THE JOURNAL OF PHILOSOPHY

VOLUME CVII, NUMBER 8
AUGUST 2010

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

All communications to the Editors and Trustees and all manuscripts may be sent to Alyssa Timin, Managing Editor, at submissions@journalofphilosophy.org.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2010

SUBSCRIPTIONS (12 issues)

Individuals	$45.00
Libraries and Institutions	$100.00
Libraries and Institutions (print and online)	$180.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset by Dartmouth Journal Services, Waterbury, VT, and printed by The Sheridan Press, Hanover, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The Journal of Philosophy (ISSN 0022-362X) is published monthly by The Journal of Philosophy, Inc., MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027. Periodicals postage paid at New York, NY, and other mailing offices.

POSTMASTER: Please send address changes to *The Journal of Philosophy* at MC4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME CVII, NO. 8, AUGUST 2010

MAXWELL'S DEMON*

This paper is dedicated to the memory of Itamar Pitowsky.

> [Classical thermodynamics] is the only theory of universal content concerning which I am convinced that, within the framework of the applicability of its basic concepts, it will never be overthrown.
>
> —Albert Einstein[1]

Einstein's opinion quoted above expresses, more or less, the prevalent view about thermodynamics. Maxwell, however, thought otherwise. Maxwell devised his famous thought experiment known as *Maxwell's Demon* in the setting of classical mechanics as a counterexample of the second law of thermodynamics.[2] He realized that a truly mechanistic worldview has consequences that are incompatible with thermodynamics, and that such a worldview means that there is no "framework of applicability" (to use Einstein's expression) which is not subject to the laws of mechanics. By this he expressed a view which seems to counter Einstein's. Since at his time the theoretical tools needed to derive this insight from the principles of mechanics were not available, Maxwell framed his view by appealing to his picturesque thought experiment of the Demon. Since Maxwell, writers agreeing with Einstein have made numerous attempts to counter his argument.[3] Most of these attempts have focused on the

*We thank David Albert, Dan Drai, Tim Maudlin, and especially Itamar Pitowsky for very helpful comments. This research is supported by the Israel Science Foundation, grant number 240/06.

[1] Albert Einstein, "Autobiographical Notes," in Paul Arthur Schilpp, ed., *Albert Einstein: Philosopher-Scientist* (La Salle, IL: Open Court, 1970), p. 33.

[2] James Clerk Maxwell to P. G. Tait, 1868, in Cargill Gilston Knott, *Life and Scientific Work of Peter Guthrie Tait* (Cambridge: University Press, 1911), pp. 213–14.

[3] See various attempts and discussion in Harvey S. Leff and Andrew F. Rex, eds., *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing* (Philadelphia: Institute of Physics Publishing, 2003).

0022-362X/10/0708/389–411

construction of various devices, and the rejection of Maxwell's idea was based on the details of these devices. We believe that focusing on these details obscured the heart of the matter. Regardless, during these hundred and fifty years or so no general proof or disproof of Maxwell's idea has settled the issue.

Ten years ago, however, David Albert gave a general argument that a Maxwellian Demon is compatible with the principles of mechanics, thus supporting Maxwell.[4] Our discussion in this paper follows and extends Albert's argument in the most general terms, and refrains from examining particular devices. We will argue that a Maxwellian Demon is compatible not only with the principles of mechanics, but also with the principles of statistical mechanics.

The question of Maxwell's Demon raises and illustrates several important philosophical issues about the project of statistical mechanics in general. If we take seriously the idea that the world can be described completely by a mechanical theory (classical or quantum), then there must be an explanation of our experience and our statistical mechanical probabilistic considerations on the basis of the laws of mechanics. In this paper, we propose a schematic statistical mechanical account of the way in which our experience of thermodynamic phenomena arises in the framework of classical mechanics. We show that this account is consistent with a Maxwellian Demon.

Whether Maxwellian Demons can be constructed in the world is a question of fact which cannot be settled by turning to the laws of statistical mechanics. The reason, as we will show, is that the laws of statistical mechanics are consistent both with worlds in which there are Maxwellian Demons and with worlds in which there are no Maxwellian Demons and the second law of thermodynamics is true. Whether Demons are possible in our world depends on the details of the dynamics and the initial conditions of our world. It may be that

[4] Albert, *Time and Chance* (Cambridge: Harvard, 2000), chapter 5. Albert's argument is formulated in the framework of Boltzmann's approach to statistical mechanics. Defending this approach is beyond the scope of this paper. Some arguments are given in *ibid.*; Sheldon Goldstein, "Boltzmann's Approach to Statistical Mechanics," in Jean Bricmont et al., eds., *Chance in Physics: Foundations and Perspectives* (New York: Springer, 2001); Craig Callender, "Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy," this JOURNAL, XCVI, 7 (July 1999): 348–73; Joel L. Lebowitz, "Statistical Mechanics: A Selective Review of Two Central Issues," *Review of Modern Physics*, LXXI, 2 (1999): S346–57. For a critical historical introduction to Boltzmann's work and references, see Jos Uffink, "Boltzmann's Work in Statistical Physics," *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Winter 2008). URL: http://plato.stanford.edu/entries/statphys-Boltzmann/.

The foregoing Maxwellian demon has no counterpart in the Gibbsian framework to statistical mechanics since the argument hinges on the Boltzmannian notion of entropy as given by the phase-space volume (Lebesgue measure) of a *macrostate*.

the dynamics and the initial conditions of our world are of the kind described by, for example, Lanford's theorem. In that case, our world will behave thermodynamically in the way spelled out by such a theorem. Although Lanford's and similar theorems require specific conditions, it is important to realize the significance of theorems of this kind. Theorems of this kind have a conditional form: *If* the world is Lanford-like, that is, if the world satisfies the conditions spelled out in the theorem, *then* it is also thermodynamic-like. Therefore, even if such general theorems were proven, we could not conclude that our world is thermodynamic-like without knowing that the antecedent of this conditional is true of our world.

In this paper, we will demonstrate that in classical statistical mechanics there are initial conditions and Hamiltonians that give rise to Maxwellian Demons. Whether or not these initial conditions and Hamiltonians are true of our world, or can be realized in laboratories, is an open question. In this sense, the Demon is a consequence of taking mechanics seriously. Another consequence of our analysis of the question of Maxwell's Demon applies to the entropy cost of information processing. We show that the Landauer-Bennett thesis concerning this cost is false.

The paper is structured as follows. In section I, we show that Maxwell's Demon is *compatible* with the principles of statistical mechanics. In section II, we discuss some restrictions on the *efficiency* of the Demon. These restrictions do not rule out the Demon as physically impossible. In section III, we show that the Demon's cycle of operation can be completed (in particular, the Demon's memory can be erased) *without* increasing the total entropy of the universe. We take this to refute the Landauer-Bennett thesis, according to which erasure of information is necessarily accompanied by a certain minimum amount of entropy increase.[5] In section IV, we draw some conclusions from our analysis.

I. MAXWELL'S DEMON

I.1. Setting the Stage. The Demon in Maxwell's original thought experiment decreases the entropy of the gas by separating particles according to their speed. The Demon manipulates the gate between two chambers, thereby allowing the fast particles to enter one chamber

[5] See Rolf Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM Journal of Research and Development*, V, 3 (1961): 183–91; Charles H. Bennett, "The Thermodynamics of Computation—A Review," in Leff and Rex, eds., *op. cit.*, pp. 283–318; Bennett, "Notes on Landauer's Principle, Reversible Computation, and Maxwell's Demon," *Studies in History and Philosophy of Modern Physics*, XXXIV, 3 (September 2003): 501–10.

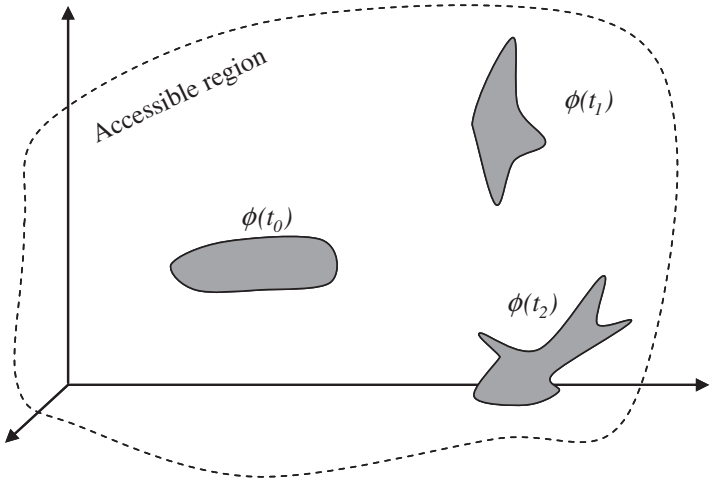


Figure 1: Dynamics. $\phi(t_0)$ is the dynamical blob at the initial time t_0 and the trajectories that start out in it evolve by the equations of motion to the regions $\phi(t_1)$ at t_1 and later $\phi(t_2)$ at t_2. The volume of the blob $\phi(t)$ is conserved at all times, according to Liouville's theorem. The dynamical evolution is restricted to the accessible region.

and the slow ones to enter the other. Consequently, whereas initially the states of the particles are distributed according to the Maxwell-Boltzmann energy distribution, the final distribution is different. By this thought experiment Maxwell captured an intuition which we shall now explain in general terms.

The state of a classical mechanical system is represented by a point in the system's phase space Γ at a given time. Γ contains a subspace consisting of *all* the microstates that are consistent with external constraints, which may include boundary conditions such as volume and limitations such as total energy (see Figure 1). This is the system's *accessible region*.[6] Some of the constraints may change with time, but the actual state of the system at any given time is necessarily confined to the region which is accessible to it at that time.

The time evolution of the system is given by a trajectory in phase space which is a continuous sequence of points obeying the classical

[6] If the dynamics is such that the region accessible to the system is metrically decomposable into dynamically disjoint regions each with positive measure (as in KAM's theorem; see Grayson H. Walker and Joseph Ford, "Amplitude Instability and Ergodic Behavior for Conservative Nonlinear Oscillator Systems," *Physical Review*, CLXXXVIII, 1 (Dec. 5, 1969): 416–32), we can consider the effectively accessible region as determined by the system's initial state.

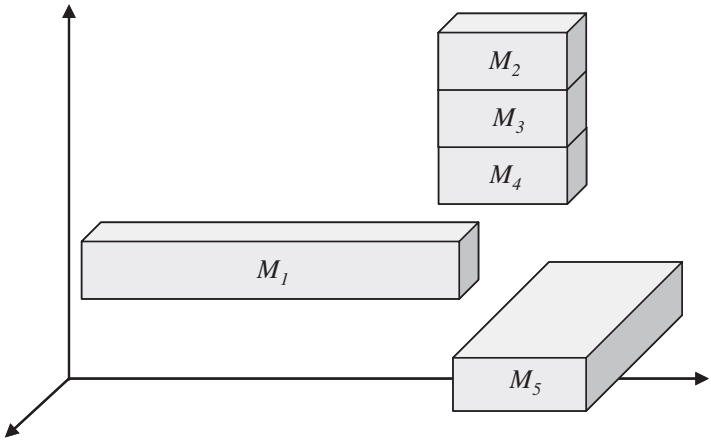


Figure 2: Macrostates. The whole phase space (and in particular the accessible region) is partitioned into macrostates, some of which are $M_1 \ldots M_5$.

equations of motion. A useful tool in mechanics is to consider the time evolution of a set of points (call them *dynamical blobs*) $\phi(t)$ corresponding to various *possible* microstates of the system at a given time t. The time evolution of these points is given by a bundle of trajectories. By Liouville's theorem the Lebesgue measure of $\phi(t)$ is conserved under the dynamics, although its shape may change over time (see Figure 1).

Γ is also partitioned into subregions which form the set of *macrostates* (see Figure 2).[7] Macrostates correspond to the values of some classical macroscopic *observables*. By this term we mean sets of microstates each of which forms an equivalence group which reflects measurement or resolution capabilities of some observer (human or other). A system is said to be in a given macrostate at time t if its actual

[7] This idea is expressed, for example, by Richard Chace Tolman, *The Principles of Statistical Mechanics* (New York: Dover, 1979 [1938]), p. 167 (although Tolman usually works in a Gibbsian framework). In Boltzmann's original work, as interpreted by Paul Ehrenfest and Tatiana Eherenfest, *The Conceptual Foundations of the Statistical Approach in Mechanics* (Mineola, NY: Dover, 2002 [Ithaca, NY: Cornell, 1959]), the macrostates in Γ express equivalence groups in μ space relative to some given resolution power with respect to a molecular state. More generally, the partition of Γ into macrostates can be described by a mapping that determines the region to which any point in Γ belongs and that satisfies two conditions. (a) All the subsets of Γ in this partition are given by some measurable function defined over Γ. This condition is necessary in order to make sense of the idea that the entropy of a system is the measure of its macrostate. (b) The measurable subsets have to be disjoint and cover all of Γ. That is, each point must belong to one, and only one, measurable set of points in Γ. This condition ensures that the system has well-defined macroscopic properties at all times.

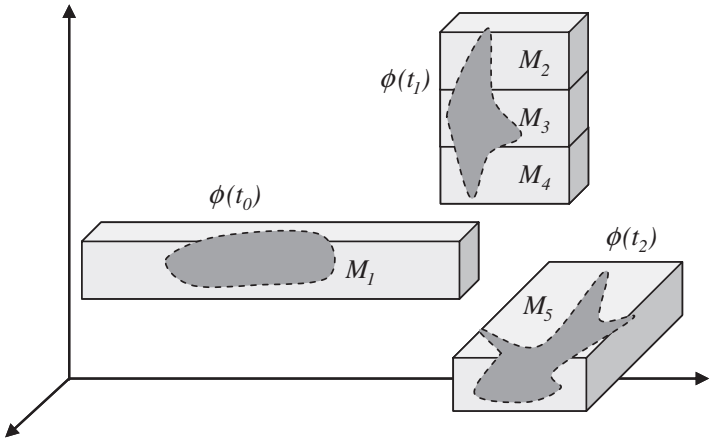


Figure 3: A macroscopic description of the dynamics is obtained by superimposing Figures 1 and 2. At the initial time, the dynamical blob $\phi(t)$ is within the macrostate M_1, and so the observer describes the system as being in M_1. At the time t_1, $\phi(t)$ partially overlaps with three macrostates: M_2, M_3 and M_4, and the observer describes the system as being in one of them, namely, the one containing the actual microstate of the system. If the observer knows the dynamical evolution of the system and the extent to which $\phi(t)$ will overlap with the different macrostates, then the transition probability assigned to the macrostates will be as follows: At t_0: $P(M_1)=1$. At t_1: $P(M_1)=0$, $P(M_2,t_1|M_1,t_0)\approx 1/3$, $P(M_3,t_1|M_1,t_0)\approx 1/3$, $P(M_4,t_1|M_1,t_0)\approx 1/3$. At t_2: $P(M_5,t_2|M_1,t_0)\approx 1$.

microstate at t (which is a point in the dynamical blob $\phi(t)$ at t) belongs to that macrostate. In these terms, statistical mechanics describes the relationship between the time evolution of the dynamical blobs and the macrostates.[8] Figure 3 illustrates the way in which an observer with the resolution capabilities given by Figure 2 sees the dynamical evolution of Figure 1. The distinction between a dynamical blob and a macrostate has implications with respect to the notion of probability in statistical mechanics to which we now turn.

