, both S and W are closed systems. Information concerning $P(S)$ and $P(W)$ at t_0 is the only external constraint that we have. Therefore $P(D)$ is the distribution with maximal entropy such that (a) it obeys the constraints implicit in the dynamical laws relating $(S \times W)$ at t_0 and $(S \times W)$ at t_1, (b) its projection onto $S(t_0)$ gives the provided distribution $P_{t_0}(S)$, and (c) its projection onto $W(t_0)$ gives the provided distribution $P_{t_0}(W)$. The projection of this maximal entropy distribution onto $W(t_1)$ represents what we can infer about $W(t_1)$, given only $P(S)$ and $P(W)$ at t_0.

[6]In general, we allow almost any kind of external system to interact with S. Therefore, for example, we do not know even how many particles there are in the external system. Rather than work with a probability distribution over kinds of external systems (as well as over the phase space of each such system), we require that W be big enough to encompass as a subset of itself any kind of external system with which the memory is likely to interact. Different kinds of external systems being in contact with S correspond to different subsets of W, i.e., they correspond to different subregions of Γ_W. In this way we can restrict our probability distribution to a single phase space, Γ_W.

Some comments are in order. First note that it is not the entropy of $P(D = S \times W \times S \times W)$ which we want to maximize, but rather the entropy of $P(\Gamma_D) \equiv P((\Gamma_S \times \Gamma_W)_{t_0} \times (\Gamma_S \times \Gamma_W)_{t_1})$, the probability distribution over the product phase space associated with D. However in general, when (as in this paper) fine-graining is assumed, we cannot directly maximize the entropy over $P(\Gamma_D)$. To see this, write $P(\Gamma_S \times \Gamma_W)$ as $P(\Gamma_{S \times W})$ for short. Let γ be any point which lies in the support of $P(\Gamma_{S \times W})$ at t_1. Now examine the line $(\gamma \in (\Gamma_{S \times W})_{t_1}, x \in (\Gamma_{S \times W})_{t_0})$ in the space Γ_D traced out by fixing γ and letting x vary. Since evolution through $\Gamma_{S \times W}$ is deterministic, γ corresponds to a unique point in the support of $P(\Gamma_{S \times W})$ at t_0, γ'. Therefore $P(\Gamma_D)$ is zero everywhere along the line (γ, x) except where $x = \gamma'$. In other words, $P(\Gamma_D)$ has the form of a delta function.[7] This means that the fine-grained entropy over the space Γ_D has the form

$$-\int d(\gamma_{t_0})\, d(\gamma_{t_1}) \{\ldots \delta[\gamma_{t_1} - \Phi_{(t_1 - t_0)}(\gamma_{t_0})]\}$$

$$\times \ln \ldots \delta[\gamma_{t_1} - \Phi_{(t_1 - t_0)}(\gamma_{t_0})]\} \sim -\ln[0] = \infty$$

(γ_{t_i} is the phase coordinate of the distribution for time t_i, and $\Phi_\tau(\cdot)$ is the injective phase space evolution operator taking γ_{t_i} to $\gamma_{(t_i + \tau)}$.) Therefore we cannot work directly with this entropy.

To avoid these difficulties in calculating fine-grained entropies, it is conventional to exploit Liouville's theorem, which tells us that the entropy of $P(\Gamma_{S \times W})_{t_0}$ equals the entropy of $P(\Gamma_{S \times W})_{t_1}$, and *define* the entropy of $P(\Gamma_D)$ as being the entropy of either $P(\Gamma_{S \times W})_{t_0}$ or $P(\Gamma_{S \times W})_{t_1}$.[8] It is this entropy which we maximize.

[7] Another way of seeing this is to note that due to the injectivity of evolution in $\Gamma_{S \times W}$, the support of $P(\Gamma_D)$ has the same dimension as the support of $P(\Gamma_{S \times W})$, i.e., $P(\Gamma_D)$ is nonzero only over a hypersurface through Γ_D. This in turn means that along certain projections $P(\Gamma_D)$ has the form of a delta function.

[8] There is another possible way around the impasse that for any allowed system the fine-grained entropy over all Γ_D is infinite. This is to integrate $\rho \ln \rho$ only over the hypersurface in Γ_D over which it is allowed to be nonzero. With this procedure, instead of evaluating a volume integral over a delta-function distribution, we evaluate a surface integral over an everywhere-finite distribution. Presumably this procedure of extremizing fine-grained entropy over this hypersurface in Γ_D is equivalent to the procedure of extremizing entropy over either $(\Gamma_{S \times W})_{t_0}$ or $(\Gamma_{S \times W})_{t_1}$. If these two procedures turned out *not* to be equivalent, then unless there is some way to choose between them, the entire technique of MaxEnt as it applies to statistical mechanics would have a problem. To wit, should we maximize entropy over a single phase space at a single time, Γ_t, as is conventional (Jaynes, 1957*a*, *b*, 1982; Smith and Erickson, 1989; Skilling, 1989*a*, *b*), or should we maximize it over the allowed hypersurface in some product space $\Pi_i \{\Gamma_{t_i}\}$?

As an aside, note that if coarse-graining were used, then in effect dissipation would be occurring within the partition elements of S and W. We would no longer be evolving strictly

Second, note that the constraints under which we must calculate entropy are of two types. The first is that of following Hamiltonian evolution in $\Gamma_S \times \Gamma_W$ when going from t_0 to t_1. This constraint is simply a boundary on the region of allowed states in the space $(\Gamma_S \times \Gamma_W)_{t_0} \times (\Gamma_S \times \Gamma_W)_{t_1}$. The second type of constraint is an external constaint, like information concerning $P_{t_0}(S)$. A particularly illuminating example of this second type of constraint is the situation in which we are given that $P_{t_0}(S)$ equals zero for all S values except for one, σ. In this case our external constraint is just like the constraint of Hamiltonian evolution; it, too, is simply a restriction of the allowed states in Γ_D to a certain subregion. The intersection of the two subregions associated with these two constraints gives the region Q in Γ_D over which we must maximize entropy. There is no other restriction on the probability distribution, which means that the entropy is maximized when the distribution is flat across Q and zero elsewhere. (More precisely, given our modified definition of entropy, we require that the distribution is flat across either of the two projections of Q onto $(\Gamma_S \times \Gamma_W)_{t_0}$ or of Q onto $(\Gamma_S \times \Gamma_W)_{t_1}$.)

This full mathematical structure is necessary to carry out formal proofs (see for example Appendix B). Fortunately, one does not need to use the full structure to justify (1.1). To start, note that *a priori*, we do not know if there was an interaction between S and W in the time between t_0 and t_1, our knowledge of $P(S)$ at t_0, by itself, usually tells us little about $W(t_1)$. (Nonetheless, there are certain cases where it can be relatively informative. A discussion of this special case of so-called "b-type" memory systems can be found in Appendix A.) Therefore, in general, to infer a lot about $P(W)$ at t_1, we need some extra information besides $P_{t_0}(S)$. For the rest of this subsection, attention will be restricted to the case where our extra information concerns $P(W)$ at some moment other than t_1. (The other case is dealt with in Sections 2 and 3.)

Now we always can evolve W deterministically for those times it is not interacting with S. Therefore, information concerning $P_t(W)$ for any time t on the same temporal side as t_1 of $\{W$'s interaction with $S\}$ is equivalent to information concerning $P_{t_1}(W)$. Similarly, information concerning $P_t(W)$ for any time t on the same temporal side as t_0 of $\{W$'s interaction with $S\}$ is equivalent to information concerning $P_{t_0}(W)$. Therefore, without loss of generality, we can take "$P(W)$ at some moment other than t_1" to mean $P_{t_0}(W)$. (Similarly, in Section 2 we will take "$P(S)$ at some moment other than t_0" to mean $P_{t_1}(S)$.) This justifies requirement 1 of (1.1).

---

according to the Hamiltonian of $S \times W$, and Liouville's theorem would not apply. However, now Liouville's theorem would not be needed, since $P(\Gamma)_D$ would not have the form of a delta function and its entropy would be well-defined.

We are given some information concerning $P_{t_0}(W)$ and $P_{t_0}(S)$. Therefore we are given some information concerning $P_{t_0}(V \equiv S \times W)$. We are given no other information. To deduce $P_{t_0}(V \equiv S \times W)$ exactly from our partial information concerning it, we use the principle of entropy maximization. (It having been argued above that we should maximize the entropy of $P(\Gamma_V)$ rather than $P(\Gamma_D)$.) Hamiltonian dynamics then maps $P_{t_0}(V \equiv S \times W)$ to a probability distribution over $W(t_1)$. This justifies the claim that memory systems which exploit information concerning $P_{t_0}(W)$ of necessity meet the second requirement of (1.1). The third requirement of (1.1) is axiomatic to all memory systems. The fourth requirement is really just a matter of semantics, and was justified in the previous subsection.[9] Therefore we have justified in full our hypothesis that memory systems which exploit information concerning $P_{t_0}(W)$ are of necessity c-type.