Suppose that we measure the size of sets of microstates in Γ by some measure, say the Lebesgue measure. We now define the *probability* of a macrostate at a given time t_1 relative to an initial macrostate at t_0

[8] The thermodynamic magnitudes are defined only for equilibrium states. A general theory of macrostates would have to give precise definitions of the macroscopic observables in terms of microphysical correlations and equivalence groups thereof that obtain between the observer states and the states of the observed systems. This is the sense in which we understand the term *macroscopic* observable in the classical context.

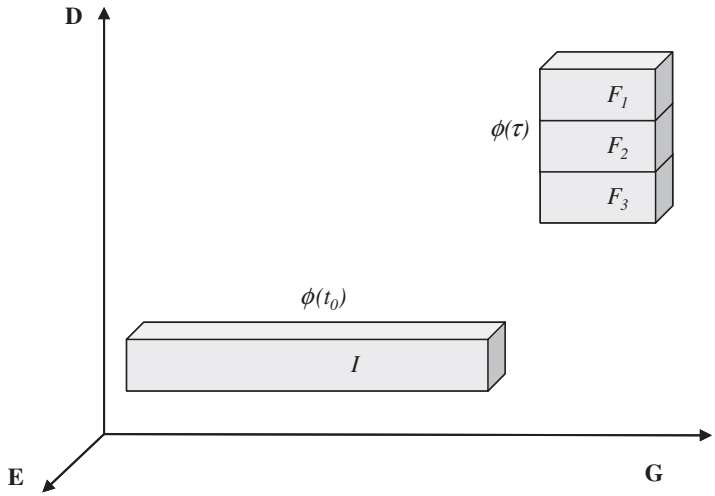


Figure 4: A Demonic evolution. At the initial time t_0 the dynamical blob $\phi(t_0)$ fully overlaps with macrostate I (therefore we draw only the macrostate, for simplicity of presentation). At time τ $\phi(\tau)$ fully overlaps with region $F_1+F_2+F_3$, such that it partially overlaps with each of these macrostates. The actual macrostate at τ is either F_1 or F_2 or F_3. The volume of the blob $\phi(\tau)$ is equal to the initial volume of $\phi(t_0)$ in accordance with Liouville's theorem, but the volume of the final macrostate is smaller than the volume of the initial macrostate, so that the entropy of D+G+E has decreased.

as follows (see Figure 4). Take a system S in an initial macrostate I at t_0. This means that the dynamical blob $\phi(t)$ of S coincides at t_0 with the macrostate I. Consider the time evolution of $\phi(t)$ from t_0 to t_1. At t_1, the time evolved $\phi(t_1)$ partially overlaps with some of the macrostates of S. The probability at t_1 of each macrostate is given by the relative Lebesgue measure of the subset of $\phi(t_1)$ which belongs to that macrostate at time t_1. This definition of probability as *transition probability* seems to us to fit the aim of statistical mechanics, which is to give macroscopic probabilistic predictions for *finite* times based on the *dynamics* of the system and on any macroscopic information we may have about the system. Indeed, the definition above seems to us the only definition of probability that satisfies this aim.[9]

[9] Here is the concise argument. (i) Approaches based on behavior in the infinite time limit (for example, ergodicity) do not yield predictions for finite times, since any finite time behavior is compatible with ergodic dynamics. (ii) Approaches based on ignorance or combinatorial considerations cannot justify the choice of the measure relative to which probability is distributed and are incompatible with the locality of

Under which conditions can the Lebesgue measure of a macrostate be identified with its probability at time t in the above sense? In the above terms, the answer is clear. The conditions are such that the dynamical blob $\phi(t)$ should be spread at time t over the accessible region in such a way that the Lebesgue measure of the blob's subregions contained in the different macrostates are proportional to the Lebesgue measure of the macrostates themselves. We shall call *normal* a dynamical evolution that satisfies this condition during a time interval Δt.[10] Of course, this condition depends on the way the shape of the blob changes by the dynamics; more precisely, whether or not an evolution is normal during a given time interval depends on the way in which the blob $\phi(t)$ spreads over a given set of macrostates at the times in question.

In these terms, one of the most important projects in the foundations of statistical mechanics is to find out the details of the dynamical conditions under which the probability of a macrostate coincides with its Lebesgue measure. In general, even if at some time the probability of a macrostate coincides with its Lebesgue measure, there is no guarantee that this condition will hold at other times. This essentially is the significance of the objections by Loschmidt and by Zermelo to Boltzmann's early theory.

I.2. The Construction of a Demon. Consider the phase space Γ of some isolated subsystem S of the universe illustrated in Figure 4. Each point in Γ describes a *microstate* of S. We divide the degrees of freedom of S into three sets: D, G, and E (for Demon, Gas, and Environment, respectively). We assume throughout that these subsystems, and in particular the Demon, are purely mechanical systems that invariably satisfy all the laws of the underlying mechanical theory, that is, in our case, classical mechanics. This means that the properties of D, G, and E are completely described in the phase space of S (by means of the generalized position and momentum and their functions), and they evolve in time in accordance with the deterministic and time-reversal invariant classical dynamics. In this sense, the Demon indeed is *not* supernatural, as emphasized by Maxwell himself (1868 letter to P. G. Tait; see note 2).

---

classical mechanics. (iii) The future macrostate depends dynamically and probabilistically on the present (or past) macrostate, and therefore probabilistic predictions must take the latter into account. In other words, the probabilities we are after are supposed to predict and explain macroscopic behavior for *finite* times, and are *conditional* on present information.

[10] A uniform probability distribution at time t is a special case of the final state of what we call a normal evolution.

Take now the phase space Γ of S and consider its partition into the macrostates. Suppose that S is prepared initially in the macrostate I. This means that initially the dynamical blob $\phi(t_0)$ exactly coincides with I. Let the microdynamics be such that $\phi(t_0)$ evolves after a *finite* time interval $\Delta t=\tau$, in such a way that at time $t=\tau$ $\phi(t)$ coincides with the region F, which is the union of the three macrostates F_1,F_2,F_3 (we often denote each of these macrostates by F_i where $i=1,2,3$). This means that if S starts out in some microstate in macrostates I at $t=0$ then, at time $t=\tau$, the microstate of S will be in *one* of the regions F_i (that is, F_1 or F_2 or F_3). If, say, S ends up in F_1, then F_2 and F_3 contain only points that belong to counterfactual trajectory segments of S. Such a microdynamics is compatible with Liouville's theorem, since the volume of the union $F_1 + F_2 + F_3$ is equal to (or larger than) the volume of I. Figure 4 illustrates the simple case in which the volume of $F_1 + F_2 + F_3$ is equal to the volume of I.

A dynamical evolution of this sort is Demonic, if the Lebesgue measure of each of the F_i states is smaller than the measure of I. The reason is two-fold. (i) This evolution is entropy reducing since the entropy of S at time t is defined as the logarithm of the Lebesgue measure of the macrostate of S at time t. (ii) The probability for decrease of entropy for the system that starts out in macrostate I is higher than the Lebesgue measure of each of the F_i states. By contrast, if the evolution were normal during Δt (as defined above), the trajectories that start out in I roughly would spread over the accessible region at time $t=\tau$, and, therefore, the probability that S would evolve from I to F (or to every subregion of F_i) during Δt would be proportional to the Lebesgue measure of F (or to the Lebesgue measure of F_i).

Let us sum up. By the concept of probability in statistical mechanics described above, the probabilities we assign to the macro-behavior of a system should be dictated by the behavior of the trajectories over time and, in particular, by the behavior of *finite* segments of trajectories that start out at time t_0 in some known initial macrostate I. At any given time $t > t_0$ the macroscopic behavior of S is determined by the overlap in Γ at time t between the time-evolved blob $\phi(t)$ and the various regions corresponding to the macrostates in Γ. As we said before, the dynamical evolution of S is called normal if and only if the measure of the finite segments starting out in I at time t_0 and arriving into each macrostate F_i at time t is proportional to the Lebesgue measure of each of the F_i. Since in our construction the probability that S arrives into any given F_i is higher than the Lebesgue measure of each of the F_i, the dynamics is Demonic in precisely the sense that it reduces the entropy of S with probability higher than the Lebesgue measure of the target macrostate F_i.

Once it is realized that probabilities and dynamics are intertwined in the way described above, two crucial points immediately follow. First, the Demonic evolution is compatible with *any* probability distribution over initial conditions, say, the distribution over the microstates in the initial macrostate *I*. Second, it is compatible with what we called a normal evolution in the following sense. Recall that a normal evolution means that after a finite time ΔT_n the probability of any macrostate *M* is proportional to the Lebesgue measure of *M*. Given a normal evolution, it is possible to tailor Hamiltonians that will be Demonic for times $\Delta t = \tau$ shorter than the time interval ΔT_n yet still normal at time T_n. This means that our Demon is consistent with some standard probabilistic assumptions of statistical mechanics, in particular the assumption of a uniform probability distribution over the microstates in *I* relative to the Lebesgue measure.[11]

We conclude from this discussion that a Demon is possible. Let us now explain what we mean by 'possible'. First, as we said, a Demon is possible in the sense that it is consistent with the principles of statistical mechanics. Second, a Demon is possible in the sense that it is conceivable that some future segment of the actual evolution of the universe (or of some isolated subsystem of it) will be Demonic. By saying that such an evolution is conceivable, we mean that it is perfectly consistent with all our past experience and with the laws of statistical mechanics. In other words, it might be that the past macroscopic behavior of the universe (as we know it) is not indicative of its future macroscopic behavior, and yet the principles of statistical mechanics hold at all times. That is, it might be that in the short term the evolution will be Demonic while the long-term evolution is normal, and vice versa.

Finally, let us re-describe the Demonic evolution in traditional terms concerning Maxwell's Demon. The partition of Γ into the macrostates *I* and F_1, F_2, F_3 shows that there is a difference in the way that the entropies of the subsystems *D*, *G*, and *E* change in the course of the evolution from *I* to the F_i. Consider the subspace of *S* consisting of the *G* degrees of freedom, which is represented in Figure 4 by the *G* axis. Take the projection of the macrostate of *S* onto this subspace; call the measure of this projection, *relative* to this subspace,[12] *the entropy of the gas*; and similarly for the *D* and *E* subspaces and entropies. The measure

[11] Note that in the dynamical approaches of Boltzmann's equation and its modern successors (for example, Lanford's theorem), the attempts to derive a monotonic entropy increase concern certain Hamiltonians and certain initial conditions. To the extent that they are successful, they show that under these specific conditions and times the evolution is not Demonic for some designated time intervals.

[12] Relative to the whole universe this measure is zero.

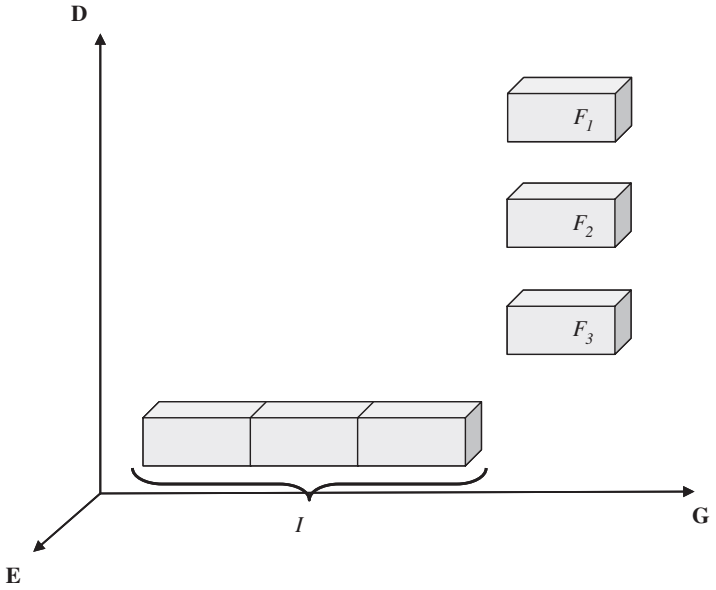


Figure 5: Albert's Demonic construction. The phase space regions corresponding to the macrostates F_1, F_2 and F_3 are topologically disconnected, and – by construction – the union of these regions fully overlaps with the blob $\phi(\tau)$. This entails that $\phi(\tau)$ must have already been topologically disconnected at t_0, when it overlapped with the macrostate I. The region covered by macrostate I is, then, divided into three disconnected regions, and these together form the initial blob.

of the projection of I onto G (relative to the subspace G) is larger than the measure of the projection of any of the F_i regions (F_1 or F_2 or F_3) onto G, and this means that the entropy of G decreases with certainty, so that the gas ends up in a certain predictable *low* entropy macrostate. The entropy of D, by contrast, is unchanged by this dynamics: The projections of the regions I,F_1,F_2,F_3 onto D all have the same measure. The macrostate of E is also unchanged throughout the evolution. Thus, the entropy of the gas has decreased, whereas the entropies of the Demon and of the environment have been conserved. And since this outcome is perfectly macroscopically predictable, it is a Demonic evolution. (We discuss the question of completing the operation cycle below.)

I.3. Remarks on Topology. In the Demonic set-up illustrated in Figure 4, the F region (consisting of the three macrostates F_i) is topologically connected. Albert's original set-up is different (see Figure 5). In his set-up, the dynamics is such that the region F consists of topologically

disconnected regions (the F_i's). Since the dynamical transformation is continuous and time-reversal invariant, this construction implies that the region I must also consist of three topologically disconnected regions. In fact, if the F regions are topologically disconnected, then the whole of the phase space is *decomposable* into dynamically disconnected regions.[13] This means that the dynamics in this case is not ergodic in the Birkhoff-Von Neumann sense of the term, and this is the reason why we prefer our set-up of Figure 4 (in which the phase space can be metrically indecomposable and the dynamics can be ergodic in this sense).

We want to make a clear distinction between topologically disconnected regions and regions which make up different macrostates. The latter are determined by observation capabilities, and it seems to us perfectly conceivable and even reasonable that observers cannot distinguish between any two topologically disconnected regions. To illustrate this point, consider a system whose dynamics is metrically indecomposable (ergodic) in the Birkhoff-Von Neumann sense. Since any phase point must belong to some macrostate, and since macrostates have a positive measure, it follows that if a system is metrically indecomposable there must be macrostates which contain points that belong to two topologically disconnected regions (one of which has measure zero, and the other has measure one). For this reason, region I can be a single macrostate in Albert's set-up, and therefore his set-up is Demonic.

II. SOME CONSTRAINTS

The above construction shows that a Demon is possible. However, the classical dynamics imposes two restrictions on the *efficiency* of the Demonic evolution, as follows.

II.1. Efficiency versus Predictability. In the above scenarios (Figures 4 and 5) of the Demon (as stressed by Albert) there is a tradeoff between a reliable entropy decrease and macroscopic predictability.[14] We now want to draw another linkage, namely, a linkage between the predictability of the Demonic evolution and the *efficiency* of the

[13] In Albert's set-up the dynamics is unstable at both the macro and micro levels, whereas in our set-up the dynamics is unstable only at the macro level.

[14] To avoid confusion, it is essential that notions such as measurement, prediction, and so on be described in purely statistical mechanical terms. This can be done if we think of prediction, for instance, as a sort of computation carried out by a Turing machine, where the machine states and the content of its (long enough but finite) tape are given by the macrostates of D and G, and the evolution between the states and along the tape is determined by the projection of the Universe's dynamics on the corresponding axes.

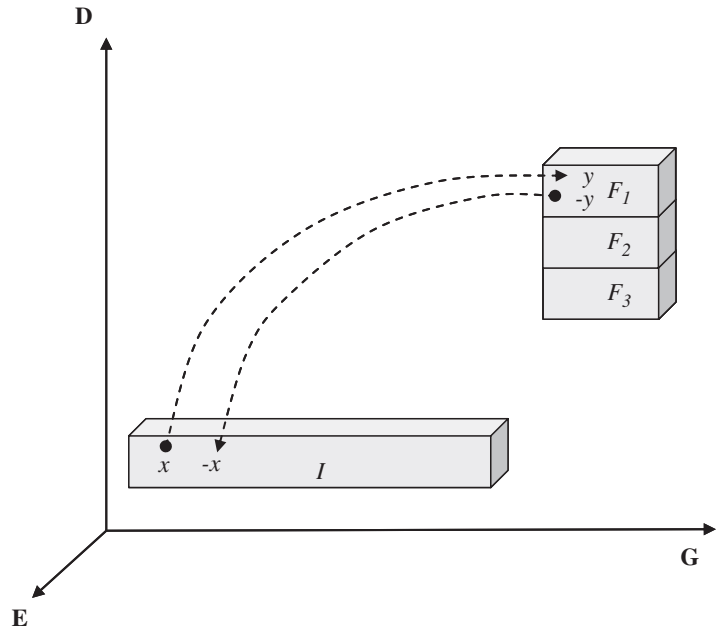


Figure 6: Efficiency of Demon.

Demon in reducing entropy. By efficiency of an operation we mean the entropy difference between the initial and final macrostates.

Consider one of the microstate points in the I macrostate; call it x (see Figure 6). In a Demonic evolution, as described above, this point must sit on a trajectory that takes it to a microstate (call it y) in one of the F_i macrostates (say, F_1) after τ seconds. Consider now the microstates which are the velocity reversals of x and y; call them $-x$ and $-y$, respectively. In many interesting cases (but certainly not in all cases[15]), microstates that are the velocity reversals of each other belong to the same macrostate. Suppose that both x and $-x$ belong to I, and similarly, that both y and $-y$ belong to F_1. This puts constraints on the efficiency and macroscopic predictability of the Demonic evolution. Let us see why.

Since the dynamics is time-reversal invariant, if the trajectory starting out in x in I takes S to the microstate y in F_1, then the trajectory

[15] Consider, for example, the macrostate in which half of the gas molecules move to the right, something we might feel as wind blowing in the right direction. Relative to the phase-space partition that corresponds to our senses, this macrostate is easily distinguishable from the case where the wind blows to the left. However, in many cases it is extremely plausible that x is indistinguishable from $-x$ (consider the air in your room).

that starts out in $-y$ in F_1 takes S back to the microstate $-x$ in I after τ seconds. As the mapping from x in I to y in F_1 reduces the entropy of S, the reversed evolution from $-y$ in F_1 to $-x$ in I increases the entropy of S. However, if S starts off in I and evolves to, say, F_1 (thereby decreasing its entropy), we want it to remain in the low entropy state F_1, avoiding points like $-y$ which take S back to the higher entropy state I after τ seconds. If we wish to make S remain in F_1 we can do one of the following things:

(i) *Stability versus Efficiency.* We can increase the volume of each of the F_i target states (while keeping their number fixed) and thereby increase the *total volume* of the F region. In this case, the relative measure of the set of $-y$ points in F_1 will decrease, and so the probability of the F_1-to-I evolutions will similarly decrease. The reason is that F_1 will include longer trajectory segments which map F_1 to itself. But the larger the volume of F_1, the smaller is the entropy difference between I and F_1. Here, there is a tradeoff between the efficiency of the Demon (that is, the *amount* of entropy decrease) and the *stability* of the low entropy state.

(ii) *Stability versus Predictability* (for a given efficiency). We can increase the *number* of the F states (given a fixed measure of each of the F_i states), so that each of $F_1,F_2,F_3,\ldots$ will still have a small volume (relative to the volume of I), but the total volume of the F region will increase. In this case, the measure of trajectories that arrive into each of the F_i's will be small relative to the volume of the F_i's. The entropy of S will decrease in every cycle of operation, and, moreover, the low entropy final macrostate will be relatively stable. However, as the number of the F_i states increases, the macroscopic evolution of S becomes less predictable. So there is a tradeoff between the stability of the low entropy state and the macroscopic *predictability.*

(iii) According to the optimal interplay between the three factors of stability, predictability, and efficiency, we can combine strategies (i) and (ii), that is, increase the measure of each of the F_i states *and* their number. It is easy to see how this interplay comes about by focusing on the special case in which the volume of I is *equal* to the total volume of the union of the F_i states (as illustrated in Figures 4 and 5). In this case, no matter how much we increase the number of the F_i states, since their total volume is equal to that of I, it follows from the time reversal of the dynamics that the trajectory of S will oscillate between the I and F_i states with frequency $1/2\tau$. Note that the Lebesgue measure of the x-type points is equal to the measure of the $-x$-type points, since the time reversal operation is measure preserving. Yet, none of these constraints undermines the fact that the above scenarios correspond to genuine Demonic evolutions.

II.2. Preparation. In order to display a Demonic behavior as in the above scenarios, *S* must start out in macrostate *I*. Once it reaches the state *I*, it will evolve Demonically spontaneously in the way we spelled out above. But note that *S* will exhibit the Demonic behavior only if and when it reaches macrostate *I*. How can *S* arrive into this initial macrostate? It is a consequence of Liouville's theorem that the measure of *I* cannot be greater than 1/2 of the measure of the entire accessible region. In particular, *I* cannot be an equilibrium state in the combinatorial sense of the term, since the volume of an equilibrium state usually takes up almost the entire accessible region in the phase space. Any macrostate whose measure is small enough can be part of a Demonic set-up. Yet, since the universe is in a low entropy state right now, this constraint does not really undermine the possibility that the universe will evolve Demonically in the future.

III. COMPLETING THE OPERATION CYCLE

By the definition we gave earlier, a system is Demonic if its entropy decreases with probability higher than that determined by the standard Lebesgue measures of the initial and final macrostates. However, some writers argue that this is not sufficient; they add the requirement that an evolution be considered Demonic only if, in addition, the cycle of operation is completed.[16] We do not want to go into the question of whether or not this requirement is justified. Instead, we will show now how it *can* be satisfied by our construction.

III.1. Three Requirements. What is a completion of an operation cycle? Once the cycle is completed we do not want the system to return exactly to its initial macrostate, since in particular we want the entropy of the gas to remain low. Instead, the idea is that at the end of the cycle the situation will be as follows. Take our *total* system *S*, consisting of the degrees of freedom *D*, *G*, and *E*. *G* must end up with entropy lower than its initial entropy, while *D* and *E* must end up with entropy not higher than their initial entropy. *D* must end up in its original initial macrostate as well as retain its initial entropy. *E*, by contrast, must retain its initial entropy, but it may end up in a macrostate that is different from its initial macrostate. This latter requirement is in accordance with the standard literature. For example, Bennett and Szilard argue that completing the cycle of operation involves dissipation in the environment, and therefore the environment's final

[16] For more details concerning the cyclic nature of the second law of thermodynamics, see Uffink, "Bluff Your Way in the Second Law of Thermodynamics," *Studies in History and Philosophy of Modern Physics*, XXXII, 3 (September 2001): 305–94.

macrostate is *a fortiori* different from its initial macrostate.[17] (For them, not only does the macrostate of the environment change; there is an increase in the entropy of the environment. We allow for a different macrostate with the same entropy.) Moreover, the overall final macrostate (at the end of the operation cycle) of S must be such that a subsequent entropy-reducing operation cycle can start off, and once the second operation is completed another one can start off, and then another, perpetually. These requirements are often stated in terms of three properties that the final macrostate of S (at the end of the operation cycle) should have:

(i) *Low Entropy*. The total entropy of S at the end of the operation cycle must be lower than its entropy in the initial macrostate I.

(ii) *Reurn of* D *to Ready State*. At the end of the operation cycle, D must return to its initial ready macrostate so that a new cycle of operation can start off.

(iii) *Erased Memory*. The final macrostate of S must be erased in the sense that at the end of the operation cycle it must be *macroscopically uncorrelated* with the F_i macrostates. In other words, at the end of the cycle there should be no macroscopic records of whatever sort that will allow retrodicting which state among the F_i was the actual macrostate of S prior to the erasure. Obviously, the requirement of erasure refers to the *macroscopic* level, since the classical *micro*dynamics is incompatible with erasure at the microscopic level because it is deterministic and time-reversal invariant.[18] Note that requirement (iii) is stronger than requirement (ii), since the memory could be stored in systems other than D.

Before we proceed to showing how all these requirements can be achieved, it is instructive to consider two attempts that do not work. The first attempt does not obey Liouville's theorem, and the second increases entropy.

Consider a dynamics which takes S from I to one of the F_i states (as before). Then: S evolves to a macrostate A (see Figure 7) such that D goes back to its initial state (requirement ii) while leaving G in its

[17] See Bennett, *op. cit.*; and Leo Szilard's 1929 paper, "On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings," in John Archibald Wheeler and Wojciech Hubert Zurek, eds., *Quantum Theory and Measurement* (Princeton: University Press, 1983), pp. 539–48.

[18] By contrast, the quantum *micro*dynamics is consistent with microscopic memory erasure (requirement iii). The information carried by the value of a quantum mechanical observable of a system in state $|\psi\rangle$ can be erased by measuring observables that do not commute with $|\psi\rangle\langle\psi|$. However, the quantum microdynamics cannot satisfy both requirements (ii) and (iii) at the microscopic level without violating unitarity. Here we only consider a classical erasure.

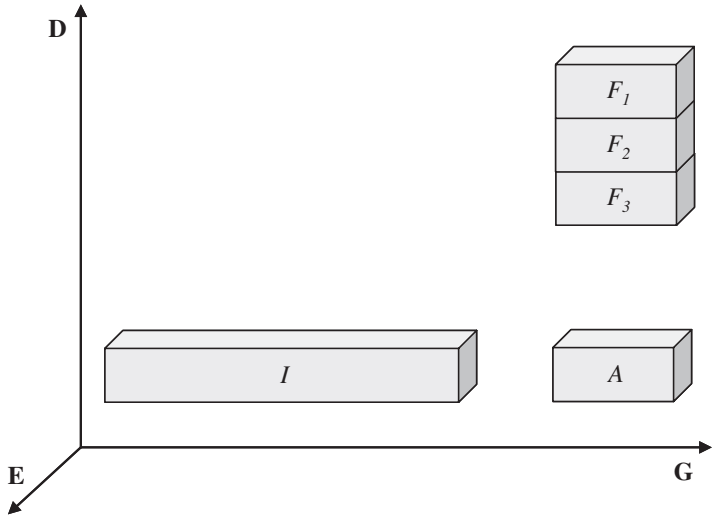


Figure 7: Erasure that violates Liouville's theorem.

low entropy state (requirement i), and E is unchanged. Such dynamics erases memory (requirement iii), since from the final macrostate A it is impossible to retrodict which among the F_i macrostates was the previous macrostate of S. However, this dynamics violates Liouville's theorem, since it maps the entire blob $F_1 + F_2 + F_3$ into A, whose volume is smaller than the volume of $F_1 + F_2 + F_3$. Therefore, such a process is impossible. This, in essence, is the difficulty addressed by Landauer.

The second attempt maps the macrostates F_1,F_2,F_3 to the region A (see Figure 8), where region A now has the following properties. It contains all the microstates at which the trajectories leaving region F arrive after τ' seconds (thus obeying Liouville's theorem); G retains its low entropy state, and D returns to its initial ready state (requirement ii). Memory is erased, since from the information that S is in macrostate A, it is impossible to infer which macrostate it was in before (requirement iii). However, due to Liouville's theorem, the entropy of E increases, and so the final entropy of S is the same as the initial entropy in macrostate I. The achievement of reducing the entropy by the transformation from I to one of the macrostates F_1 *or* F_2 *or* F_3 is lost, contrary to requirement (i).

We now turn to show, by way of construction, how requirements (i), (ii), and (iii) can be achieved together without violating any principle of mechanics.