What properties should $P_{t_0}(W)$ have in order to result in a maximally efficacious memory? To answer this question, first assume for simplicity that our knowledge of $P_{t_0}(S)$ is that S is in the state σ at $t = t_0$. Now let x be either of the two variable $\{S \text{ at } t_1\}$ or $\{W \text{ at } t_0\}$. It is proven in Appendix

[9]There is a subtle issue concerning requirement 4 which is brought to the fore by the analysis of this subsection. To require that "for the provided information concerning $W(t_0)$, the conclusion about $W(t_1)$ must vary with s_0" formally means that making infinitesimal variations to $P_{t_0}(S)$ while keeping $P_{t_0}(W)$ unchanged results in changes to our deduced $P_{t_1}(W)$ (similarly for the requirement that "for the provided s_0, the conclusion about $W(t_1)$ must vary with varying the information concerning $W(t_0)$.") There is no problem with this formalization of requirement 4 of (1.1) in the real world, since the microscopic laws of physics are reversible, and one can always evolve *any* probability distribution across V backward or forward in time with impunity. However, when we are operating in an "abstract world," as in Sections 1.1 and 1.2, it might be that changes to s_0 without corresponding changes to $W(t_0)$ result in a contradiction. Such will be the case if the "abstract world" is irreversible. For example, consider the code segment ". . . , SAVE_LOOP = LOOP; . . . ." Let the memory system S be the abstract RAM housing SAVE_LOOP. Varying the state of $\{S$ after the code segment has executed$\}$ without also varying the state of $\{W$ after the code segment has executed$\}$ results in a contradiction, since with such a variation SAVE_LOOP after the segment no longer equals LOOP after the segment, despite the fact that the program says that they were copied to be equal. For such situations, "for the provided information concerning $W(t_0)$, the conclusion about $W(t_1)$ must vary with s_0" means that when infinitesimal variations to $P_{t_0}(S)$ are made *and* $P_{t_0}(W)$ *is varied accordingly*, we get changes to our deduced $P_{t_1}(W)$. Alternatively, one can just restrict consideration to abstract *reversible*, digital, finite computers. For such computers, changes to s_0 unaccompanied by changes to $W(t_0)$ cannot result in a contradiction, and we do not need to require that "$P_{t_0}(W)$ *is varied accordingly*." To see why this is so, examine the mapping over the space of {all possible patterns in the data segment of memory} induced by a single step of a program running on such a reversible, digital, finite computer. Due to reversibility, the mapping must be injective. Due to the finiteness of the space of {all possible patterns in the data segment of memory}, it therefore follows that the mapping is surjective as well, i.e., the mapping must be bijective. Therefore, for these systems, any S and any W can be evolved together in any direction in time without getting a contradiction, so (in particular) s_0 can be varied while $W(t_0)$ is left unchanged.

B that a necessary condition for $P(x)$ to have the strongest possible effect on our knowledge of $W(t_1)$ (i.e., a necessary condition for maximizing the Shannon information of $P(W)$ at t_1) is that $P(x)$ be a delta function. In other words, using MaxEnt with a distribution $P(x)$ results in minimal entropy over $P_{t_1}(W)$ iff $P(x)$ is a delta function. As a result, we want $P_{t_0}(W)$ to be as close to a delta function as possible. (For similar reasons, for p-type memory systems we will want $P_{t_1}(S)$ to be as close to a delta function as possible—see Sections 2 and 3.)

1.4. Time-Symmetry and c-Type Memory Systems

It is important to note some distinctions between the memory systems discussed above and human memory. An abstract computer's RAM, when used as a c-type memory to give information concerning a previous state of the computer, is not a p-type memory system in the sense of definition (2.1), given below. On the other hand, as is discussed later, human memory seems to be p-type,[10] and at a minimum certainly is not c-type—to remember

[10]Human memory is not well understood. However, some general comments can be made. First, it appears that human memory is anatomically localized to certain regions of the brain (e.g., the hippocampus). In addition, human memory is associative. Moreover, current wisdom says that the storing of a memory in those regions corresponds biologically to the modification of synaptic weights between the neurons in those regions, perhaps according to a Hebb-type rule. The memories themselves, be they of abstract thoughts or of sensory impressions of the world outside the brain, represent information from outside of these regions. Second, rather than worrying about the precise biological process involved in storing and recalling a memory in the memory centers of the human brain, we can abstract those centers and describe them as an input–output mapping taking certain neural firing patterns fed in from the outside and transforming them into other neural firing patterns which are then fed back to the outside. Each input–output pair in such a mapping serves as a memory system. Usually one thinks of the input to such a memory system as a particular question, and output as an answer to the question. For example, the input could be "My first grade teacher's name:" and the output of the memory could be "Mr. Smith" (the complete input–output pair is "My first grade teacher's name: Mr. Smith.") Different memory systems are delineated by different input questions. New memories are stored by modifying the appropriate input–output pair, i.e., by having the appropriate memory system interact with the external world and thereby induce a correlation between the state of the memory system (i.e., the output corresponding to its input) and the relevant aspect of that external world. We have confidence in a memory system only to the degree that we believe the completed input–output pair truly reflects "information concerning the world external to the memory system at a time different from the present." Note the necessity in all this of a special "I don't know/remember" output for those inputs whose corresponding output from the outside has not yet been recorded. In terms of the framework presented in this paper, this "I don't know/remember" output corresponds to the state $P_{t_1}(S)$. This special state is the known pre-{interaction with the outside world} state of the memory system. All human memory systems have their output component initialized to this special "I don't know/remember" state at first—exactly as in p-type memory systems.

something concerning the environment outside of our brains, we humans do not need to deterministically evolve that environment outside of our brains. Another difference between human memory and computer memory is associated with the fact that computers can be run in a completely reversible manner whereas human brains are time-asymmetric. (We think forward in time, not backward.) Similarly, even if human and computer memory systems used similar mechanisms to store a memory (i.e., had similar laws governing how W affects S), the way the stored memory is used to infer something concerning the external world is fundamentally different in the two types of systems. To prove this, we need do no more than note that the human brain can only remember the past. (A fact which is also true of all p-type memory systems—see Sections 2 and 3.) In contrast, c-type memory systems can remember in either direction of time: c-type memory systems work by evolving the joint system $S \times W$ via Hamiltonian dynamics, and that evolving can be into the future as easily as into the past.

This symmetry of c-type memory systems is completely overt in truly symmetric reversible computers, where, tautologically, no thermodynamic or computational phenomenon can distinguish between the two directions in time, and therefore memory cannot distinguish between the two directions. What is interesting is that this symmetry also holds for nonreversible computers; i.e., it holds *even if the program using the memory is logically irreversible.* To have some abstract RAM act as a future memory, rather than backtracking the entire computer from t_0, one now "forward-tracks" it: simply evolve forward from the current state of the entire computer and prove that the current contents of that RAM give information concerning some data which in the future will exist in a certain place in the computer external to that RAM.

For example, consider the following code segment:

```
...
Y=Z;
Y=Y+1;
...
```

Let t_0 be the moment just before the code segment has executed, and t_1 the moment just after it has executed. The state of {the abstract RAM containing Z} at t_0 gives information concerning the state of something external to that {RAM containing Z} at t_1 (namely, the contents of the abstract RAM containing Y at t_1). This arises via exploitation of knowledge of the state of Z's external world (i.e., the computer program) at t_0. Clearly, modifying the state of {the abstract RAM containing Z} modifies our conclusion about the future state of that world external to {the abstract RAM containing Z}. Similarly, modifying the program (and therefore the state of the external

world of {the abstract RAM containing Z}) modifies our conclusion about the future state of that world external to {the abstract RAM containing Z}. Therefore {the abstract RAM containing Z} meets all four requirements for being a c-type memory system, *of the future*. We have simply "forward-tracked" rather than "backtracked" the entire abstract computer.[11]

The many-to-one mappings of an irreversible computation can terminate deterministic backward evolution of the program after the evolution has only proceeded a finite distance into the past. This can occur, for example, if sometime in the past the contents of the RAM one is remembering was set by the CPU to all 0's. In such a case, one cannot evolve further back in time than when that action was taken by the CPU. (Or, to put it another way, any conclusion concerning $W(t_1)$ for times t_1 previous to such an action by the CPU is independent of $S(t_0)$. Therefore $S(t_0)$ does not serve as a memory of such a $W(t_1)$.) Many-to-one mappings can set a limit on how far into the past one can remember.

In a similar manner, many-to-one mappings can erase the contents of the RAM that is being "remembered" after the evolution gets only a finite distance into the future. This is how they can set limits on how far forward a computer's memory of the future can go. Many-to-one mappings play no temporal favorites; they are as free to restrict memory of the future as they are to restrict memory of the past.

It could conceivably be extremely helpful to have a temporally symmetric memory. Therefore the question arises; why did not evolution design human memory to be c-type? To answer this question, note that to use a c-type memory we have to understand and be able to calculate evolution in W space, as well as evolution in S space. Moreover, we would like to have $P_{t_0}(W)$ be a delta function (see the end of the previous subsection). If W possesses simple dynamics and has relatively few degrees of freedom, then a c-type memory system can be feasible. This is the case with memory in abstract computers. However, in any environment not as compact and well-ordered as a computer, $P_{t_0}(W)$ will never be a delta function, and the deterministic evolution in $\Gamma_S \times \Gamma_W$ is so unwieldy and requires the manipulations

[11] This process of forward-tracking assumes that nothing happens to render its proof concerning the future state of the external world invalid (e.g., a human does not come along at sometime in the future and twiddle the bits by hand, overriding the program). But in a similar manner, the backtracking memory assumed that nothing happened to render *its* proof invalid (e.g., a human did not come along at some time in the past and twiddle the bits by hand, overriding the program). Remembering in either direction requires the assumption that the computer as a whole is logically closed throughout the evolution of the program. Without this assumption, any "memory" is just so much informationless noise. All of this applies to probabilistic computer memories (Pearl and Crolotte, 1980) as well as to more common deterministic abstract computer memories.

of so much information extraneous to the actual memory itself as to verge on the useless. As an example of this, just try backtracking (à la c-type memory) from the current state of the atmosphere to ascertain what the sky looked like over Iowa yesterday at 9:15 AM. (Contrast this with the ease with which a p-type memory system (like a photograph) can tell one what the sky over Iowa looked like yesterday at 9:15 AM.) It is due to this difficulty with large W's that humans do not use c-type memory. We want to be able to use our memory even when W contains the entire physical universe external to our brains.

In general, in a real computer there are two places that entropy change (associated with p-type memory, many-to-one mappings, initialization, and time-asymmetry) might occur *of necessity*: in the mechanism of the input–output devices, or in the memory mechanism of the human operator, via his or her memory of the initial state of the machine. Both of these are interactions with a system external to the computer as a whole. The actions of an abstract CPU operating on an abstract RAM, independently of the outside world, do not necessarily cause a change in entropy. This is because abstract Turing machines can be constructed which are completely invertible so that no information is lost during a calculation. For such a machine, retrodiction and prediction are equally reliable, there is no sense in which there exists backward memory but not forward memory, and there does not exist an arrow of time.