III.2. Low Entropy and Return to Ready State. We begin with requirements (i) low entropy and (ii) return to the ready state. Consider

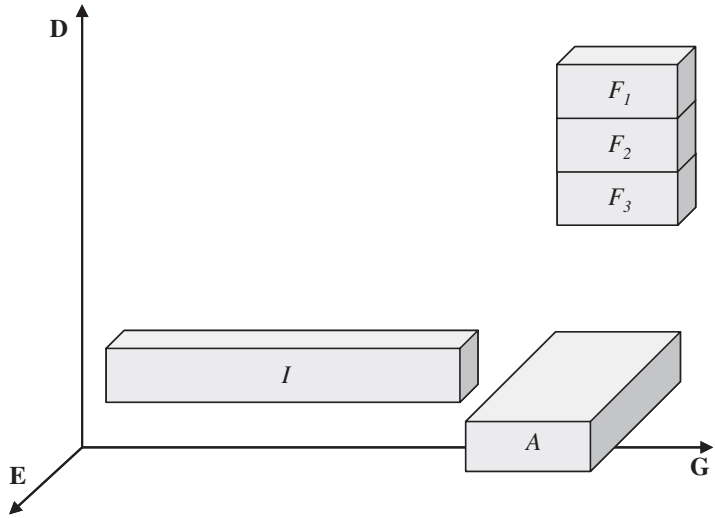


Figure 8: Dissipative erasure.

Figure 9. The region A is partitioned into three disjoint macrostates A_1, A_2, and A_3 such that the union of their volumes is at least as large as the union of the volumes of F_1, F_2, and F_3. In the simplest case, illustrated in Figure 9, the volumes of A_1, A_2, and A_3 are all the same and are equal to the volumes of F_1, F_2, and F_3. We now require that the dynamics maps the F_i states (after a certain time interval) to the macrostates A_i (i=1, 2, 3).[19] The *actual* final state of S will be *one* of the A_i macrostates, and the volume of that macrostate is, by construction, equal to the volume of each of the F_i macrostates and smaller than the volume of the initial macrostate I. This means that the total entropy of S during the evolution from F to A does not change, and in particular it does not increase. So the evolution satisfies requirement (i) of low entropy.[20]

Let us see now what this entropy-conserving transformation implies for the three subsystems separately: G, D, and E. The A_i macrostates are chosen such that the projection along the G axis is the same as in the F_i macrostates, and so G retains its low entropy. The projection

[19] If the regions F_1,F_2,F_3 are topologically disconnected, then so will be the regions A_1,A_2,A_3. This will put some constraints on the dynamics of the erasure; see below. Since in our set-up the F_i regions are connected, this problem does not arise.

[20] The partition into thermodynamic macrostates might even yield smaller and more numerous A macrostates such that the entropy of the final macrostate at the completion of the cycle would be even smaller than it was at t=τ; but this is more than we need right now.

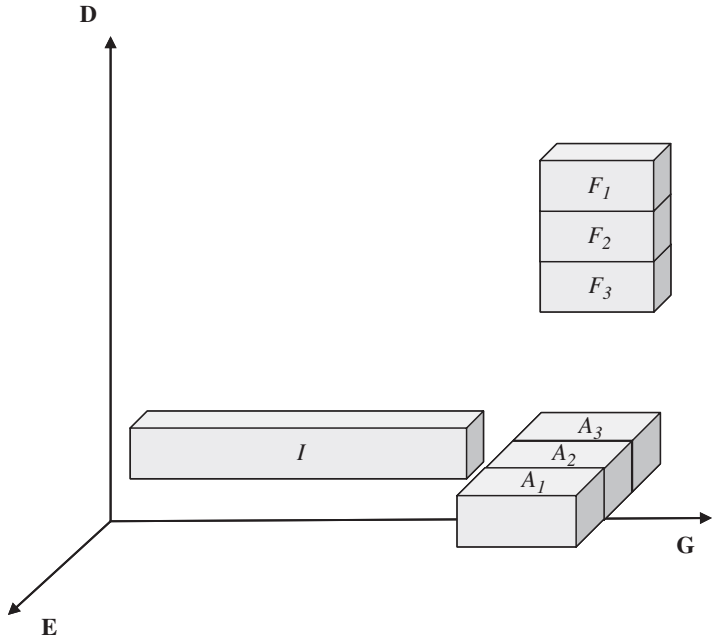


Figure 9: Entropy conserving erasure I.

along the *D* axis is the same as in the initial macrostate *I*, and so by this dynamics *D* returns to its initial ready state. So requirement (ii) of return to the ready state is satisfied for *D*. Moreover, the entropy of *D* has not changed throughout the evolution.

Along the *E* axis, there are by construction three regions corresponding to three possible final macrostates *E1*, *E2*, *E3* of *E*. The *entropy* of *E* in each of these macrostates is the same as it was in the initial state *I*, although *E*'s final macrostate is different from its initial one. As we said above, the fact that *E* ends up in a macrostate different from its initial macrostate is not a problem and is in accordance with the standard requirements in the literature. Moreover, we can construct the evolution from *F* to *A* such that the entropy of *E* will *decrease* by taking a partition of *A* into more numerous and smaller subsets. In this case, obviously, *E* not only need not but cannot return to its initial macrostate. So requiring that it will return to its initial macrostate is superfluous.

III.3. Memory Erasure. We will now show, by explicit and general phase-space construction, that it is possible to construe the *A* macrostates such that our dynamics will result in a genuine memory erasure without increasing the total entropy of *S* or violating Liouville's theorem.

So far, nothing in our construction corresponds to memory erasure, since it is possible that the A_1, A_2, and A_3 macrostates are one-to-one correlated with the F_1, F_2, and F_3 macrostates, so that from the final A_i macrostate it is possible to retrodict the F_i macrostate. However, such a correlation easily can be avoided, as follows. Consider a dynamics such that 1/3 of the points in each of the regions F_1, F_2, and F_3 are mapped onto each of the regions A_1, A_2, and A_3. Conversely, this dynamics entails that among all the points that arrive into each of the A_i regions from the F_i regions, *1/3* arrive from each of the F_i regions.

By this construction, the A_i macrostates are *not* macroscopically correlated to the F_i macrostates, and in this sense they bear no information about their macroscopic history. Given the final A_i macrostate, it is impossible to retrodict the F_i macrostate. In particular, given the A_i macrostate of S, say it is A_1, it is impossible to reconstruct the historical macrostate F_i, since the dynamics maps *sub*-regions of the F_i macrostates to *sub*-regions of the A_i macrostates. Therefore, the F-to-A transformation is a memory erasure, and moreover, as we just saw, it is a *dissipationless* memory erasure (relative to the carving up of the phase space into the F_i and A_i macrostates). More generally, relative to any given set of macrostates, there is an erasing dynamics (in finite times) of the kind spelled out above which is perfectly compatible with Liouville's theorem and with the requirements of low entropy and return to the ready state. At the same time, the actual final A_i macrostate, and in particular the projection of A_i onto the E axis, is macroscopically unpredictable given the previous F_i state of S.

By this construction we have demonstrated that the cycle of operation in a Demonic evolution can be completed *in the right sense of completion*. The initial and final macrostates of S are indeed different; by the end of each cycle the number of macrostates of E which overlap with the blob is (in our set-up) *tripled*. But this is irrelevant to the questions of Maxwell's Demon and memory erasure. More generally, our construction shows that the exponential increase in the number of macrostates is perfectly compatible with a reliable, regular, and repeatable entropy decrease and genuine memory erasure.

According to the Landauer-Bennett thesis, memory erasure is necessarily accompanied by a compensating entropy increase of $kln2$ per bit of lost information. Landauer and Bennett base their thesis on Liouville's theorem. Our F-to-A dynamics is a *counterexample* directly refuting the thesis.

Finally, consider a more refined partition of the F and A regions into macrostates (see Figure 10). Instead of macrostate F_1, for example, we have three macrostates, F_{11},F_{12},F_{13}; instead of the macrostate A_1 we have A_{11},A_{12},A_{13}; and so on, such that F_{11},F_{12},F_{13} are mapped to

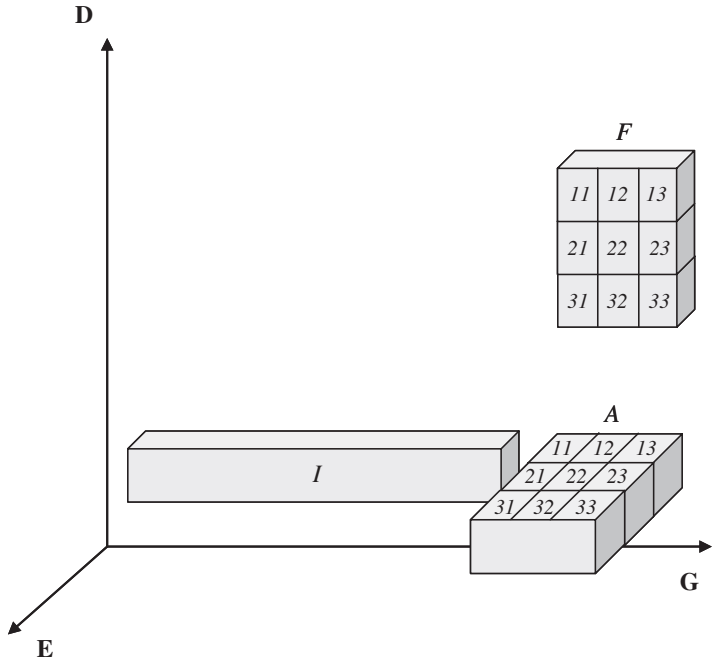


Figure 10: Entropy conserving erasure II.

A_1; F_{21},F_{22},F_{23} are mapped to A_2; and F_{31},F_{32},F_{33} are mapped to A_3. Conversely, A_{11},A_{21},A_{31} are mapped to F_1, and so on.[21] Relative to this partition, the dynamics described above is *not* a memory erasure, since it is possible to retrodict from the actual macrostate A_{ij} the macroscopic history of *S*. But an erasure dynamics *can* be constructed relative to this partition, in an essentially similar way to the one above. We see that a memory erasure is *relative* to a partition of the phase space. However, note that there is *no* universal erasure, that is, an erasure applicable to all possible partitions, however refined, since a universal erasure would require dynamics that is *maximally* mixed in a *finite* time interval. This is impossible, because it is impossible that after a finite time interval *every* set of positive measure in *every* macrostate contains end points that arrived from *all* the other macrostates.[22]

[21] In our set-up, the macrostates F_{11},F_{12},F_{13}, and so on need not correspond to topologically disconnected regions (for the same reason we have argued before; see section 1). However, if the macrostates F_1,F_2,F_3 are topologically disconnected, then so will be their sub-regions F_{ij} (that is, F_{11},F_{12},F_{13}, and so on) and the regions A_1,A_2,A_3 and their sub-regions A_{ij}.

[22] This is an implication of the locality of classical mechanics.

IV. CONCLUSION

> The law that entropy always increases—the second law of thermodynamics—holds, I think, the supreme position among the laws of Nature.... [I]f your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation.
>
> —Arthur Eddington[23]

We believe that Eddington was wrong. We have just shown that Maxwell's Demon is compatible with (classical) statistical mechanics, and therefore the second law of thermodynamics is not universally true if statistical mechanics is.

Nevertheless, if we take our experience as a guide, we cannot construct Demons. How can we explain this, given that Maxwellian Demons are in principle possible? As we said, whether or not Maxwellian Demons can be constructed in our world depends on the kinds of Hamiltonians we can construct and the initial conditions we can control. Maxwellian Demons are possible in cases where there is the right sort of harmony between the dynamics (the evolution of the dynamical blobs $\phi(t)$) and the partition of the phase space into macrostates. One can construct Demons either by finding the right sort of dynamical evolution to match a given set of macrostates (by constructing the Hamiltonian), or by finding the right set of macrostates to match a given dynamics (by constructing the right measuring devices).[24] If we could achieve such a Demonic harmony, we could extract work from heat and, contrary to the Landauer-Bennett thesis, perform a logically irreversible computation without dissipation.

It seems to us that the difficulties in actually constructing a Demonic system involve practical issues such as controlling a large number of degrees of freedom and their initial conditions, interventionist considerations,[25] and so on. Since the issues involved here are merely practical, ruling out the possibility of the Demon in advance

[23] Arthur Eddington, *The Nature of the Physical World* (London: Everyman's Library, J. M. Dent, 1935), p. 81.

[24] Compare Adolf Grünbaum's suggestion: "...[F]or any specified ensemble there will plainly be coarse-grainings that make the ensemble's entropy do whatever one likes, at least for finite time intervals." See Lawrence Sklar, *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics* (New York: Cambridge, 1993), p. 357.

[25] In quantum mechanics, interventionist constraints presumably would be related to decoherence effects. On the role of decoherence in statistical mechanics, see Hemmo and Shenker, "Can We Explain Thermodynamics by Quantum Decoherence?" *Studies in History and Philosophy of Modern Physics*, XXXII, 4 (December 2001): 555–68, and "Quantum Decoherence and the Approach to Equilibrium (II)," *Studies in History and Philosophy of Modern Physics*, XXXVI, 4 (December 2005): 626–48.

on the basis of, say, the second law of thermodynamics, is circular reasoning.[26] Certainly, the fact that Demons have not been observed in the past does not by itself entail that they will not be observed or constructed in the future.

MEIR HEMMO

University of Haifa

ORLY SHENKER

The Hebrew University

[26] In the case of erasure, the dissipation of *klog2* per bit is so small that it cannot be measured given present technology.

A MODAL THEORY OF FUNCTION*

The function of a trait token is usually defined in terms of some properties of other (past, present, or future) tokens of the same trait type. I argue that this strategy is problematic, as trait types are (at least partly) individuated by their functional properties, which would lead to circularity. In order to avoid this problem, I suggest a way to define the function of a trait token in terms of the properties of the very same trait token. To able to allow for the possibility of malfunctioning, some of these properties need to be modal ones: a function of a trait is to do *F* just in case its doing *F* would contribute to the inclusive fitness of the organism whose trait it is. Function attributions have modal force. Finally, I explore whether and how this theory of biological function could be modified to cover artifact function.

I. ARTIFACT FUNCTION AND BIOLOGICAL FUNCTION

The function of my corkscrew is to open bottles. The function of my heart is to pump blood. These two function-attributions are of different kinds. My corkscrew is an artifact, whereas my heart is a biological organ. Artifact function seems to be the easier of the two kinds to analyze. The standard way of explaining artifact function is with reference to the notion of design. My corkscrew has the function to open wine bottles if and only if it was designed to open wine bottles. If we wonder what the function of an artifact may be, we should just ask the designer to get an answer.

This explanatory scheme will not work in the case of biological functions, as there is no designer who designed biological traits (or, in any case, there is no one we could ask). Thus, it seems that in spite of the fact that we talk about functions both in the artifact and in the biological case, these two kinds of function are very different indeed: one is fixed by design, whereas the other is not.