2. p-TYPE MEMORY SYSTEMS

Recall that there are two useful types of memory system: those which rely on $P_{t_0}(W)$, and those which rely on $P_{t_1}(S)$. Memory systems which rely on $P_{t_0}(W)$ were investigated in Section 1. The current section investigates memory systems which rely on $P_{t_1}(S)$. First, this section presents a real-world example of such a memory system. After this, this section generalizes from this example to present (2.1), the definition of p-type memory systems. Then this section presents information-theoretic and statistical mechanical arguments justifying the assertion that all memory systems which rely on $P_{t_1}(S)$ are p-type.

In both this section and Section 3 several naturally occurring p-type memory systems (i.e., systems for which W is the entire physical universe) will be considered. Since the precise mechanism by which the human brain remembers is not well understood, only inorganic naturally occurring memory systems (e.g., a photograph) will be investigated in detail. However,

there is no reason to doubt that human memory as well is a p-type memory system,[12] and in the rest of this paper it will be assumed that it is.

2.1. Memory in a Photograph

Consider a photograph on a piece of film. That film is a memory system in that its current state (the image in the photograph) gives us information concerning an interaction the film once had with something outside of it (namely, a set of photons).

For this system $t_1 < t_0$ and, as will be shown below, $J = P_{t_1}(S)$. There are a number of salient features of this memory system.

First, the film memory system is designed so that its state is relatively stable after the interaction. This allows us to not worry about how long it has been since the exposure of the film (i.e., since the interaction of the memory system with its outside world) when we try to deduce details of the external interaction from the memory system's current state.

Second, the possibility exists both of the film being exposed and therefore having a memory (i.e., of the memory system being open in the past) and of the film not being exposed and therefore not having a memory (i.e., of the memory system being closed in the past).

Third, the film memory system is designed so that it cannot have spurious memories of the outside world arising if there had in fact been no external interaction. This follows from the fact that we can always distinguish between a state reflecting a real external interaction and the state which arises if there had not been any such interaction with the outside world. (Without the ability to make this distinction, we would never be able to tell if a state of the system truly reflected an interaction with the outside world.) More precisely, film memory systems avoid the possibility of spurious memories by having the preinteraction state of the film, the initially unexposed black state of the film, be predetermined, stable, and distinguishable from the states the system is allowed to occupy if there had been an external interactions. The setting up of the film memory system in its predetermined preinteraction state will be referred to as the system's "initialization" (when observed, this initialized state of the memory system can be thought of as

[12]Nonetheless, it is conceivable (though highly unlikely) that the memory mechanism in the human brain is b-type memory, the third, particularly weak form of memory described in Appendix A. It is also conceivable (and a bit more likely) that although human memory is p-type, there are b-type "prior knowledge" aspects to how the distribution $P_{t_1}(S)$ is inferred (see Section 2.3). The important point for the psychological arrow of time is that, whether humans use p-type memory or use b-type memory, their memory must be temporally asymmetric, since both types of memory rely on the second law. (By the discussion at the end of Section 1 and in footnote 10, we know that human memory cannot be c-type, which is the only symmetric type of memory.)

corresponding to the reply "I do not know" in response to a query of that memory system). The film memory system is initialized when it is originally coated with a uniform layer of unexposed photosensitive material. This initialized state, whose existence is necessary if we are not to have spurious memories, corresponds to a peak in the distribution $P_{t_1}(S)$.

Finally, for us to have confidence in our initializer, we want it to be robust. We want it to be able to initialize the system regardless of the system's precise state prior to the initialization. (For example, we do not want to have to be concerned with the state of the components of the piece of film prior to its construction.) We want to be confident in our conclusion of what relationship exists between the memory system's current state and its past history. This confidence is only as strong as our confidence that the initialization took place regardless of the state of the system prior to initialization. Such confidence in the initialization process is what allows us to have no qualms about "resetting" a memory (i.e., reinitializing it) and using it again.

Summarizing, we want the state of the memory system to be stable when not interacting with the outside world; we do not want to preclude either the possibility of an external interaction or the possibility of no external interaction; we require the existence of a special initialized state which serves as a reference state signifying no external interaction; and we want to have confidence that the system was in that initialized state at t_1 whether or not we know anything about the system's state for times previous to t_1.

2.2. p-Type-Memory Systems

Although the preceding example was presented as memory of the past, we can easily generalize from it to get a time-symmetric definition of memory systems which work via the same logical mechanism as memory in photographs. This generalization, (2.1), defines p-type memory systems. It is presented and discussed in this subsection. We can hypothesize that not only photograph memory systems, but in fact *all* memory system which exploit $P_{t_1}(S)$ are necessarily p-type. Arguments establishing this hypothesis are presented in the subsection following this one.

(2.1) If a system S is in a state s_0 at a time t_0, the difference between s_0 and another state s_1 is said to be a *p-type memory*, be it of the past or of the future, and the system is said to be a *p-type memory system* if:

1. It is known that at a time $t_1 \neq t_0$ the system was in the state s_1.

2. For any moment in the temporal interval extending from t_1 to t_0, the state of the system is stable if it is not interacting with the external world. In particular, in the absence of interactions with the external world between times t_0 and t_1, s_0, the state of the system at t_0, would equal s_1.

3. With the possible exception of the precise value of s_1 provided, nothing precludes the memory system's being closed for the time period between t_0 and t_1, and nothing precludes the memory system's being open for that time period.

4. If $t_1 < t_0$, there exists a multiplicity of states the system could have been in for times previous to the process forcing the system to be in state s_1 at time t_1. If $t_1 > t_0$, there exists a multiplicity of states the system might be in for times following the process forcing the system to be in state s_1 at time t_1.

The rationale for requirement 3 is that *a priori* we want to allow both the possibility that a memory is stored in S and the possibility that S is "empty," containing no memory of its external world.

If $s_0 = s_1$, we will sometimes loosely say that we "have no memory" in the interval between t_1 and t_0. More usually, $s_0 \neq s_1$, and the difference between the two, being evidence of an external interaction between t_0 and t_1, constitutes the memory. Note that given only the information that $S(t_0) = s_0$, and not also that $S(t_1) = s_1$, one could not conclude that at some point in its history S must interact with W (this is because one could deterministically evolve $S(t_0) = s_0$ to any time t_1 using the assumption that S is closed, and not arrive at a contradiction). This is why it is crucial to meet requirement 1 of (2.1).

Requirement 2 of (2.1) is made for calculational convenience; it makes it easier to use the memory. In certain unusual circumstances it is possible to weaken it into the requirement that we know the precise evolution law for the memory system in the absence of external interactions. We will not consider such weakenings of requirement 2 in this paper.

The state-space collapsing down to the state s_1 at the time t_1 (i.e., the process of meeting requirement 1, a process alluded to in requirement 4) is referred to as the *initialization* of the memory, even if $t_1 > t_0$, so that the multiplicity of states being forced into s_1 occur *after* t_1 rather than before it.[13] Given just the knowledge of the current state of the memory system, knowledge of which state is the initialized one is not only necessary for us to conclude that there was (or will be) an external interaction, but also for us to infer any details of that external interaction. The reliability of the memory depends on our confidence in that knowledge of which state is the initialized one, i.e., it depends

[13] "Initialization" might seem a poor choice of word, since it carries temporally asymmetric connotations of a "before" and an "after." However, *all* human language carries such connotations to some degree, since human thought is time-asymmetric. The policy in this paper is to highlight all terms which are being treated in an explicitly time-symmetric manner in spite of their connotations in common human speech. In general, even for terms not so highlighted, it should be assumed that a time-symmetric meaning is being imposed unless the context makes it explicitly clear that this is not the case.

on our confidence that requirement 1 is met by the system's world-line. Requirement 4 of (2.1) reflects this desire of ours to have a high confidence that requirement 1 is met. It also reflects our desire that the initializer be able to function a number of times in succession, starting with differing initial conditions, and each time function with identical behavior as a memory system.

The multiplicity of states temporally adjacent to the initialized state are referred to as *pre-initialized* states, again, even if $t_1 > t_0$. (As always, despite the temporally asymmetric colloquial connotations of the terminology, here the definitions are made in an explicitly symmetric form.)

Note that no assumption has been made in (2.1) that $t_0 > t_1$. If $t_1 < t_0$, the initialization consists of a many-to-one mapping; if $t_0 < t_1$, it consists of a one-from-many mapping, the temporal inverse of a many-to-one mapping. (See Wolpert (1990) for a discussion of such mappings.) No implicit arrow of time has been assumed.

Note that c-type memory systems assume complete knowledge of the outside world. In contrast, p-type memory systems need no information about the outside world whatsoever to operate. Note also that there is no need in (2.1) for a requirement similar to requirement 4 in (1.1). This is because it is automatically and trivially true that either {changing the value of $S(t_0)$ while leaving $S(t_1)$ alone} or {changing the value of $S(t_1)$ while leaving $S(t_0)$ alone} will result in a different conclusion concerning the change in S caused by W. As a result, either modification will result in a different conclusion concerning $P_{t_1}(W)$.

In general, with p-type memory systems care must be taken in determining what is the memory system and what is the external world. For example, imagine that we have a plate which we know was once white but is now splattered with black ink. Clearly, we have a memory (of the past), and intuition might say that it is the plate which serves as the memory system. This is incorrect, however—the plate itself has not been altered in the slightest by its interaction with its outside world, and it provides no information from the past. The plate is simply now physically adjacent to some black ink. By itself, it gives no information concerning its past interaction with the ink (i.e., concerning the splattering). In point of fact, it is the ink which is the p-type memory system in this case, with its initialized state being "none of the ink is on the plate". It is the spatial configuration of the ink which has been changed by the interaction and which provides us with information from the past concerning the interaction.

2.3. Why Memory Systems Which Exploit $P_{t_1}(S)$ Are p-Type

This subsection presents information-theoretic and statistical mechanical arguments that all useful memory systems which rely on $P_{t_1}(S)$ are

p-type (i.e., must work via the same logical mechanism as a photograph).

Any memory system which is not either b-type or c-type must exploit information about S at times other than t_0. As was mentioned in the discussion of c-type memory systems, for simplicity we can have that "time other than t_0" be t_1. Moreover, Appendix B tells us that we want $P_{t_1}(S)$ to be infinitely peaked about some point s_1. This establishes the first requirement of (2.1), the definition of p-type memory systems.

Just as c-type memory systems, starting with $S(t_0)$ and $W(t_0)$, rely on knowledge of the dynamics of $S \times W$, p-type memory systems, starting only with $S(t_0)$, rely on knowledge of the dynamics in S. Since we are setting up the memory system, the dynamics in S is more or less under our control. To keep things as calculationally tractable as possible, we configure S space dynamics to preserve the S value from the end of the interaction all the way up to t_0. Similarly, we ensure that $S(t_1)$ is preserved in the case that there was not an interaction. This establishes the second requirement of (2.1).

Requirement 3 of (2.1) is axiomatic to all types of memory.

Therefore we only have yet to establish requirement 4 of (2.1). A fully rigorous exposition of the reasons for this requirement is beyond the scope of this (already lengthy) paper, and is in fact an extremely subtle issue. Here only the outline of such an exposition will be presented.

Essentially, requirement 4 follows from the need to meet requirement 1, i.e., our need to deduce that $\{S(t_1) = s_1\}$. To make any deductions about anything, we only have *direct* information from the present, t_0, at our disposal. In other words, we must deduce $\{S(t_1) = s_1\}$ using current information alone. However, by hypothesis we cannot conclude $\{S(t_1) = s_1\}$ solely from observing $\{S(t_0) = s_1\}$, since (in the terminology of the subsection of the Introduction on memory) we are here interested in memory systems relying on information $J \notin P_{t_0}(S)$. Therefore we need more current information than just $\{S(t_0) = s_0\}$ in order to meet requirement 1. Viewed another way, our goal is to be able to conclude that $\{S(t_1) = s_1\}$ *whether or not S interacts with W between* t_0 *and* t_1. However, knowledge of $S(t_0)$, by itself, is not sufficient to make this conclusion, since it cannot tell us if there is or is not such an interaction (and given only $S(t_0) = s_0$, the precise conclusion reached concerning the state of S at t_1 depends critically on this question of whether or not there is such an interaction).

Note that any scheme to deduce $\{s(t_1) = s_1\}$ is completely independent of the working of the memory; the memory is simply a device designed to facilitate using the provided (!) information $\{S(t_1) = s_1\}$, along with the other constraints of Hamiltonian dynamics and $\{S(t_0) = s_0\}$ and along with the procedure of MaxEnt, to set $P_{t_1}(W)$. Deducing the information that $\{S(t_1) = s_1\}$, like observing $P_{t_0}(W)$ in c-type memory systems, is a procedure not directly related to the ultimate use of that information in the memory.

In general, one can deduce $\{S(t_1) = s_1\}$ from current information either through *direct* inference or through *indirect* inference. As it turns out, both of these kinds of inference involve many-to-one mappings if $t_1 < t_0$ (one-from-many mappings if $t_1 > t_0$).

Indirect inference means we do not directly infer from current information that our particular system S was in the state s_1 at t_1. Rather, with indirect inference our information beyond $\{S(t_0) = s_0\}$ is the prior knowledge that other systems identical to ours are invariably in s_1 at t_1. Such indirect inference usually involves the examination of the behavior of many other systems of the same type as ours.[14] Note, however, that if indirect inference is used, we no longer have a purely p-type memory system, but rather have a memory system with some b-type aspects to it. Since a detailed investigation of b-type memory systems is beyond the scope of this paper, indirect inference will be considered no further, except to note that in practice it invariably necessitates requirement 4 of (2.1).[15]

For simplicity of the discussion, for the moment restrict the analysis so that S is a memory of the past. Then direct inference that our particular

[14]Note that, in general, this examination will itself involve memory; a full analysis of the indirect inference necessarily includes an analysis of that memory involved in the inference.

[15]Some simple observations can be made concerning the manifestation of b-type memory systems in indirect inference. First note that an indirect inference cannot be based simply on prior knowledge that systems identical to S are invariably in the state s_1. This is because such a correlation would not allow the systems to ever be pushed out of s_1, and would therefore violate requirement 3 of (2.1). This means that any indirect inference that $\{S(t_1) = s_1\}$ must be based, at least in part, on prior knowledge concerning how S is being used in concert with W as a memory system, i.e., on some aspect of (2.1). Very often such prior knowledge concerns the state information available to a p-type memory system: $S(t_0) = s_0$. However, if the indirect inference is based on a correlation that systems identical to ours are in s_1 at t_1 whenever they are in s_0 at t_0, then of necessity there is also the correlation that $W(t_1)$ is a state which forces s_1 at t_1 to s_0 at t_0. Moreover, all the full p-type memory system can tell us is precisely this information that $W(t_1)$ is a state which forces s_1 at t_1 to s_0 at t_0, and this information concerning $W(t_1)$ follows from (and is only as strong as) our prior knowledge of the correlation between $\{S(t_0) = s_0\}$ and $P_{t_1}(W)$. Therefore we could just as well skip the step of examining the correlation between $\{S(t_1) = s_1\}$ and $\{S(t_0) = s_0\}$ and examine the correlation between $\{S(t_0) = s_0\}$ and $P_{t_1}(W)$ directly. In other words, this kind of indirect inference can be viewed as dependent on $P_{t_0}(S)$ and independent of $P_{t_1}(S)$, i.e., as b-type memory. As a result, like b-type memory systems in general, this kind of indirect inference invariably involves state-space-collapsing mappings. In fact, *any* kind of indirect inference invariably involves such a mapping. As an example, consider the following special version of the photograph memory system: Suppose we have before as a completed photograph and that we have no *direct* evidence concerning the state of the film before it was a completed photograph. (For example, we never saw the film being put into the camera, never saw it being made, etc.) Then to use the photograph as a memory system we have to infer, from everyday experience of photography, that our particular piece of film was constructed in such a way that it was initially blank, no matter what the state of its constituent components before its construction. In other words, we have to (indirectly) infer a many-to-one mapping.

system is in state s_1 at t_1 means that there is an external system Z, not part of W, which interacted with S before the start of S's interaction with W, and such that the state of Z at t_1 allows us to infer that S was in state s_1 before its interaction with W. (For ease of analysis, it is implicitly assumed that we can deduce the state of Z at t_1 exactly from an observation of Z at the present, t_0.) Let the moment when Z lost contact with S be t_a and let the moment when Z and S came into contact be t_b; $t_b < t_a \leq t_1$. We want the known state of Z at t_1, z', to tell us that the state(s) S might occupy at t_b evolve into the unique state s_1 at t_1. (If S is a memory of the future, then all of this holds with $t_b > t_a \geq t_1$.) In effect, we are using Z as a b-type memory of $S(t_1)$.

Now in general, just given the observation that $Z = z'$ at t_1, we do not have enough information to specify the state of the joint system $Z \times S$ at t_b precisely. In other words, there is a multiplicity of states in $Z \times S$ at t_b all of which are consistent with the external constraint that $Z(t_1) = z'$, and therefore all of which are allowed (i.e., there is a multiplicity of states which get mapped by evolution in $\Gamma_{Z \times S}$ to states at t_1 whose Z space projection is z'). However, by hypothesis the dynamics of the joint system must force that multiplicity of states to have the S space projection s_1 at t_1. In other words, that dynamics must force the state of $Z \times S$ at t_1 to be unique, with the value (z', s_1). Therefore the dynamical evolution of the joint system $Z \times S$ must take a multiplicity of states to a single state; it must be many-to-one.[16]

If the memory is of the future, the conclusion is instead that the interaction between Z and S will be one-from-many. In either case, the need to "conclude that the system was (will be) in state s_1 at t_1" necessitates requirement 4 of (2.1). This is the last of the requirements of (2.1); we have established that memory systems relying on $S(t_1)$ in addition to $S(t_0)$ must meet (2.1) in full.

2.4. Comments

In Appendix A, the same kind of reasoning used in the preceding subsection to justify requirement 4 of (2.1) is used in concert with the central

[16]Since the mapping is many-to-one, by the central assumption of the theory of reversible computation, it increases entropy. Therefore, the entropy of the joint system $Z \times S$ does not obey Liouville, and we know that MaxEnt, as it is used with memory systems to infer $P_{t_1}(W)$ from $P_{t_0}(S)$ and J, cannot be directly applicable to the process of indirect inference. (There are a number of possible causes of this inapplicability. For example, the joint system $Z \times S$ might be connected to a heat bath. Or the constraints on $Z \times S$ might be of a different type from those used in the analysis of memory systems.) Indeed, if we *could* analyze $Z \times S$ in the same way we analyze $S \times W$, then we would simply maximize entropy over $Z \times S$ at t_1, subject to the single external constraint that $Z(t_1) = z'$. We would thereby reach the conclusion that ρ is constant across Γ_S. In other words, for such a case we could not conclude that $S(t_1) = s_1$ from the sole observation that $Z(t_1) = z'$. To make our direct inferences, none of the states $S(t_1) \neq s_1$ can be allowed.

assumption of the theory of reversible computation to make some general points about b-type memory systems. Taking our cue from Appendix A, we can apply the central assumption of the theory of reversible computation to requirement 4 of p-type memory systems. This assumption implies that the process of initialization results in higher entropy at t_1 than at t_b. This means that such a process violates the second law if $t_1 > t_0$. Therefore we can conclude that p-type memory systems must be of the past. Under the assumption that human memory is p-type, we have explained why humans can remember the past but not the future. We have shown explicitly that the psychological and thermodynamic arrows of time are correlated.

We can come to the same conclusion without directly invoking the central assumption. To do this, note that for direct inference the dynamics must be such that no state $(s \neq s_1, z')$ in $Z \times S$ at t_1 can exist, i.e., can be evolved from a state (s'', z'') in $Z \times S$ at t_b. (This is because if such a state *could* be evolved this way, then MaxEnt over $Z \times S$ would force us to have a nonzero ρ over that state.) This fact alone leads to the conclusion that evolution in $Z \times S$ must be irreversible. With the second law, this in turn means that $t_b < t_1$, and that p-type memory systems can only be of the past.