I focus on biological function in this paper, but at the end I will come back to the notion of artifact function and reevaluate the

*I presented earlier versions of this paper at the Pacific APA in March 2006 as well as at the University of British Columbia (April 2006) and at Syracuse University (November 2006). I am grateful to my commentator at the APA, Robert Richardson. I am especially grateful to Mohan Matthen, John MacFarlane, and Mark Heller for comments and discussion. I am equally grateful for the useful comments I received from three referees of this JOURNAL.

0022-362X/10/0708/412–431

standard analysis of artifact function in light of the argument presented in the case of biological function.

After considering some important desiderata every theory of function needs to satisfy (section II), I point out that the function of a trait token is usually defined in terms of some properties of other (past, present, future) tokens of the same trait type. I argue that this strategy is problematic, as trait types are usually individuated (at least partly) in terms of their functional properties, which would lead to circularity (sections III–V). In order to avoid this problem, I suggest a way to define the function of a trait token in terms of the modal properties of the very same trait token (sections VI–VII). Finally, I explore whether and how this theory of biological function could be modified to cover artifact function (section VIII).

II. THREE DESIDERATA FOR A THEORY OF FUNCTION

There may be many more than three desiderata for a theory of biological (or artifact) function, but I will mention three of these, which I take to be the most important ones and which apply in both the biological and the artifact cases.

First, a trait can have two (or more) functions at a time. The function of my mouth is both to eat and to speak, for example. A theory of function should be able to allow for the possibility that one trait has two (or more) functions.

Second, function attributions can depend on the explanatory project at hand. The function of my left eyelid is to blink, but its function is also to keep my left eye moist. It depends on the explanatory project at hand which function attribution we will opt for. Suppose that we are concerned with the anatomy of the eyelid, regardless of its relation to the eye. In this explanatory project, it will be irrelevant whether the eye is kept moist or not: the function of the eyelid is to contract and expand in a certain way: to blink. In some other explanatory projects, however, where we analyze the moistness of the eye and we are not concerned with the anatomy of the eyelid, the function of the eyelid in this explanatory scheme will be to keep the eye moist. It should not be a surprising claim that function attributions can depend on the explanatory project. It has been argued that the function of a trait explains why this trait is the way it is.[1] If, however, explanations are considered to be responses to why-questions[2] and, therefore, themselves

[1] L. Wright, "Functions," *Philosophical Review*, LXXXII, 2 (1973): 139–68; Paul E. Griffiths, "Functional Analysis and Proper Functions," *British Journal for the Philosophy of Science*, XLIV, 3 (1993): 409–22.

[2] Bas C. van Fraassen, *The Scientific Image* (New York: Oxford, 1980).

depend on the explanatory project, then the function we attribute to a trait will also depend on the explanatory project.[3]

Third, any theory of function must be able to account for the phenomenon of malfunctioning. A trait can have a function but fail to perform this function. If my heart skips a beat, it still has the function to pump blood, but at that moment it fails to perform this function: it malfunctions.

III. THE ETIOLOGICAL THEORY OF BIOLOGICAL FUNCTION

The most widespread notion of biological function is the following: a trait of an organism has function *F* if and only if its performing *F* has contributed to the survival of the ancestors of this organism. This notion of function is usually referred to as 'etiological': what determines the function of a trait is its history. The function of the human heart is to pump blood because the fact that the heart pumped blood contributed to the survival of our ancestors.[4]

According to a widely accepted version of the etiological theory, the "modern history theory of functions," in order for a trait to have a function it must be the case that this trait has *recently* contributed to the survival of the organism's ancestors.[5] If a trait contributed to the survival of an organism's ancestors in the distant past but has not contributed since, it does not have a function. The human appendix, for example, has not contributed to our survival recently; thus, it does not have any function. To sum up, the etiological view of function asserts that the function of a trait is determined by its *recent* history.

Note that this theory of biological function restores the continuity between the explanation of biological and artifact functions. The function of my corkscrew is to open wine bottles because it has been designed to open wine bottles, whereas the function of my heart is to

[3] This feature of function attributions also explains why biological function is sometimes thought to be indeterminate (see, for example Peter Godfrey-Smith, "A Modern History Theory of Functions," *Noûs*, XXVIII, 3 (1994): 344–62, at p. 356). In the distant past, the human appendix had the function to decompose celluloid. Now it no longer has this function. At some point in our evolutionary history, the human appendix ceased to have this function. But it is difficult to see what could be the criterion for the exact time when it no longer had this function. If function attribution depends on the explanatory project, then, depending on with which explanatory project we are engaging, the human appendix may or may not have the function to decompose celluloid.

[4] Ruth G. Millikan, *Language, Thought and Other Biological Categories* (Cambridge: MIT, 1984); Karen Neander, "Functions as Selected Effects: The Conceptual Analyst's Defense," *Philosophy of Science*, LVIII, 2 (1991): 168–84; Neander, "The Teleological Notion of 'Function'," *Australasian Journal of Philosophy*, LXIX, 4 (1991): 454–68; Griffiths, *op. cit.*; Godfrey-Smith, *op. cit.*; Wright, *op. cit.*

[5] Godfrey-Smith, *op. cit.*

pump blood because it has been selected for pumping blood. In both cases, function is fixed by the past: past design or past selection.

Probably the most famous objection to the etiological view is based on the swampman thought experiment. A very direct consequence of the etiological definition of function is that what fixes the function of a trait is its past, not its present. Hence, if an organism that is molecule-for-molecule identical to me (the swampman) were created by chance, its organs would not have any functions, since it would lack the evolutionary history that would fix the function of these organs. Without engaging with the Byzantine swampman literature, I raise a more serious objection to the etiological theory in the next section and then generalize this objection to other theories of function in section v.

IV. A NEW OBJECTION: THE INDIVIDUATION OF TRAIT TYPES

The etiological definition of function presupposes that *trait types* can be individuated in an unproblematic manner. The trait whose function is to be defined and the traits that have been selected for in the past must be *of the same type*. But how can we individuate trait types? What makes hearts different from nonhearts?

I will argue that there is no coherent, noncircular way of individuating trait types that is available to the etiological theory of function.

The question, then, is how trait types are individuated. I will consider three options and point out that none of them is available to the etiological theory of function.

IV.A. Functional Criteria. The most widely accepted account of trait-type individuation holds that tokens of a certain trait type all have the same function. A token object belongs to trait type *T* if and only if it has certain functional properties: if it has the function to do *F*. Those entities are hearts that have the function of pumping blood. Those entities that do not have this function are not hearts.

As Karen Neander puts it: "Most biological categories are only definable in functional terms."[6] This account of individuating trait types is widely (though not universally) accepted as a general suggestion both in philosophy of biology and in philosophy in general. Tyler Burge, for example, writes that "to be a heart, an entity has to have the normal evolved function of pumping blood in a body's circulatory system."[7]

[6] Neander, "Functions as Selected Effects," *op. cit.*, p. 180; see also Morton Beckner, *The Biological Way of Thought* (New York: Columbia, 1959), p. 112, and Tim Lewens, *Organisms and Artifacts: Design in Nature and Elsewhere* (Cambridge: MIT, 2004), p. 99.

[7] Tyler Burge, "Individuation and Causation in Psychology," *Pacific Philosophical Quarterly*, DCCVII, 4 (1989): 303–22.

It is important to note, however, that the etiological theory of function cannot help itself to this way of individuating trait types when defining function without running into circularity.

As we have seen, the etiological definition of function presupposes an account of trait-type individuation. Now, if we want to avoid circularity, we cannot use the notion of function in order to explain trait-type individuation. When we are explaining function, the claim that x^* (the trait whose function we are explaining) is a token of type X (the traits that have been selected in the past) is part of the *explanans*. Hence, we cannot use the *explanandum* (function) to explain part of the *explanans* (why x^* is a token of type X).[8]

Thus, if we want to talk about trait types in the definition of function, we need some other way of individuating them.

IV.B. Morphological Criteria. A simpler suggestion is that we can use morphological criteria for individuating trait types. The proposal is that a token object belongs to trait type T if and only if it has certain morphological properties. An entity is a heart if, for example, it has a certain shape, size, and color, and it is not a heart otherwise.

One problem with this suggestion is that trait types need to range over different species, but the hearts of different species have very different morphological properties.[9]

But even if we only want to individuate a trait type within a certain species, we still cannot use morphological criteria. A malformed heart that does not have the morphological properties hearts have is a heart all the same (it is a malformed *heart*, after all). What keeps hearts and nonhearts apart cannot be a set of morphological properties.

A possible suggestion would be to say that hearts are those entities that play a certain causal role, and those entities that do not play this causal role are not hearts.[10] This proposal, however, would not work in the malformed heart case: a malformed heart does not have the causal role hearts have, but it is a heart nonetheless. What

[8] This problem is acknowledged by some of the defenders of the etiological theory of function. See Neander, "Types of Traits: Function, Structure, and Homology in the Classification of Traits," in Andre Ariew, Robert Cummins, and Mark Perlman, eds., *Functions: New Essays in the Philosophy of Biology and Psychology* (New York: Oxford, 2002), pp. 402–22, especially p. 403; see also Griffiths, *op. cit.*, and Paul Sheldon Davies, *Norms of Nature: Naturalism and the Nature of Functions* (Cambridge: MIT, 2001).

[9] Neander, "Functions as Selected Effects," at p. 180.

[10] I discuss this proposal here because its most natural rendering falls under the morphological account of trait-type individuation, but it is worth noting that it could also be interpreted as a version of the functional account—if we conceive of function in a way Cummins does in his "Functional Analysis," this JOURNAL, LXXII, 11 (November 1975): 741–64, and in his "Neo-Teleology," in Ariew, Cummins, and Perlman, eds., *op. cit.*, pp. 157–73.

matters is not what the heart does (or how it looks), but what it is supposed to do.

To sum up, the suggestion that morphological criteria could be found for individuating trait types does not work.[11] This leaves us with a third alternative.

IV.C. Homological Criteria. A third possible answer is the following. One could argue that what guarantees that two traits are tokens of the same type is that they are homologues: they have common descent; they are members of the same "reproductively established family."[12] To make this suggestion as plausible as possible, we should not confuse it with the view that homologous traits are 'coded by' the same gene, for the simple reason that no trait should be taken to be 'coded by' a gene. The way a trait turns out depends partly on the gene, but it also depends on the intra- and extra-cellular environment during the developmental process. Assuming that the genetic factor can be singled out from this complex causal network is biologically very implausible.[13] The most plausible homological accounts of trait-type individuation are not committed to this gene-centric view.[14] The suggestion is that whether a trait belongs to a homologically established trait type depends not only on what gene this trait was coded by, but also on the developmental process.

The homological account of trait-type individuation is vulnerable to a serious objection. Take the following example. The forelimbs of vertebrates, such as the wings of birds, and the forelegs of ancient amphibians are homologous: the wings of eagles belong to the same reproductively established family as the forelegs of some ancient amphibians. According to the suggestion for individuating trait types we are considering here, they must be of the same trait type.

But the wing of the eagle and the foreleg of an ancient amphibian are clearly not tokens of the same trait type. One of them is a wing, and the other is a foreleg. They belong to very different trait types indeed.

[11] The morphological account of trait-type individuation may be supplemented with a homological one (Ron Amundson and George V. Lauder, "Function without Purpose: The Uses of Causal Role Function in Evolutionary Biology," *Biology and Philosophy*, IX (1994): 443–69; Lauder, "Homology, Form, and Function," in Brian K. Hall, ed., *Homology: The Hierarchical Basis of Comparative Biology* (San Diego: Academic Press, 1994), pp. 151–96). As we will see, however, this homological account raises serious worries.

[12] See for example Amundson and Lauder, "Function without Purpose," *op. cit.*; Lewens, *op. cit.*, pp. 99–100; Millikan, *op. cit.* The term "reproductively established families" was introduced by Millikan (*ibid.*, p. 23ff).

[13] Griffiths and R. D. Gray, "Developmental Systems and Evolutionary Explanation," this JOURNAL, XCI, 6 (June 1994): 277–304, especially pp. 298ff.

[14] V. Louise Roth, "On Homology," *Biological Journal of the Linnean Society*, XXII (1984): 13–39, especially p. 17; Günther P. Wagner, "Homology and the Mechanisms of Development," in Hall, ed., *op. cit.*, pp. 273–99.

The defender of the homological way of individuating trait types may argue that these two traits do belong to the same trait type; both are forelimbs after all. A trait token can be typed in many different ways, and typing the eagle's wings as forelimbs is a perfectly valid way of doing so. The real problem is that this way of individuating trait types cannot individuate trait types narrowly enough: it will not even be able to differentiate between wings and forelegs. More importantly, such a broad way of typing traits does not help in our definition of function, as using this way of talking about trait types would attribute the function of crawling to the eagle's wings.

But perhaps such old members of a reproductively established family (for example, the traits of our ancient amphibian) just do not count. A possible suggestion would be to say that wings belong to the same trait type because they are all *recent* members of a reproductively established family. In other words, two token objects belong to the same trait type if and only if they are *recent* homologues: *recent* members of a reproductively established family.

The problem with this suggestion is that there is no noncircular way of cashing out what is meant by 'recent'. We would be happy to say that the eyes of the eagle and the eyes of the ancient amphibian are tokens of the same type. Then why can't we do the same with forelimbs? What is so different in the two cases that makes us sort the two token traits under the same type in the latter case but not in the former?

The only thing that differentiates the example of the eye from the example of the forelimb is that the selection pressure changed in the latter case but not in the former. Forelimbs have been selected for doing something *different* in the bird population and in the ancient amphibian population. Eyes, on the other hand, have been selected for doing *the same thing* in the bird population and in the ancient amphibian population.

Thus, if we want to make sense of the suggestion that two trait tokens belong to the same trait type if and only if they are *recent* homologues, then we will have difficulties defining what is meant by the term 'recent'. In defining the eagle's trait types, 'recent' includes the ancient amphibian population when we are analyzing the eye example, but it certainly does not include the ancient amphibian population when it comes to forelimbs. The bottom line is that what 'recent' amounts to depends on what the trait in question has been selected for; what 'recent' amounts to depends on the etiological function of the trait. The suggestion boils down to the claim that what makes a trait token a token of a certain trait type is that it is a homologue of trait tokens that were selected for doing the same thing as this token (or, in other words, that had the same etiological function as this token).

Hence, this way of individuating trait types collapses into the functional account of trait-type individuation. But we have seen above that the functional account of trait-type individuation cannot be used in the definition of function without running into circularity.