Note that the combined system $(Z \times S)$ can be viewed as the memory system of W; we are given the value of $(Z \times S)$ at the present, $(Z \times S)(t_0)$, and are using it to come to a deduction concerning $W(t_1)$. In a certain sense, $(Z \times S)$ can almost be viewed as a special type of b-type memory system, where the inference concerning $W(t_1)$ is made via $S(t_1)$, which in turn is made via $Z(t_1)$.

When analyzing memory systems, there are a number of useful techniques to keep in mind. One, exhibited in Appendix A for b-type memory systems, is interchanging S and W. Another is time-reversal. As an example of this second technique, consider again the process of directly inferring knowledge about $S(t_1)$. For $t_1 < t_0$, we are saying that the system Z interacted with S before t_1 and that Z's current state tells us something about the state of S after its interaction with Z. Fair enough; now consider the case where $t_0 < t_1$. For such a case we are positing a system Z which *will* interact with S and whose current state tells us something about the state of S *before* its future interaction with Z. This time-reversal of the process is clearly preposterous. However, the second law is the only time-asymmetric law of physics. Therefore, it is the only law upon which a process can be built such that the time-reversal of that process is illegal while the process itself is allowed. As a result, without any further analysis of the direct inference process and of why its time-reversed version is impossible, we can conclude that the second law must be involved somehow.

We can extend this kind of reasoning. Human memory is asymmetric. Assuming this is not due to a design flaw, it must be a consequence of the

second law. Therefore, independent of the analysis of this paper, entropy increase *must* be involved in human memory somehow.

3. HOW REAL-WORLD p-TYPE MEMORY SYSTEMS ACHIEVE INITIALIZATION

This section is an examination of how some naturally occurring p-type memory systems operate, and in particular it is an examination of how such systems are initialized and of how the second law is involved in that initialization. This examination is intended to serve as an intuitive complement to the (relatively) formal analysis of Section 2 and Wolpert (1990).

3.1. Internal Initialization

The state-space collapsing of the initialization of p-type memory systems can either occur with S closed or with S open. (For example, in the language of Section 2, for the case where $\{S(t_1) = s_1\}$ is inferred directly, one of the following two possibilities holds: all of the state-space collapsing of the joint system $Z \times S$ occurs at some time between t_b and t_a, while S is still coupled to Z, or some of it occurs when the two systems are decoupled at sometime between t_a and t_1.) Colloquially, we will say that initialization can be achieved either internally or via an interaction with an external system. (Such an initializing external interaction is not to be confused with the external interaction recorded in a memory.)

Whether achieved internally or externally, initialization entails collapsing all possible pre-initialized states, be they in the past or the future (corresponding to memory of the past or the future, respectively), to s_1, the initialized state. In other words, the initialized state serves as an attractor, of the past if remembering backward and of the future if remembering forward.[17]

The rest of this subsection is an examination of internal initialization; external initialization is treated in the next subsection. An example of a self-initializing memory system is the surface of the moon, given the information that sometime in the past the surface managed to relax and therefore became

[17]Note that the typical one-to-many mapping of a chaotic system is *not* an attractor of the future. It is not the temporal inverse of a many-to-one mapping. To see this, examine a one-to-many mapping going from a state s at time t to one of a multitude of states at $t' > t$. It is *not* true for such a mapping that regardless of the precise state s' at t', we can conclude that the system must have been in the state s at t. In fact, there are in general many possible pre-images of the mapping. But for s to be an attractor of the future, it must be the *only* pre-image of the mapping. This is exactly analogous to the necessity that a state be the only post-image of a mapping if it is to be an attractor of the past. For an elaboration of this distinction between the various kinds of mappings, see Wolpert (1990).

relatively smooth (and therefore, in particular, given the information that no lunar geological process exists which disturbs the relaxation of the lunar surfaces to smoothness). The ancient smoothing of the moon's surface is the initialization of requirement 4 of (2.1). Our knowledge of such a smoothing process in the past is how we meet requirement 1 of (2.1), and the fulfillment of the other requirements is obvious. Present-day craters on this surface contitute the system's memories of having interacted with external systems (i.e., meteors) sometime in the past, after the initialization to smoothness. Since such surfaces with craters have lower entropy than smooth, level (i.e., initialized) surfaces,[18] the second law allows us to conclude that they are evidence of an external interaction in the past. This is how they serve as memories of the past.

For a self-initializing system, in the real world (where $t_1 < t_0$), the attractor state s_1 is invariably the state of maximal entropy, as in the moon example. For such systems, s_0 and s_1 are usually correlated with the entropy of the system. In these systems it is always given that at the end of initialization the memory mechanism has been closed and relaxing long enough so that, due to the second law of thermodynamics, it is at maximal global entropy. The relaxing to the state of maximal entropy is the initialization of the memory system. Using the second law of thermodynamics, we can use these systems as backward memories. For example, if the memory system's states are indeed correlated with the entropy of the system, and if s_1 is the state of maximal entropy, then the second law tells us that if at anytime after initialization the system has low entropy, then there must have been an external interaction sometime in the past, after the initialization. Entropy would be high if the system had remained closed— since it is not high, the system must have been open. This evidence of an external interaction (e.g., a crater) is the memory. Note that the initializing process of these kinds of memory systems, being based on the second law, cannot work for $t_1 > t_0$; in our universe, self-initializers have $t_1 < t_0$.[19]

3.2. External Initialization

In general, when the initializing is external (so the system is open while collapsing to s_1), there is no need for the memory system to depend on

[18]Technically speaking, it is only the entropy of the memory system, i.e., of the matter making up the moon *before the impact*, which has shrunk. The meteor has added mass and therefore entropy to what we colloquially call "the moon"—this extra entropy must be subtracted off to get the change in entropy of the memory system by itself.

[19]It might be thought that strict time symmetry would allow anomalously *high* entropy to be used as a means of very accurately inferring the future in such self-initializing systems, just as *low* entropy allows inferring of the past. To do this, one would assume a system is initialized in the future with very low entropy, and then, if at a *previous* time the system has higher entropy, one would be able to conclude that the system was open at some moment between

differences between the entropies of s_0 and s_1. This is because the collapsing to an initial state, although increasing the entropy of the combined initializer–memory system, often does not result in maximal entropy of the memory system by itself. As a result, s_0 and s_1 can have the same entropy and be distinguishable, and yet s_1 can still be stable in time (requirement 2). An example of this kind of memory is a digital recording of some data, in an Electronically Erasable Programmable Red-Only Memory (EEPROM) computer chip, say. No two binary strings in the EEPROM differ in hardware-level entropy—all states are equally relaxed. Yet if we are given that the EEPROM was initialized to all 0's, for example, and that it now contains some 1's, we know that it has had an interaction with the outside world sometime since initialization. The meeting of requirements 1–4 is immediate and obvious. We have a non-entropy-based memory mechanism.

Another example of an externally initialized system is an ocean-side beach. The ocean, wind, and rain sweeping over the sand serve as the external initializer of the beach. The difference in state between such a swept beach and (for example) the same beach with a square etched on it is the evidence that the beach with the square interacted with an external system (i.e., a square-etcher) sometime since it was initialized. It is trivial to verify that this system for remembering square etchings meets all four requirements of (2.1). The photographic plate of film is yet another example of an externally initialized system. For this system, the initializing is done by the manufacturing process which produced the plate.

For such non-entropy-based memory systems, initialization does not occur through the assumption that the system has sufficiently relaxed to be in a defnite state at t_1. Instead it occurs through the assumption that an external system has interacted with the memory system, and left it in the definite state at t_1. Now, all such external initializing interactions will occur in some nonzero time interval Δt. And by requirements 1 and 3 of (2.1), we want to have the system fixed in its initialized state at the temporal end of Δt which is closest to t_0. Given the real world, this is enough to force $t_1 < t_0$; we can design initializing interactions which always *end* with the memory system in a definite, predetermined state regardless of its state at the beginning of the interaction, but we cannot design interactions which always *begin*

its initialization and the observation of its high entropy. Unfortunately, due to the second law, an external interaction is required anyway, between t_1 and t_0, just to meet requirement 1 of (2.1) and ensure the low-entropy initialization in the future. This violates requirement 3 of (2.1). We have a "memory" whether we want one or not. Entropy-based backward memory mechanisms as used in nature, on the other hand, actually use the second law to their advantage, initializing the memory without necessitating an external interaction between t_1 and t_0. Initialization using external interactions occurring outside of the window $[t_0, t_1]$, which also makes use of the second law, is considered below.

with the memory system in a definite, predetermined state regardless of its state at the end of the interaction. Therefore $t_1 < t_0$.

As one might suspect, the asymmetric fact that real-world external initializing interactions can end in a predetermined state but cannot begin in one has its base in the second law and in how real-world external initializing interactions make use of this law. Real-world interactions which initialize by leaving the memory system in a definite state s_1 work by coupling the memory system and the initializer into a composite system which then relaxes to a higher (joint) entropy. After this relaxing, the external initializer is removed. The state of the memory system corresponding to this maximal joint entropy is unique and is s_1, whereas there are many possible states of the memory system corresponding to a lower joint entropy (e.g., all states the memory system could have been in prior to the initializing interaction). Note that this initialization to s_1 was dependent on the second law of thermodynamics. As a result, this process can be used to end the initialization of a memory system in a definite state, but not to start it in one. Just as with self-initializers, for externally initialized memory systems it is maximizing entropy that serves to direct memory state space flow to a unique attractor state s', and thereby initialize the system. Therefore, just as with self-initializers, for externally initialized memory systems the moment of initialization must precede the moment when the memory is read.

3.3. Discussion

Note that it is always easier to convince someone that a system was once in one of *many* states than that it was once in *one particular* state. In their collapsing of the state space the initializations of p-type memory systems exploit this. Such initializations allow the observer to convince himself or herself only that the system was once in one of the many preinitialized states, all of which get mapped to the single state s_1. With mappings that collapse state space, it is not necessary to narrow things down to the state s_1 directly. (In the language of Section 2, with direct inference z' only needs to convey sufficient information to allow us to conclude that at t_b, $Z \times S$ is in one of the multitude of states which get mapped to $S = s_1$ at t_1. Similarly for indirect inference.)

As an example of how it is easier to convince someone that a system was once in a *volume* in state space rather than once at a particular *point* in state space, reconsider the self-initializing example of the moon's surface. To use the lunar surface as a p-type memory system, we only have to make the relatively easy inference that the surface of the moon was once quite hot (and as a result managed to relax gravitationally). Because this is such a

weak and imprecise inference, it is easy to gather independent evidence corroborating it. Now consider how things would differ if the initialization had not been through a many-to-one state-space collapsing process. In such a scenario, any lunar profile at all could serve as the initialized surface. Smoothness is irrelevant in this case, so for illustrative purposes pick any one particular craggy lunar profile as the initialized state. In this scenario, using the surface of the moon as a memory would necessitate inferring from current data that the surface of the moon had had the particular (initialized state) craggy profile at time t_1. (The lunar surface memory would then work by observing whether or not the current profile of the moon's surface is identical with this craggy initialized profile.) Such an inference of one particular craggy profile in the distant past is exceedingly difficult, to say the least. This kind of difficulty makes essentially useless any scheme which tries to infer $\{S(t_1) = s_1\}$ without the benefit of state-space collapsing. (The discussion of Section 2 goes further and argues that such a scheme is actually impossible.)

It has long been believed by many researchers that the asymmetry of memory in humans must follow from the thermodynamic arrow. However, it is only now, with the exposition of this section and of Section 2, that the connection between the two arrows of time can be seen in a detailed and formal manner. In particular, the analysis of this section implies that both externally and internally initialized p-type memory systems depend critically on the second law. It is not just that such memory mechanisms *obey* the second law, as argued by Hawking (1988), for example, and as implied by the analysis of Section 2. Rather, such memory mechanisms apparently *rely* on the second law—without a law like the second law of thermodynamics, there could be no p-type memory, of either direction in time. This is because p-type memory systems use the second law itself as the state-space-collapsing process (the second law takes a *multiplicity* of low-entropy states to a *single*-entropy state, namely $S_{\max}$—see Wolpert (1990)).[20]

[20]There are several interesting (but as yet unproven) conjectures which follow from the central importance of the second law to the working of p-type memory systems. For example, because of the inherent uncertainty in determining that $S(t_1) = s_1$, does it necessarily follow that no p-type memory system is 100% reliable? More generally, is it true that no memory system of any type, if it has no access to $W(t_0)$, can fix $P_{t_1}(W)$ to a delta function, whether $t_1 < t_0$ or vice versa? The conjecture is obviously true if $t_1 > t_0$. What is interesting is that it might hold even if $t_1 < t_0$. After all, even if one could envision a system in which one's confidence that $S(t_1) = s_1$ is as strong as one's confidence that entropy cannot decrease with time, since the second law says only that there is a very small—but nonzero—probability that entropy can shrink with time, there would be a very small—but nonzero—probability that $S(t_1) \neq s_1$. If this conjecture is true, then it would imply, for example, that one could never be 100% convinced that a particular photograph is an accurate depiction of a scene.

4. THE PSYCHOLOGICAL ARROW OF TIME AND ASYMMETRIC MEMORY

This section presents an argument that the psychological arrow of time is a consequence of the temporal asymmetry of human memory.

Just as there is no privileged position which we can call "here" which is always moving in a particular direction, there is no privileged time "now" which is always moving in a particular direction (toward the future). This is demanded by relativity's insistence on the equivalence of space and time. "Now", along with "here", is just a reference to a particular point on a world-line. All moments in time "exist"; there is no law of physics describing a passing of the baton of reality from one moment to the next, no "flow" of time[21]. No moment in time is picked out by the laws of physics as being special, just as no position is picked out by the laws of physics as being special.

The human mind does not perceive this, however. To humans, the world seems to have a privileged present, moving forward through time. Despite relativity, time, unlike space, appears to be asymmetric. This perceived temporal asymmetry is the famous psychological arrow of time.

To explain the psychological arrow, it is first necessary to relate it to a phenomenon which is amenable to mathematical analysis. Without such a mathematical handle, arguments will be over words rather than over physics. The "mathematically amenable" phenomenon most readily related to the psychological arrow is backward memory. It turns out that this phenomenon is sufficient to induce the psychological arrow.

To understand how asymmetric memory causes the psychological arrow of time, first note that every moment in time is, to the human mind occupying that moment, "now." There are an infinite number of these "nows," and information is transferred among them via one's backward memory. As a result, a human at one "now" can remember having thought at the "now" 2 seconds in the past, "What will I be observing in 2 seconds? Will I be observing what I think I will?" The human at the original "now" can also note what its present observations really are, thereby testing how predictions from the past stand up in the future. This allows mental processes to "go" from the now of 2 seconds ago to the current one, but not vice versa.

There are two important properties of this memory-aided prediction-testing. First, it can be applied to any moment in time (i.e., at every moment

[21] A "flow" of something is a derivative of a variable with respect to time. Since the two variables defining such a derivative have to be distinct (lest the "flow" have the fixed rate of 1), a "flow of time" can only be defined if there are two dimensions of time. Since there is no second dimension of time, as is implied by the term "the flow of time", "the flow of time" must be a completely meaningless term.

in time, one can remember the previous moments in time leading up to the current one). Second, all such prediction-testings have the same orientation: toward the future. As a result of these properties, the overlappings of all the prediction-testings result in an impression of time "moving forward." Due to backward memory, forward is the only direction mental processes can go.[22]

Note that no asymmetric assumption has been used in these arguments. If instead it were only the future of the world-line that could be remembered, time would appear to "move" into the past. Similarly, imagine that at all points on your world-line you could remember the events to your left on your world-line, but none of the events to your right. Imagine that in addition your world-line never doubled back on itself spatially (just as, due to your world-line's never exiting the light-cone, it never doubles back on itself temporally). Under these conditions, space would appear to you to "move" to the right.

All of the universe's world-line exists and is fixed, and the laws of physics are simply relations between different points on that world-line. There is no experimental evidence of a dynamic "evolving" of the universe, going from the past to the future, which transforms the universe from an undetermined to a determined state.[23] The subjective impression of such a dynamic evolving therefore must be an illusion caused by a "static" law relating the prefixed points on the world-line. And it is almost impossible to conceive of any objectively verifiable "static" law, other than asymmetric

[22]As an aside, note that this correlation of the subjective arrow of time with memory and therefore with the second law of thermodynamics throws a big monkey-wrench into the idea of time travel: unless one could somehow get around the second law, in going back in time one would lose all memory of the present. This inability to travel at will through time does not break the equivalence between time and space, however. This is because, strictly speaking, there is no such thing as "space travel" at will either, in the sense that due to the universe's being deterministic, we do not have the freedom to fix our future position in space any more than we can fix our past position in space. The reason we *think* we have the freedom to determine our own future position can be traced to our subjective impression that we have free will. This perception of free will, in turn, is part of the subjective impression of a psychological arrow of time, which (it is argued here) is simply a side effect of the asymmetry of our memory. According to the point of view of this paper, the perception of "free will" is nothing more than an illusion caused by this fact that to us the past (being remembered) appears fixed, whereas the future (not being remembered) appears free to be set. Hence *free* will.

[23]Although this analysis is presented in a classical framework, its conclusions still follow in a quantum mechanical framework. For example, Everett's (1957) famous many-worlds interpretation shows how quantum mechanics does not necessitate an "evolving" of the universe from an undetermined to a determined state. To Everett, the illusion of such an evolving in quantum mechanics is a consequence of the fact that we only have access to a subset of the entire universe (i.e., only a subset of all the worlds) rather than to the entire thing.

memory, which could account for an illusion of movement along one's world-line in a particular spatiotemporal direction.[24]

5. CONCLUSIONS

In this paper it is shown that the psychological arrow of time can be reduced to the temporal asymmetry of human memory. This reduction allows us to discuss the psychological arrow in a mathematical manner, since (the relevant aspects of) human memory systems can be defined rigorously, precisely, and time-symmetrically. Such a definition is presented here—a careful exploration of real-world memory systems serves to motivate this definition, while information-theoretic arguments serve to establish its logical necessity. An analysis of this definition shows that there are two types of memory system, each being appropriate for remembering in a different kind of universe. In particular, this paper argues that any memory system recording the external universe of humans, be it a memory system of the past or of the future, must rely on a process which collapses state-space flow to an attractor state. (This attractor state is the state of the memory at its initialization.) The central assumption of the theory of reversible computation tells us that in our universe such a collapsing process must involve the second law of thermodynamics (Wolpert, 1990). The temporal asymmetry of the second law thereby necessitates an asymmetry in the state-space collapsing process. This means that human-type memory itself is necessarily asymmetric in our universe, despite its time-symmetric definition. In this way this paper demonstrates, in a fully rigorous and precise manner, the correlation between the psychological arrow of time and the thermodynamic one.

This paper also discusses the relation between human memory systems and the other type of memory system, which is appropriate for highly constrained universes. This second kind of memory is best exemplified by the (abstract) RAM chips in a universe consisting of the rest of the (abstract) computer. Such (abstract) computer memory systems, which record only the

[24]The asymmetry of human memory is indisputable, as is the fact that this asymmetry alone would result in some kind of asymmetry in our mental processes. As a result, the burden of proof is actually on those who would dispute the explanation of the psychological arrow as being due to the asymmetry of memory. To dispute an explanation of an observation (the psychological arrow) when that explanation is based solely on an indisputable phenomenon (the asymmetry of our memory), it is necessary to some up with some aspect of the observation which is not accounted for by the indisputable phenomenon. In other words, unless one comes up with an objectively verifiable aspect of the psychological arrow which cannot be explained by the asymmetry of our memories, one cannot dispute the hypothesis that the asymmetry of our memories is the sole cause of the psychological arrow.

state of the computer (and not the state of the world outside the computer), do not require initialization and therefore do not require any state-space collapsing. As a result, in accord with the theory of reversible computation, such computers can remember their future as easily as their past. This is true even if the program running on the computer is not logically reversible. Unfortunately, computer-type memory systems are feasible only when severe restrictions exist concerning the universe external to the memory (i.e., external to the RAM chips). It is due to these restrictions that time-symmetric computer-type memory systems cannot be used by humans, for whom the external world is the entire physical universe.