To sum up, the etiological theory of function cannot rely on any of the three ways of individuating trait types. Since the etiological notion of function requires an unproblematic way of individuating trait types, we need to dispose of this theory of function.[15]

V. BEYOND ETIOLOGY: A MORE GENERAL PROBLEM

If the argument I presented here is correct, then the etiological theory of function cannot stand, for it has to rely on an independent account of individuating trait types, and no such account is available for the etiological theory. So, the etiological theory should be disposed of, and we should look for some other theory of function. The problem is that *all* alternatives to the etiological theory of function rely on an independent account of individuating trait types.

The main alternative to the etiological theory of function has been the so-called propensity theory, which claims that it is not the past but the future of the organism that fixes the function of a trait.[16] The function of a trait is what *will (be likely to) contribute* to the survival of the organism. In other words, function is a propensity.

According to the propensity definition, a trait "has a (biological) function just when it confers a survival enhancing propensity on a creature that possesses it."[17] In other words, the function of a trait is its propensity to contribute to the fitness of the organism: "when we speak of the function of a character, therefore, we mean that the character generates propensities that are survival-enhancing in the creature's natural habitat."[18]

Several objections have been raised against this view.[19] Whether or not these objections are conclusive, it needs to be noted that the propensity definition of function also presupposes an unproblematic

[15] One could argue at this point that these three ways of individuating trait types are not exhaustive. More specifically, one could argue for some kind of hybrid account (see for example Neander, "Functions as Selected Effects," at p. 178, and "Types of Traits," especially pp. 403–04). It can be pointed out that these hybrid accounts of trait-type individuation would either collapse into a homological account or raise the same problems about circularity as the functional account of trait-type individuation.

[16] John Bigelow and Robert Pargetter, "Functions," this JOURNAL, LXXXIV, 4 (April 1987): 181–96.

[17] *Ibid.*, p. 192.

[18] *Ibid.*

[19] Godfrey-Smith, *op. cit.*, especially pp. 352–53; Neander, "The Teleological Notion of 'Function'"; Millikan, *White Queen Psychology and Other Tales for Alice* (Cambridge: MIT, 1993); Denis M. Walsh, "Fitness and Function," *British Journal for the Philosophy of Science*, XLVII, 4 (1996): 553–74.

account of how trait types are individuated: a trait token has the function to do *F* if and only if the fact that traits *of the same type* will do *F* will contribute to the organism's survival.[20]

There is a third theory of function that we should consider: the relational theory of function.[21] It has been argued that we cannot talk about the function of a trait in general, but only the function of a trait relative to a certain selective regime.[22] Here is Walsh's definition: "The/a function of a token of type *X* with respect to selective regime *R* is to *m* iff *X*'s doing *m* positively (and significantly) contributes to the average fitness of individuals possessing *X* with respect to *R*."[23] Again, this definition presupposes an independent way of individuating trait types.

In other words, the most important candidates for defining function presuppose an account of individuating trait types. Thus, we have a serious worry. We are left with no plausible theory of function.

VI. A MODAL THEORY OF FUNCTION

We have seen that every definition of function that talks about trait types faces the trait-type individuation objection. An obvious way to avoid this objection would be to define function without referring to trait types at all. If we accept a notion of function that does not rely on the prior individuation of trait types, then we obviously do not need to worry about trait-type individuation. If we could define function without appealing to trait-type individuation, then we could use this definition of function to individuate trait types without running into circularity.

Note, however, that if a definition of function does not rely on an account of trait-type individuation, then the function of a token trait must be determined entirely by the properties of *that very trait token* and not by the properties of other tokens of the trait type to which this token belongs. In that case, however, it is difficult to see how a trait can malfunction. When a trait malfunctions, it is supposed to do (that is, it has the function to do) *F*, but it does not do *F*. My heart malfunctions when it does not pump blood (though it is supposed to/has the function to do so). If we define the function of a trait

[20] As Godfrey-Smith pointed out, the propensity theory oscillates between talking about the propensity of a trait token (see Bigelow and Pargetter, *op. cit.*, p. 192) and taking about the propensity of a trait type (*ibid.*, pp. 194–95). He convincingly argues that the only plausible reading is the latter (Godfrey-Smith, *op. cit.*, at p. 360).

[21] Walsh, "Fitness and Function," *op. cit.*

[22] By selective regime Walsh means "the total set of abiological and biological (including social, developmental and physiological) factors in the environment of the trait which potentially affect the fitness of individuals with that trait" (*ibid.*, p. 564).

[23] *Ibid.*

token in terms of the properties of that trait token alone, then it is difficult to see how the function can be different from what the trait token actually does. In other words, it is difficult to see how such an account of function could explain malfunctioning.

One possible way to explain how a trait can malfunction is by attributing modal force to claims about function. Trait *x* may not perform *F*, but if it were to perform *F*, this would contribute to the survival of the organism with *x*. Thus, at first approximation, doing *F* is a function of *x* if and only if it is true that if *x* is doing *F*, then this *would* contribute to the survival of the organism with *x*.

Thus, the suggestion is that the tense of 'contribute' in the definition of function is not past tense as in the etiological account. It is not future tense either—this would be the suggestion of the propensity theory. And, finally, it is not present tense, which would be the way the relational theory defines it. According to my account of function, instead of 'contributed', 'will contribute', or 'contributes', we have to use 'would contribute'. Function attributions have modal force.[24]

Some further clarifications are needed about this general suggestion. First, the talk about contribution to the survival of an organism, which is a standard way of analyzing function, is vague. What really matters in natural selection is not the survival but the inclusive fitness of the organism. Further, if a trait's doing *F* contributes to the survival of an organism, the trait is doing *F* at time *t*, but what it contributes to, that is, the organism's survival, is at some other time, *t**. But many things can happen between *t* and *t**. Presumably some kind of appeal to *ceteris paribus* clauses could go around this problem, but to keep things simple, instead of talking about contribution to the survival of an organism, I will talk about contribution to the inclusive fitness of an organism. This will also allow me to define the function of a trait at time *t* in terms of some (modal) facts at time *t*.

Second, I define function with the help of a counterfactual. Any theory of counterfactuals could be used to fill in the details of this definition, but, for simplicity, I will use Lewis's theory.[25] Using a Lewisian account of counterfactuals, my definition of function would amount to the following. Performing *F* is a function of *x* if and only if some possible worlds where *x* is doing *F* and this contributes to the survival of organism *O* are closer to the actual world than any of those possible worlds where *x* is doing *F* but this does not contribute to *O*'s survival.

[24] It has to be noted that at least some versions of the etiological notion of function could also be interpreted as carrying modal force—whether they do depends on how we interpret the concept of 'contribution' in the definition of function.

[25] David Lewis, *Counterfactuals* (Cambridge: Harvard, 1973).

Now, some of these possible worlds may be frightfully distant. It would contribute to any organisms' survival if their scratching their ear killed off any approaching predators. Still, this is not a function of scratching one's ear, because those possible worlds where scratching one's ear can kill off predators are very far away—the project of explaining the function of scratching one's ear should not take into consideration such distant possible worlds.

Thus, the set of possible worlds that we are considering when determining whether the counterfactual that defines function is true or not should be restricted to 'relatively close' possible worlds.

> *Performing* F *is a function of organism* O*'s trait* x *at time* t *if and only if some 'relatively close' possible worlds where* x *is doing* F *at* t *and this contributes to* O*'s inclusive fitness are closer to the actual world than any of those possible worlds where* x *is doing* F *at* t *but this does not contribute to* O*'s inclusive fitness.*[26]

It is important to note that I only intended to define the function of a token trait. Sometimes we talk about the function of trait types: the function of hearts is to pump blood. I will not give a definition for the function of trait types, as this definition would depend on how we individuate trait types, which, as we have seen, is a very difficult question.[27]

If *x* is not doing (or even cannot do) *F* in the actual world, but in a 'relatively close' possible world it is doing *F* and its doing *F* contributes to the organism's inclusive fitness, then we can still attribute function *F* to *x*. This is exactly what happens if a trait is malfunctioning: if it fails to perform its function.

A *prima facie* worry about this modal account of function is that it proliferates functions: there is no limit to the various potential functions F_1, F_2, ... F_n that are such that, if *x* were to do F_i, doing so would contribute to *O*'s inclusive fitness. Note, however, that according to my definition, function attribution implies that some 'relatively close' possible worlds where *x* is doing *F* and this contributes to *O*'s inclusive

[26] Note that this way of defining function individuates function quite narrowly. Grasping in general is not a function of my hand, because it may be the case that the closest possible world where my hand is grasping is one where its doing so does not contribute to my survival. If we want to attribute a function to my hand, we have to give a much more specific characterization: grasping food when I am hungry, for example. My hand, of course, has many functions: grasping food when I am hungry is only one of these. The fact that my definition gives a very specific characterization of the functions of a trait (and not a general specification, like grasping) is an explanatory advantage of my account.

[27] I also have misgivings about the overuse of trait types in evolutionary biology in general. See Bence Nanay, "Population Thinking as Trope Nominalism," *Synthese*, forthcoming.

fitness are closer to the actual world than any of those possible worlds where *x* is doing *F* but this does not contribute to *O*'s inclusive fitness. In other words, if we find a 'relatively close' possible world where *x* is doing *F* and this contributes to *O*'s inclusive fitness, this by no means guarantees that performing *F* is a function of *x*. What is also needed is that some of those worlds where *x* is doing *F* and this contributes to *O*'s inclusive fitness are closer to the actual world than any of those possible worlds where *x* is doing *F* but this does not contribute to *O*'s inclusive fitness.

Take the following example.[28] My left foot could serve as a paddle for swimming fast. This might improve my inclusive fitness—for example, if it made me sexually attractive (or famous). But if so, does this make paddling a function of my left foot according to the modal theory?

The answer is the following. If there is a 'relatively close' possible world where the *F*-ing of my left foot contributes to my inclusive fitness, this does not make *F*-ing a function of my left foot. What is required for my left foot to have a function *F* is that those worlds where it does *F* and this contributes to my inclusive fitness are closer to the actual world than the ones where it does *F* without contributing to my inclusive fitness. It is not enough for function attribution to find a possible world where doing *F* contributes to my inclusive fitness. We need to compare this world to those where it does not. By these standards, paddling would be disqualified from the elite circle of the functions of my left foot.

More slowly: the modal definition of function was this: some 'relatively close' possible worlds where *x* is doing *F* and this contributes to *O*'s inclusive fitness are closer to the actual world than any of those possible worlds where *x* is doing *F* but this does not contribute to *O*'s inclusive fitness. Let us apply this to paddling: some possible worlds where my left foot serves as a paddle for swimming fast and this contributes to my inclusive fitness are closer than any of those possible worlds where it serves as a paddle for swimming fast but this does not contribute to my inclusive fitness. However, there are lots of very nearby possible worlds where my left foot serves as a paddle for swimming fast without this making any difference to my inclusive fitness. In fact, most nearby possible worlds are extremely likely to be such worlds. There are some possible worlds, of course, where it does (say, where I am an Olympic swimmer). But, at least in my case,

[28] I am grateful to Mohan Matthen both for the example and for pushing me to address this general line of objection.

it is extremely unlikely that these worlds would be closer to the actual world than the boring possible worlds where paddling does nothing for my inclusive fitness.

There may be some people for whom there are possible worlds where their feet serve as paddles for swimming fast and this does contribute to their inclusive fitness. Michael Phelps may be one of them. When we apply the modal theory in the case of his feet, we may have to attribute the function of paddling to them, at least in some explanatory projects. But I do not think that we should find this surprising. After the 2008 Summer Olympics, the media was full of commentaries about why he wins all the races, and the commentators literally talked about how his various body parts function to help him swim faster (the function of his palms, the function of his relatively short legs, and so on).

A last worry about paddling: how is it possible that it is not a function of my feet to paddle, but it is a function of Phelps's feet to paddle? This is an important worry because it highlights a crucial aspect of the modal theory of function. The modal theory of function attributes function to trait tokens, not trait types. The function(s) of a token trait is (are) defined in terms of modal facts about this very token trait. Hence, there is no guarantee that my feet and your feet will have the same functions.

Here is another important question about the modal theory of function. There are famous disputed cases of function attribution. Does the modal theory help us to resolve these? One such case is this. When there is a conflict between two male baboons, sometimes one of them picks up an infant (this phenomenon was first observed among Barbary macaques). What is the function of this behavior? There are (at least) two candidates. The first is that the baboon who picks up the infant is using the baby to protect himself from the other male, who does not want to risk hurting the baby because if he does the female baboons will start attacking him. This is called the 'agonistic buffering hypothesis'.[29] The alternative is that the function of this behavior is parental care. It has been observed that infants are most often picked up by a long-term resident male baboon, while the male he is having the conflict with is usually a recent immigrant. In other words, the holder could possibly be a father of the infant, and he is protecting it from the other, possibly infanticidal male.[30]

[29] J. M. Deag and J. H. Crook, "Social Behaviour and 'Agonistic Buffering' in the wild Barbary Macaque *Macaca sylvana* L," *Folia Primatologica*, XV, 3/4 (1971): 183–200.

[30] Curt Busse and William J. Hamilton III, "Infant Carrying by Male Chacma Baboons," *Science*, CCXII, 4500 (June 12, 1981): 1281–83.

How should we decide whether the function of this behavior is self-defense or parental care? The first thing to notice is that according to the modal theory of function the behavior of picking up the infant has both functions: doing both would contribute to the organism's inclusive fitness. Thus, we should not ask which one is the function of this behavior: both are. Rather, the question should be framed in terms of which one of the two is the more relevant/important function of this behavior. And in order to answer this question we need to examine, unsurprisingly, the modal facts. Suppose that a male baboon *A* is picking up his son *C* during his fight with *B*. In order to decide what the (primary, most relevant) function of this behavior is, we need to consider the possible world where *A* is picking up *D* (not *C*), who is not his son. Does this behavior contribute to *A*'s inclusive fitness in this possible world? If it does, then we have reason to believe that the primary function of this behavior is self-defense. But if it does not, then parental care seems to be a more relevant function. This example could be elaborated further, but this sketchy treatment should be enough to underline the importance of modal facts in resolving problematic cases of function attribution.[31]

I left open what counts as a 'relatively close' possible world in the above definition. The short answer is that what counts as a 'relatively close possible world' depends on the explanatory project. As we have seen, function attribution can depend on the explanatory project. One way of explaining this would be to say that different explanatory projects focus on different sets of possible worlds where *x* could be doing *F*. It needs to be spelled out, however, what this dependence on explanatory projects really means.

It depends on the explanatory project how we should analyze the function of my eyes in an environment where it is pitch dark. There is a possible world where everything is the same as in this one except that it is not pitch dark. If we count this possible world as 'relatively close', then my eye does have a function. If we are analyzing the function of my eye in my bedroom with the lights off, then it seems to be a good idea to include the possible world where it is not pitch dark. If we are analyzing a scenario where photons suddenly disappeared from the universe, then we probably should not include the possible world where it is not pitch dark.

In some explanatory projects, it is irrelevant what *x* would or could do if it had different intrinsic properties—we are interested in the

[31] Primatologists, of course, cannot do fieldwork in possible worlds. But they can infer the function of *A*'s token behavior from observing other, similar instances of this behavior (taking for granted some unproblematic way of typing behavior).

function of *x* as it is. In these cases, the set of 'relatively close possible worlds' would amount to the set of possible worlds where the intrinsic properties of *x* are the same as in the actual one. Other things about these possible worlds and, most importantly, the environment *x* is in, can vary. A possible example for such explanatory projects would be to find the function of a seemingly functionless trait by extrapolating environments where this trait does contribute to the organism's inclusive fitness.

In some other explanatory projects, what *x* does or can do in environments different from the present one is irrelevant. In these cases, the set of 'relatively close possible worlds' means the set of possible worlds where the environment is the same as in the actual world. The function of *x*, then, is relative to the environment. Strictly speaking, in such explanatory projects we should talk about the function of *x* in environment *E*—just as the relational theory of function does. Examples where the same trait can do very different things that would contribute to the inclusive fitness of the organism in different environments are possible examples for explanatory projects of this kind.[32]

VII. OBJECTIONS

We need to make sure that the modal theory of function satisfies the three desiderata I enumerated in section II.

Any theory of function needs to be able to explain malfunctioning. As we have seen, a trait malfunctions if and only if it has a function but fails to perform it. This is perfectly possible in my account, since even if *x* is not doing *F* in the actual world, it may still be true that if *x* were performing *F* then this would contribute to the inclusive fitness of the organism that possesses *x*.

The other two desiderata are also satisfied. A trait can have two or more functions, as there may be many things the trait does that

[32] As we have seen, the relational theory of function talks about function relative to a selective regime, that is, to "the total set of abiological and biological (including social, developmental and physiological) factors in the environment of the trait which potentially affect the fitness of individuals with that trait" (Walsh, *op. cit.*, p. 564). Relativizing function to a selective regime may be thought of as being the same as relativizing it to an environment. The relational view is nevertheless different from this special case of my definition of function in two very important respects. First, as we have seen, the relational theory of function defines function in terms of the contribution of *trait types*, whereas my definition does not talk about trait types and defines function entirely in terms of the properties of the token trait. Second, the relational notion of function does not carry any modal force (see especially *ibid.*, section v.1), whereas my notion does.

would contribute to the organism's inclusive fitness. And, as we have seen, the attribution of functions depends on the explanatory project, since the explanatory project determines which nearby possible worlds we should take into consideration when assessing the function of a trait.

Let us see how this proposal can deal with the cases that are problematic for the etiological approach. If the swampman's heart pumped blood then this would contribute to the inclusive fitness of the swampman (this follows from the supposition that the swampman is molecule-by-molecule identical to a human being); hence, the swampman's heart has the function to pump blood, in spite of the fact that he lacks history.

My notion of function is obviously not vulnerable to the trait-type individuation objection, because it does not use trait types when defining function. It defines the function of a token trait entirely in terms of the properties of this token trait. To sum up, if we conceive of function the way I suggested, some of the worrying consequences of the etiological view disappear.

Finally, one could argue that this new theory of function is susceptible to new objections. More precisely, one may worry that this definition does not capture the notion of function, but rather the notion of usefulness.

My response is to bite the bullet: function may have a lot to do with usefulness. But it is important to distinguish usefulness from use. It would indeed be a worrying consequence of my view if it ended up assimilating function to use: to whatever the trait is being used for. But this is not the case. What a trait is being used for is determined by what goes on in the actual world. Function (and, arguably, usefulness), in contrast, depends on what goes on in nearby possible worlds. Function is a modal concept; use is not. As long as we clarify that usefulness is not the same as use and that it should be conceived of as a modal concept, it may not be such a bad idea to claim that function has a lot to do with usefulness.

The main consideration against thinking about function as usefulness is that the notion of function is generally taken to be tied to the notion of design, which is very different from usefulness.[33] As Philip

[33] It has been argued recently that if we conceive of function as usefulness we may avoid some undesirable consequences of conceiving of function as design (Wayne D. Christensen and Mark Bickhard, "The Process Dynamics of Normative Function," *Monist*, LXXXV, 1 (January 2002): 3–28; Richard Cameron, "How to Be a Realist about *Sui Generis* Teleology Yet Feel at Home in the 21st Century," *Monist*, LXXXVII, 1 (January 2004): 72–95).

Kitcher put it, "the function of *S* is what *S* is designed to do."[34] This seems to be a very widely accepted view.[35]

The main motivation for interpreting function as design comes from the artifact case: in the case of artifact function, design fixes function, so, if we want to maintain the continuity between biological and artifact function, we should expect something very similar in the case of biological function. If we manage to point out that even in the artifact case function has little to do with design, then the main motivation for this objection ceases to exist. This is exactly what I intend to do in the next section.

VIII. BACK TO ARTIFACT FUNCTIONS

I outlined a theory of biological function. However, if this theory is correct, then the explanation of biological function is very different from the explanation of artifact functions. Artifact function is fixed by design, whereas biological function is fixed by modal facts. Some would see this as a weakness of my account. One of the attractions of the etiological theory of function was that it could provide a theory of biological function that is continuous with the way we usually explain artifact function.[36]

My response is to say that instead of constructing a theory of biological function that would mirror the standard way of thinking about artifact function, we should reevaluate the standard understanding of artifact function. In short, my claim is that artifact function is not, or at least not always, fixed by design. It is important that this section is not intended to give a full account of artifact function, but rather to attempt to explore the possibility of modifying the modal theory of biological function in such a way that it would cover artifact functions.

[34] Philip Kitcher, "Function and Design," *Midwest Studies in Philosophy*, XVIII, 1 (September 1993): 379–97, at p. 380.

[35] See also Millikan, *Language, Thought and Other Biological Categories*, especially p. 17; and George C. Williams, *Adaptation and Natural Selection* (Princeton: University Press, 1966), especially p. 209. Even those who aim to reconsider the role the notion of design plays in the explanation of biological function (for example, Collin Allen and Marc Bekoff, "Biological Function, Adaptation, and Natural Design," *Philosophy of Science*, LXII, 4 (1995): 609–22; David J. Buller, "Function and Design Revisited," in Ariew, Cummins, and Perlman, eds., *op. cit.*, pp. 222–43) accept a weaker claim that if *x* is designed to do *F*, then the function of *x* is to do *F*.

[36] See Kitcher, *op. cit.*; Beth Preston, "Why Is a Wing Like a Spoon? A Pluralist Theory of Function," this JOURNAL, XCV, 5 (May 1998): 215–54. For a dissenting view, see Pieter E. Vermaas and Wybo Houkes, "Ascribing Functions to Technical Artefacts: A Challenge to Etiological Accounts of Functions," *British Journal for the Philosophy of Science*, LIV, 2 (2003): 261–89; and Houkes and Vermaas, "Actions versus Functions: A Plea for an Alternative Metaphysics of Artifacts," *Monist*, LXXXVII, 1 (January 2004): 52–71.

The slinky was not designed to be used as a toy that can 'walk' downstairs. It was designed to be a tension spring in a horsepower monitor for naval battleships. Nonetheless, its function now is to 'walk' downstairs. Similar examples include truck tires used for football practice and old chalkboards used as dinner tables in some trendy households.[37]

Thus, it is not true of artifacts in general that *x* has function *F* if and only if *x* was designed to do *F*. But then how can we explain artifact function?

My suggestion, not surprisingly, is that function attribution to artifacts also depends on modal facts about the token artifact. Thus, the function(s) of an artifact is fixed by what *would* contribute to the fulfillment of the goals of the agent who is using the artifact. This could be spelled out in the following way: artifact *x* has function *F* at time *t* if and only if some 'relatively close' possible worlds where *x* is doing *F* at *t* and this contributes to the fulfillment of the goals of the agent who is using the artifact are closer to the actual world than any of those possible worlds where *x* is doing *F* at *t* but this does not contribute to the fulfillment of the goals of the agent who is using the artifact.

The function of the slinky is to roll from one step to the other because some 'relatively close' possible worlds where it is rolling from one step to the other and this contributes to the fulfillment of the goals of the agent who is using it are closer to the actual world than any of those possible worlds where it is rolling from one step to the other but this does not contribute to the fulfillment of the goals of the agent who is using it. What it was designed for is irrelevant.

One may be slightly suspicious of the reliance on the notion of 'the fulfillment of the goals of the agent who is using the artifact', so I need to make some explanatory remarks about this notion. What if nobody is using the artifact at the moment? Would it follow that the artifact has no function? No. As we have seen, artifact function is defined by what *would* contribute to the fulfillment of the goals of the agent who is using the artifact. If nothing contributes to the fulfillment of the goals of the agent who is using the artifact in the actual world, say, because nobody is using the artifact, this does not mean that the artifact has no function. Whether it has a function depends not just on what happens in the actual world but also on what would happen if things were different. If some 'relatively close' possible

[37] Would it be possible to consider the person who puts these artifacts to new use as the designer? This would certainly be an option, but in this case design would not explain any properties of the artifact (except for what it is being used for); thus, we would lose the main motivation for comparing function to design.

worlds where someone is using the artifact and what it is doing contributes to the fulfillment of the goals of this agent are closer to the actual world than any of those possible worlds where someone is using the artifact and what it is doing does not contribute to the fulfillment of the goals of this agent, then it does have a function.

A possible worry about this way of thinking about artifact function is the following. I could use my laptop as a doorstop, but does this make it a function of the laptop to serve as a doorstop? If so, then the account of artifact function I outlined here would make the concept of *x*'s function dangerously similar to what *x* is being used for. And every given object could be used for thousands of things. Thus, the danger is that if we accept the account I have been proposing, every object will end up having thousands of functions.

It is important to point out that if I use an object as a doorstop in the actual world then it does not follow under my definition that a function of this object would be to be a doorstop. Again, the function of artifacts is fixed by counterfactual facts. The function of an artifact is not whatever it does that fulfills the goals of the agent who is using it but what it does that *would* contribute to the fulfillment of the goals of this agent.

Again, the definition of artifact function was the following: artifact *x* has function *F* at time *t* if and only if some 'relatively close' possible worlds where *x* is doing *F* at *t* and this contributes to the fulfillment of the goals of the agent who is using the artifact are closer to the actual world than any of those possible worlds where *x* is doing *F* at *t* but this does not contribute to the fulfillment of the goals of the agent who is using the artifact. How does this definition apply in the case of my laptop? In the actual world, my laptop serves as a doorstop, and this contributes to the fulfillment of my goals. But does my laptop's serving as a doorstop contribute to the fulfillment of my goals in nearby possible words? In some, it does; in some it does not. We have no reason to believe that some nearby possible worlds where it does are closer than any possible worlds where it does not. Thus, we have no reason to attribute the function of serving as a doorstop to the laptop. The moral is that if a function of an artifact *x* is to do *F*, it is not enough that I happen to use *x* for *F*-ing in the actual world (or that I could do so). It is not even enough that if things were different, *x* would still be used for *F*-ing. In order for an artifact *x* to have a function to do *F*, it needs to be true that some 'relatively close' possible worlds where *x* is doing *F* and this contributes to the fulfillment of the goals of the agent who is using the artifact are closer to the actual world than any of those possible worlds where *x* is doing *F* but this does not contribute to the fulfillment of the goals of the agent who is using the artifact.

The conclusion is that the symmetry between biological function and artifact function could be restored if we accept a modal theory of function: both the function of artifacts and the function of biological traits are fixed by modal facts. I have only sketched, and not defended, the possibilities of a modal theory of artifact function here. The aim of this paper was to defend a modal theory of biological function.

IX. CONCLUSION

Finally, some readers may be skeptical about the modal theory of function because of the appeal to possible worlds. In conclusion, I find it important to emphasize that the modal theory of function does not presuppose realism about possible worlds; nor does it presuppose the Lewisian analysis of counterfactuals. I used the Lewisian framework because it is the most widespread nowadays and because it allowed me to make explicit some of the fine details of the modal claim. But any other account of counterfactuals could be used to fill in the details of the account. The main claim of the modal theory of function is that function attributions have modal force. This claim could be made with or without relying on possible worlds.

BENCE NANAY

University of Antwerp and University of Cambridge

BRIDGING THE MODAL GAP*

Standard Realism about *x* is the view that *x* exists and is constitutively independent of humans' cognitive responses. De re Essentialism is the view that (i) there is some (nontrivial) property *p* which some actual object *o* has essentially, and (ii) *o*'s so having *p* is a constant, internal feature of *o* which is independent of context and of human ways of viewing *o* and *p*. Kripkean Essentialists endorse both standard Realism and de re Essentialism about ordinary objects (stars, tables, wolves, rocks, paperclips, mountains, portions of water). I argue that a consequence of endorsing standard Realism is that one places oneself in an epistemic position with regard to ordinary objects such that one does not have sufficient access to ordinary objects to say, of some specific ordinary object *o* and some specific (nontrivial) property *p*, that *o* has *p* essentially. Standard Realists typically defend their endorsement of de re Essentialism by reasoning from specific cases. They say, "Here is an actual object, Elizabeth. She has her parentage essentially. That she has her parentage essentially is a constant, internal feature of her which is independent of context. Hence, de re Essentialism is true." If I am correct that a consequence of endorsing standard Realism is that one places oneself in an epistemic position with regard to ordinary objects such that one does not have sufficient access to ordinary objects to say, of some specific ordinary object *o* and some specific (nontrivial) property *p*, that *o* has *p* essentially, then standard Realists cannot use such reasoning to defend their endorsement of de re Essentialism. Since such reasoning is the main way standard Realists defend their endorsement of de re Essentialism, showing that such reasoning is flawed severally cripples the position of the standard Realist who wishes to endorse de re Essentialism.