It should be noted that this paper does *not* address the question of why the second law of thermodynamics holds, given that the microscopic laws of the universe are time-symmetric. (This is a question which some think cannot be answered at all, using present-day physics (Penrose, 1989).) In this paper the second law is just taken as a given. A careful analysis of the problem of why the second law holds and a proposed solution to it is the subject of another paper (Wolpert, 1989).

APPENDIX A. DISCUSSION OF b-TYPE MEMORY SYSTEMS

This appendix discusses b-type memory systems, those memory systems which are provided with no information other than $P_{t_0}(S)$.

Suppose there is a cave some of whose stones have been arranged so that they spell out a four-letter word. In such a situation we would conclude that it was likely that a human had entered the cave in the past to arrange the stones into the word. The state of the interior of the cave is S, and the human entering it is the interaction with an external system, W. This is a b-type memory system—we have no *a priori* information concerning the state of any system other than $\{S(t_0) =$ "stones arranged into the shape of a four-letter word"$\}$.

How do we conclude from $\{S(t_0) =$ "stones arranged into the shape of a four-letter word"$\}$ that $\{W(t_1)$ involved a human$\}$? Essentially, a Bayesian or maximum-likelihood analysis is used. If $P_{t_1}(W)$ is not particularly peaked about the state "human coming upon the cave," then the word in stones had to have formed through natural causes (i.e., through the particular deposition history of the rocks in the cave) and the probability of a word in stones at t_0 is low. On the other hand, if $P_{t_1}(W)$ *is* particularly peaked about the state "human coming upon the cave," then the probability of a word in stones at t_0 is quite substantial. Therefore it is likely that $P_{t_1}(W)$ is peaked about the state "human coming upon the cave." Another example of this type of memory is a beaver pond, $P_{t_0}(S)$ peaked about the state "pond has

a dam" being evidence that a beaver entered the pond in the past ($P_{t_1}(W)$ peaked about the state "beaver comes upon the pond").

Note that b-type memories can be of the future as well as of the past. Just interchange S and W. For example, seeing a beaver come upon a pond is a "memory" that the pond is likely to have a dam in the future. In practice, however, we will often want the memory to be the space which gets affected by the interaction—the external world should be barely perturbed at all by the process of being observed and memorized. As an extreme case, we can require the interaction to be such that $W(t_1) \to W(t_0)$ is a single-valued mapping, independent of $S(t_1)$ and $S(t_0)$. This is enough to force b-type memories to be of the past, as can be seen by the following (sketched) proof based on the formalism of Wolpert (1990).

Define $s_0 \equiv S(t_0)$. We will assume that the memory works perfectly, i.e., only one W state at t_1, w_1, is consistent with $S(t_0)$ being s_0. Now in general, having only observed $S(t_0) = s_0$, *any* state $S(t_1)$ is possible—there is always some external system state at t_1 which will force a particular state $S(t_1)$ to evolve into s_0. (For example, the state {pattern of stones spelling out a four-letter word} by itself sets no restrictions whatsoever on the state of those stones before the human entered the cave.) However, by our assumption that the memory works perfectly, only $W(t_1) = w_1$ is consistent with $S(t_0) = s_0$, and all these allowed $S(t_1)$ values correspond to the same value w_1 of $W(t_1)$. (See Wolpert (1990) and note that in general a given value of W corresponds to more than one point in the associated phase space Γ_W.) Due to our assumption that $W(t_1) \to W(t_0)$ is a single-valued mapping independent of $S(t_1)$ and $S(t_0)$, this means that there is only one value w_0 of $W(t_0)$ consistent with $S(t_0) = s_0$. Therefore we have a mapping taking a multiplicity of states in $S \times W$ at time t_1 to a single state in $S \times W$ at time t_0. In the language of Wolpert (1990), the evolution in $S \times W$ is many-to-one if $t_1 < t_0$, and one-from-many if $t_0 < t_1$. However, the central theorem of reversible computation tells us that this means the entropy at t_0 must exceed the entropy at t_1. (For example, the entropy of the total combined system of human/cave after the human has rearranged the stones exceeds the value the entropy had before the human entered the cave—the entropy gain in the human part of the combined system more than offsets the entropy loss in the position of the stones.) Therefore, by the second law of thermodynamics, t_1 must precede t_0, and such b-type memories where W is not perturbed by the interaction must be of the past and cannot be of the future. QED.

In practice, b-type memories are not particularly useful because they usually do not tell us very much about $W(t_1)$. In addition, they usually work only for a severely limited subset of the space S_{t_0}. (For example, very few of the possible patterns of stones in a cave would lead us to conclude that the cave interacted with something outside of it in the past.) As a result, b-type memory systems will not be analyzed here in any depth.

APPENDIX B. PROOF THAT A DELTA FUNCTION DISTRIBUTION OVER x IS A NECESSARY CONDITION FOR MAXIMAL INFORMATION OVER $W(t_1)$

The variable x referred to at the end of Section 1.3 is an element of a partition on a phase space. We assume that that partition is fine enough so that we can approximate a distribution over x with the same distribution over the phase space partitioned by x. We will prove that information is maximized for the case where x is $S(t_1)$ if the distribution over x is a delta function. The proof when x is $W(t_0)$ is trivial; if both S and W are known exactly at t_0, then independent of any issues of maximizing entropy, we can evolve the combined system $S \times W$ deterministically from t_0 to t_1. Therefore this distribution over $W(t_0)$ gives us maximal information about $W(t_1)$ (we know $W(t_1)$ exactly, in fact).

We start with a provided distribution over $x = (\Gamma_S)_{t_1}$. Given this distribution, we must find the maximal entropy over $(\Gamma_{S\times W})_{t_1}$ subject to the constraint that $P_{t_0}(S)$ is a delta function and subject to the constraint of the provided distribution over x. The goal is to choose a distribution over x such that the resultant distribution over $(\Gamma_{S\times W})_{t_1}$ has maximal information. (More precisely, we are interested in maximal information over the projection space $(\Gamma_W)_{t_1}$.) In other words, we must find the distribution over x such that {the maximum possible associated entropy} is minimized.

As was pointed out in the text, the first constraint ($P_{t_0}(S)$ is a delta function) is simply a constraint on the allowed region in $(\Gamma_{S\times W})_{t_1}$; it is a boundary, nothing more. Given the second constraint (the provided distribution over x), the process of extremizing the entropy becomes the following: extremize an integral of the form $\int \rho(\Gamma_{A\times B}) \ln[\rho(\Gamma_{A\times B})]\, d\Gamma_{A\times B}$, subject to a constraint of the form $\int \rho(\Gamma_{A\times B})\, d\Gamma_A = P(B)$, where the integrals are only over the allowed regions of the associated spaces. For simplicity we will rewrite this as extremizing $\int \rho(a, b) \ln[\rho(a, b)]\, da\, db$ subject to $\int \rho(a, b)\, da = P(b)$. Here $P(b)$ is the externally provided distribution over Γ_B, the phase space associated with x. Note that the normalization constraint $\int \rho(a, b)\, da\, db = 1$ is automatically taken care of by the provided distribution $P(b)$, since $\int \rho(a, b)\, da\, db = \int P(b)\, db$, and we assume that the provided $P(b)$ obeys $\int P(b)\, db = 1$. For simplicity, from now on we treat A and B as though both were $\mathbf{R}^1$; similar arguments work for higher dimensional spaces.

Using Lagrange multipliers in the usual way, we conclude that for a given B-space value b and an associated allowed range in A space, over that entire allowed range in A space, $\rho(a, b)$ is constant and independent of the A-space coordinate a. In fact, $\rho(a, b) = P(b)/|\sigma(b)|$, where $\sigma(b)$ is the allowed range in A corresponding to B value b, and $|\sigma(b)|$ is the size of that

interval. $\sigma(b)$ is determined by the boundary of the allowed region in $(\Gamma_{S\times W})_{t_1}$, which in turn is set by the first constraint that $S(t_0)$ is a delta function. For simplicity, we are assuming that the allowed region in $(\Gamma_{S\times W})_{t_1}$ is simple (and therefore simply connected) and that its border is nondifferentiable at, at most a finite number of points. (Note that this does *not* mean we are assuming that the support of $\rho(a, b)$ is a simply connected set; the topological structure of the support of $\rho(a, b)$ depends on the topological structure of the support of $P(b)$.)

The probability distribution over A space is given by the formula

$$P(a) = \int_{L(a)}^{H(a)} \rho(a, b)\, db = \int_{L(a)}^{H(a)} P(b)/\sigma(b)\, db$$

$L(a)$ is the lowest b value in the allowed region for A-space value a, and $H(a)$ is the highest such value. The only function under our control is $P(b)$. We want to choose the $P(b)$ which maximizes $\int P(a) \ln[P(a)]\, da$ (subject to the constraint that $P(b)$ is normalized). This $P(b)$ gives the distribution across x which maximizes our information about a, i.e., which maximizes our information about $W(t_1)$. We will show that given any distribution $P(a)$ corresponding to a non-delta function $P(b)$, there is another distribution over A space whose support is a subset of the support of $P(a)$, which corresponds to a delta function $P(b)$, and whose information over A space is higher than that of the original $P(a)$. In this way we will show that one should always use a delta function $P(x)$.

If $P(b)$ is a delta function, $\delta(b-\beta)$, then $\rho(a, b)$ has as support only a line segment going in (A, B) space from (the bottom of $\sigma(\beta)$, β) to (the top of $\sigma(\beta)$, β). The value of $\rho(a, b)$ is constant across that line segment: $\rho(a, b) = \text{step}(a, \sigma(\beta)) \times \delta(b-\beta)$, where $\text{step}(a, \sigma(b)) \equiv \{1/|\sigma(b)|$ if a is in the region $\sigma(b)$, 0 otherwise$\}$. Therefore for this case $P(a)$ is zero everywhere except across an interval where it is constant; $P(a) = \text{step}(a, \sigma(\beta))$. For this case the derivative of $P(a)$ is 0 everywhere except at the two ends of that interval.

For simplicity, we will now assume that the support of $P(a)$ is not a disconnected set. (Arguments similar to the following hold for the case where the support of $P(a)$ is disconnected.) We will prove that a delta function $P(b)$ gives maximal information first for the case where $P(a)$ is constant across its support, and then for the case where it is not.

If $P(a)$ is constant across its support but $P(b)$ is not a delta function, then we can consider replacing $P(b)$ with a new delta function $P(b)$ whose support is a subset of the support of the original $P(b)$. The support of the resulting new $P(a)$ will be a subset of the support of the original $P(a)$. Moreover, since the new $P(b)$ is a delta function, the new $P(a)$ will be constant across its support, just like the original $P(a)$. As a result, normalization tells us that the new $P(a)$ is everywhere greater than or equal to the

original $P(a)$. This means that the entropy of the new $P(a)$ is less than the entropy of the original $P(a)$. This completes the proof that if $P(a)$ is constant across its support but $P(b)$ is not a delta function, then replacing $P(b)$ with a delta function decrease the entropy of $P(a)$.

Now consider the case where $P(a)$ is not constant across its support. (In general, allowing $P(a)$ to be nonconstant across its support means that $\rho(a, b)$ need not be constant across *its* support, although we do know that *for a fixed* b, $\rho(a, b)$ is constant everywhere it is not zero.) Define $a' \equiv \arg\max(P(a))$. Define b' as any one of the points $b \in B$ with minimal value $|\sigma(b)|$ such that $\sigma(b)$ encloses the A-space value a'. Divide up the entire B axis into intervals of width 2ε, where ε is arbitrarily small, so we can assume that $P(b)$ is constant over any such interval. Now raise $P(b)$ over the interval $b \in [b' - \varepsilon, b' + \varepsilon]$ by a constant ratio k', while uniformly diminishing all $P(b \notin [b' - \varepsilon, b' + \varepsilon])$ by a constant ratio k, where k is chosen to maintain the normalization of $P(b)$. This will be referred to as a "delta-ing procedure."

Lemma B.1. $P(a')$ is raised by the delta-ing procedure.

Proof. First assume that b' is in fact the only $b \in B$ such that $\sigma(b)$ encloses the A-space value a', regardless of the size of $|\sigma(b)|$. Then $L(a') = b' - \varepsilon$, $H(a') = b' + \varepsilon$. Since the delta-ing procedure raises $\rho(a, b)$ over this range $[L(a'), H(a')]$, it raises $P(a')$, i.e., it raises the maximum of $P(a)$. If there are $b \in B$ besides b' whose $\sigma(b)$ enclose the A-space value a', then either $L(a') < b' - \varepsilon$, and/or $H(a') > b' + \varepsilon$. We now must prove that for this scenario as well the delta-ing procedure has raised the maximum of $P(a)$. Let N be the number of width-2ε B-space intervals, aside from the interval $[b' - \varepsilon, b' + \varepsilon]$, whose corresponding allowed interval in A space enclose the A-space value a'. Delineate the N such $P(b)$ values as $P(b_i)$, $1 \leq i \leq N$. Delineate by $P(b_j)$, $1 \leq j \leq M$, $M \geq N$, the set of *all* $P(b)$ (other than b') which do not equal 0. We have increased $P(b')$ by the ratio k', and we have decreased all the $P(b_j)$ by the ratio k. Therefore normalization says that $k'P(b') = 1 - k \sum_j P(b_j)$. Now it is always true that

$$P(a') = P(b')/|\sigma(b')| + \sum_i \{P(b_i)/|\sigma(b_i)|\}$$

so the new

$$\begin{aligned} P(a') &= k'P(b')/|\sigma(b')| + k \sum_i \{P(b_i)/|\sigma(b_i)|\} \\ &= \left\{1 - k \sum_j P(b_j)\right\} \Big/ |\sigma(b')| + k \sum_i \{P(b_i)/|\sigma(b_i)|\} \\ &= 1/|\sigma(b')| + k \left[\sum_i \{P(b_i)/|\sigma(b_i)|\} - \sum_j \{P(b_j)/|\sigma(b')|\}\right] \end{aligned}$$

The old

$$P(a') = 1/|\sigma(b')| + \left[\sum_i \{P(b_i)/|\sigma(b_i)|\} - \sum_j \{P(b_j)/|\sigma(b')|\}\right]$$

If we subtract the old $P(a')$ from the new one, we get

$$(k-1)\left[\sum_i \{P(b_i)/|\sigma(b_i)|\} - \sum_j \{P(b_j)/|\sigma(b')|\}\right]$$

By definition of b', $|\sigma(b')| \leq |\sigma(b_i)|$ for all N of the b_i. Therefore, the quantity inside the brackets is negative. Furthermore, $k < 1$. Therefore the change in $P(a')$ is positive, proving the proposition. QED.

If we keep iterating the delta-ing procedure, we get a $P(b)$ arbitrarily close to a delta function. This means that the resultant $P(a)$ is constant over its support, and by (B.1), we know that the magnitude of this new $P(a)$ over its support is greater than the maximum magnitude of the original $P(a)$. To finish the proof that $P(b)$ should be a delta function, we only have to show that the growth of the maximum of $P(a)$ means that the entropy of the new $P(a)$ is less than the entropy of the original $P(a)$. To do this, we will find a lower bound on the entropy of the original $P(a)$, given that the maximum of that $P(a)$ is $p \equiv P(a')$, and then show that this lower bound is greater than the entropy of the new $P(a)$.

Lemma B.2. The minimal possible entropy of $P(a)$ subject to the constraint that $P(a) \leq p \; \forall a \in A$ occurs when $P(a) =$ either p or 0 for all a.

Proof. Imagine that $P(a)$ has the value $p' \notin \{0, p\}$ in some arbitrarily small interval $[x, y]$. Then without violating normalization of $P(a)$, we can modify $P(a)$ so that it has the value p over the interval $[x, x + (y-x)p'/p)$, and is 0 over $[x + (y-x)p'/p, y]$. The change in entropy accompanying this modification is

$$-[\{(y-x)p'/p\}p \ln[p] - \{y-x\}p' \ln[p']] = -[\{(y-x)p'\}\{\ln[p] - \ln[p']\}]$$

Since $p > p'$, this quantity is negative. Therefore the original $P(a)$ did not have minimal entropy. QED.

(B.2) implies that a lower bound on the entropy of the original $P(a)$ is $-zp \ln[p]$, where z is the volume of the support of the new $P(a)$ (i.e., $1/p$); the lower bound $= -\ln[p]$. Similarly, the $P(a)$ corresponding to a delta function $P(b)$ is constant over its support, and therefore has entropy $-\ln[p'']$ (p'' is the maximum of the new $P(a)$ corresponding to the delta function $P(b)$). We know that $p'' > p$, so the entropy of $P(a)$ corresponding to the delta function $P(b)$ is less than the entropy of the original $P(a)$ corresponding

to a non-delta function $P(b)$. This proves that $P(x)$ should be a delta function to get maximal information about $W(t_1)$.

ACKNOWLEDGMENTS

I would like to thank N. Rosenberg, P. Stolorz, and W. Fontana for helpful comments. Some of this work was done under the auspices of the Department of Energy.

REFERENCES

Bennett, C. H. (1982). *International Journal of Theoretical Physics*, **21**, 905.
Bennett, C. H. (1988). *IBM Journal of Research and Development*, **19**(1), 16–24.
Bitbol, M. (1988). *Philosophy of Science*, **55**, 349–375.
Davies, P. C. W. (1974). *The Physics of Time Asymmetry*, University of California Press, Berkeley, California.
Everett III, H. (1957). *Review of Modern Physics*, **29**, 454–462.
Fredkin, E., and Toffoli, T. (1982). *International Jorunal of Theoretical Physics*, **21**, 219.
Gold, T. (ed.) (1967). *The Nature of Time*, Cornell University Press, Ithaca, New York.
Hawking, S. W. (1988). Cambridge University Report, 2/88 (no report number).
Jaynes, E. T. (1957*a*). *Physical Review*, **106**, 620–630.
Jaynes, E. T. (1957*b*). *Physical Review*, **108**, 171–190.
Jaynes, E. T. (1982). *Proceedings of the IEEE*, **70**, 939–952.
Landauer, R. (1961). *IBM Journal of Research and Development*, **5**, 183.
Landauer, R. (1985). *Annals of New York Academy of Sciences*, **426**, 2.1.
Layzer, D. (1976). *Astrophysical Journal*, **206**, 559–569.
Pearl, J., and Crolotte, A. (1980). *IEEE Transactions on Information Theory*, **26**, 633.
Penrose, R. (1989). *The Emperor's New Mind*, Oxford University Press, Oxford.
Popper, K. (1965). *Nature*, **207**, 233–234.
Skilling, J. (1989*a*). Classic maximum entropy, in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed., Kluwer, Dordrecht, Holland.
Skilling, J. (ed.) (1989*b*). *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, Holland.
Smith, C. R., and Erickson, G. (1989). From rationality and consistency to Bayesian probability, in *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, Holland.
Wheeler, J. A., and Zurek, W. H. (1983). *Quantum Theory and Measurement*, Princeton University Press, Princeton, New Jersey, pp. 785–786.
Wolpert, D. (1988). *Nature*, **335**, 595.
Wolpert, D. (1989). A consistent statistics for proving the second law of thermodynamics, Los Alamos National Laboratory Report LA-UR-90-954.
Wolpert, D. H. (1990). The relationship between logical erasure and thermodynamic irreversibility, Loss Alamos National Laboratory report LA-UR-90-4108.