I. STANDARD REALISM

Standard Realism about *x* is the view that *x* exists and is constitutively independent of humans' cognitive responses. To say that *x* is constitutively independent of *y* is to say that it is no part of what it is to be *x* that *y* be a certain way.[1] For instance, to say that John is constitutively

* Thanks to members of the audiences at the Australian National University (2006) and Pacific APA (2007) for helpful comments and discussion, as well as to Cody Gilmore, Michael Glanzberg, L. A. Paul, and Paul Teller for constant feedback over the years.

[1] The formulation "no part of what it is to be *x*" is from Carrie Jenkins, "Realism and Independence," *American Philosophical Quarterly*, XLII, 3 (2005): 199–209.

0022-362X/10/0708/432–443

independent of Paul is to say that it is no part of what it is to be John that Paul be a certain way. On the other hand, to say that John is constitutively dependent on DNA-structure *s* is to say that it is part of what it is to be John that DNA-structure *s* be a certain way.

Standard Realism concerns a specific kind of constitutive independence, namely, being constitutively independent of humans' cognitive responses. The notion of being constitutively independent of humans' cognitive responses may be more fully explicated if we contrast it with the more familiar notion of being constitutively independent of humans' manipulative responses.[2] An object *o* is constitutively independent of humans' manipulative responses iff it is no part of what it is to be *o* that humans' manipulative responses be a certain way. Natural objects (such as rocks, trees, planets, and wolves) are constitutively independent of humans' manipulative responses; artifacts (such as houses, paperclips, paintings, and bicycles) are not constitutively independent of humans' manipulative responses. Contra constitutive independence of manipulative responses, constitutive independence of cognitive responses concerns not what humans do, but what humans think. An object *o* is constitutively independent of humans' cognitive responses iff it is no part of what it is to be *o* that humans' cognitive responses be a certain way. Amy's friend Sue is typically thought to be constitutively independent of Amy's cognitive responses. That is, it is typically thought to be no part of what it is to be Sue that Amy's cognitive responses be a certain way. The pain in Amy's arm is typically thought to be constitutively dependent on Amy's cognitive responses. That is, it is typically thought to be part of what it is to be a pain in Amy's arm that Amy's cognitive responses be a certain way. Standard Realists about ordinary objects think that ordinary objects are like Amy's friend Sue rather than like the pain in Amy's arm. That is, they think that, for any ordinary object *x*, it is no part of what it is to be *x* that humans' cognitive responses be a certain way.

II. DE RE ESSENTIALISM

De re Essentialism is the view that (i) there is some (nontrivial) property *p* which some actual object *o* has essentially,[3] and (ii) *o*'s so having

[2] We respond manipulatively to the world when we, for example, fit pieces of wood together to create a table, carve a block of stone into a statue of Crazy Horse, or arrange sticks and mud across a stream to build a bridge.

[3] It is notoriously difficult to spell out exactly what is meant by "nontrivial." Suffice it to say, properties such as *being human*, *having mass* m, and *being made of wood* are nontrivial, whereas properties such as *being self-identical*, *having some property or other*, and *being either* q *or* ~q are trivial.

p is a constant, internal feature of *o* which is independent of context and of human ways of viewing *o* and *p*. The first clause is an existence clause. It asserts that there actually exists some object *o* and some property *p* such that *o* has *p* essentially. The second clause concerns the way in which *o* has *p*. In particular, it rules out *o*'s having *p* only in some contexts (or according to some representations, or given certain descriptions). It ensures that it is the nature of *o* itself which makes it the case that *o* has *p*.[4] Kripke describes de re Essentialism as the view that "[properties] can meaningfully be held to be essential...to the object independently of its description."[5] Kit Fine describes de re Essentialism as the view that "there is something in the object itself which sustains a distinction between its accidental and essential properties."[6]

De re Essentialism is a particularly strong version of Essentialism. There are many ways to be an Essentialist without being a de re Essentialist. Most salient among these weaker forms is what I call "Hypothetical Essentialism." Hypothetical Essentialism is de re Essentialism minus the existence clause. De re Essentialism says, "There is water, and it is essentially H_20." Hypothetical Essentialism says, "If there is any water, then it is essentially H_20." The conflict discussed in this paper is between standard Realism and de re Essentialism, not between standard Realism and other forms of Essentialism.

III. KRIPKEAN ESSENTIALISM

The standard Realist's preferred method of arguing that ordinary objects have de re essential properties—that is, by finding a particular ordinary object *o* which has a particular (nontrivial) property *p* essentially—requires us to have sufficient epistemic access to ordinary objects to say that some ordinary object *o* has some property *p* essentially:

Epistemic Requirement

We have sufficient access to ordinary objects to say, of some specific ordinary object *o* and some specific (nontrivial) property *p*, that *o* has *p* essentially.

[4] Strictly speaking, it ensures that it is either the nature of *o* itself which makes it the case that *o* has *p* or it is the nature of *p* itself which makes it the case that *o* has *p*. Perhaps, on some construals of properties and for certain properties, it is plausible to say it is the nature of *p*, rather than the nature of *o*, which makes it the case that *o* has *p*. I suspect, on most construals, such properties will be trivial; for example, perhaps it is the nature of *being* p *or not* p which ensures that all objects possess it. At any rate, since the present concern is to distinguish de re Essentialism from context-dependent forms of Essentialism, such complicating factors can be set aside.

[5] Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980), pp. 39, 41.

[6] Fine, *Modality and Tense* (New York: Oxford, 2005), p. 19.

The individual who endorses Kripkean Essentialism is not, of course, logically compelled to accept the Epistemic Requirement. Kripkean Essentialism says nothing about our access to the world. It merely makes a claim about the way the world is. Nonetheless, the individual who endorses Kripkean Essentialism would be hard-pressed to deny the Epistemic Requirement, as so doing would deprive him of the main way of arguing for Kripkean Essentialism. In practice, those who endorse Kripkean Essentialism do not deny the Epistemic Requirement. In principle, it certainly seems that they are right not to deny it—that Kripkean Essentialism is the sort of controversial view that, to be acceptable, needs to be grounded in arguments from specific cases, that is, that this object has this property essentially.

In the remainder of this paper, I will be arguing that the Epistemic Requirement does not obtain. There are potentially convincing arguments in favor of standard Realism, and there are potentially convincing arguments in favor of de re Essentialism. However, the arguments in favor of standard Realism are convincing only against the assumption that de re Essentialism is false, and the arguments in favor of de re Essentialism are convincing only against the assumption that standard Realism is false. I will be arguing that a consequence of endorsing standard Realism is that one places oneself in an epistemic position with regard to ordinary objects such that one does not have sufficient access to ordinary objects to hold that there is some specific ordinary object *o* and some specific (nontrivial) property *p* such that *o* has *p* essentially. If I am correct, then the Kripkean Essentialist's ability to provide a satisfactory explanation of how we make the leap from the actual (*o* actually has *p*) to the modal (*o* has *p* essentially) has been substantially weakened. Pending the presentation by Kripkean Essentialists of new arguments which do not rely on generalizing from specific cases, we are left to doubt whether the Kripkean Essentialist can successfully bridge the Modal Gap.

IV. KRIPKE AND THE MODAL GAP

Kripke attempts to bridge the Modal Gap via thought experiments. For instance,

> How could a person originating from different parents, from a totally different sperm and egg, be this very woman? One can imagine, given the woman, that various things in her life could have changed... But what is harder to imagine is her being born of different parents. It seems to me that anything coming from a different origin would not be this object.[7]

[7] Kripke, *op. cit.*, p. 113.

> Given that gold does have the atomic number 79, could something be gold without having the atomic number 79? Let's suppose that scientists have investigated the nature of gold and have found that it is part of the very nature of this substance, so to speak, that it have the atomic number 79. Any world in which we imagine a substance which does not have these properties is a world in which we imagine a substance which is not gold, provided these properties form the basis of what the substance is. It will therefore be necessary and not contingent that gold be an element with atomic number 79.[8]

Kripke's story is: We are empirically acquainted with ordinary objects. This allows us to denote them. Once an object has been denoted, we can think about it in counterfactual circumstances. So thinking, that is, running thought experiments, allows us to figure out whether or not object *o* would have property *p* in circumstance *c*. Having figured this out, we are in a position to judge whether or not *o* has *p* essentially. Kripke's thought experiments only carry evidential weight if we have some reason to believe that our modal intuitions trace modal facts. My concern is that if we accept the Kripkean Essentialist picture that ordinary objects exist in the world independently of us (that is, standard Realism) and have properties essentially independent of contexts of description (that is, de re Essentialism), then we have no reason to think our essentialist intuitions trace modal facts. The literature is replete with articles arguing that there is, indeed, a link between modal intuition/thought experiments and modal facts.[9] I will focus on one of the most thorough attempts to provide such a link: George Bealer's Modal Reliabilism.[10] I will argue that Bealer's attempt fails. Moreover, my arguments against Bealer generalize to any such link one might try to create on behalf of the Kripkean Essentialist. Bealer's failure to bridge the Modal Gap is not due to any idiosyncrasy of Bealer's theory, but rather is due to the underlying metaphysics—standard Realism plus de re Essentialism—of the view he is trying to defend. If I am right, then the Kripkean Essentialists' standard method of denoting a specific object and claiming it has a (nontrivial) de re essential property does not provide a satisfactory answer to the Modal Gap problem. In other words, if I am right, then the Kripkean

[8] *Ibid.*, p. 123–25.

[9] See, for instance, David Chalmers, "Does Conceivability Entail Possibility?" in Tamar Szabo Gendler and John Hawthorne, eds., *Conceivability and Possibility* (New York: Oxford, 2002), pp. 145–200; Stephen Yablo, "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research*, LIII, 1 (1993): 1–42.

[10] Bealer, "The Philosophical Limits of Scientific Essentialism," *Philosophical Perspectives*, I (1987): 289–365, and "Modal Epistemology and the Rationalist Renaissance" in Gendler and Hawthorne, eds., *op. cit.*, pp. 71–125.

Essentialists will have to find some new epistemological story to tell to defend their mutual endorsement of standard Realism and de re Essentialism.

V. BEALER'S THEORY: CATEGORY CONCEPTS AND NATURALISTIC CONCEPTS

Bealer thinks the best way to argue for the essentialist claims that Kripkean Essentialists endorse is to have the claims as conclusions of arguments which have (i) a general a priori premise which contains the essentiality, (ii) a linking premise which subsumes a naturalistic concept under a category concept, and (iii) an a posteriori premise which contains the specific empirical facts relevant to the essentialist claim.

The general a priori premises all involve category concepts. Category concepts are a priori concepts that concern how we categorize the world at the most fundamental level. Bealer gives the following examples of category concepts: identity, property, proposition, quality, compositional stuff, functional stuff.[11] Bealer argues that we determinately possess a category concept iff we possess the capacity (in cognitively ideal circumstances) to correctly apply the concept to hypothetical cases characterized exclusively in terms of other a priori concepts.[12] For instance, I determinately possess the category concept *compositional stuff* iff, were I to be in cognitively ideal circumstances and were the test objects to be described to me only in terms of a priori concepts, I would apply the concept *compositional stuff* only to those objects which really are compositional stuffs. If I determinately possess the concept *compositional stuff* and objects o_1, o_2, and o_3 are all described to me only in terms of a priori concepts, then (in cognitively ideal circumstances) I will be able to judge correctly whether or not compositional stuff o_1 could be composed of o_2 rather than of o_3. Determinately possessing a category concept is, thus, innately truth-tracking. It follows just from the fact that I determinately possess the category concept *compositional stuff* that (in cognitively ideal circumstances) I will answer "yes" to a modal question concerning the category concept *compositional stuff* iff the answer to the modal question really is "yes." In general, for any category concept c, if I determinately possess category concept c then (in cognitively ideal circumstances) I will answer "yes" to a modal question concerning category concept c iff the answer to the modal question really is "yes." Bealer's a priori premise thus not only introduces the modality

[11] Bealer, "Modal Epistemology and the Rationalist Renaissance," p. 106.
[12] Bealer, "The Philosophical Limits of Scientific Essentialism," p. 354.

which will be present in the conclusion, but also builds in a guarantee that our modal judgments (in appropriately ideal circumstances) are correct.

The linking premises all involve subsuming naturalistic concepts under category concepts.[13] Naturalistic concepts are concepts concerning natural kinds, for example, *water*, *gold*, *bear*. Bealer endorses a Moderate Causal theory of naturalistic concept possession according to which we determinately possess a naturalistic concept iff (i) we have learned it via a Kripkean causal chain, and (ii) our thought and talk about the concept is mediated by a background of appropriate category concepts.[14] Mediating our thought and talk about naturalistic concepts via category concepts ensures that when we denote an object of a certain natural kind, we are determinately denoting an object to which the relevant naturalistic concept applies. For example, the naturalistic concept *water* latches onto something with a certain composition (H_20) rather than to something with a certain function (drinkable liquid) because the naturalistic concept *water* is mediated via the category concept *compositional substance*.[15] In other words, mediating our thought and talk about naturalistic concepts via category concepts ensures that just one modal profile goes with each naturalistic concept.[16] Bealer, thus, tries to ensure that our determinate possession of naturalistic concepts is as infallible as our determinate possession of category concepts. Part of what it is to determinately possess the naturalistic concept *water*, and, hence,

[13] The following would count as linking premises according to Bealer: "Portions of water are compositional substances," and "Gold is an elemental substance." The former subsumes the naturalistic concept *water* under the category concept *compositional substance*. The latter subsumes the naturalistic concept *gold* under the category concept *elemental substance*.

[14] Regarding (i), "to possess a concept (of the right sort) a person need only be properly situated in the world; in particular, the person need only bear appropriate causal (historical, socio-linguistic) relations to items in the world," in Bealer, "The Philosophical Limits of Scientific Essentialism," p. 305. Regarding (ii), "causal theories are unsatisfactory unless supplemented with the theory that our thought and talk about naturalistic items is mediated by a background of appropriate category and content concepts for which the circumscribed rationalist theory of determinateness holds," *ibid.*, p. 349.

[15] *Ibid.*, p. 351.

[16] One might worry that the process is now entirely a priori. If I determinately possess naturalistic concept *c*, then—since my determinate possession involves knowing under which category concepts the naturalistic concept falls—have I not gotten the modality entirely a priori, and thus not gotten the necessary a posteriori claims the Kripkean Essentialist endorses? The reason the process is not entirely a priori is because determinate possession of the naturalistic concept also requires standing in a Kripkean chain to an actual object. So, one can say, "If this is water, then it has its chemical composition essentially," and this is all a priori. But saying "this is water" requires being suitably situated with respect to some object, and this suitable situation is a posteriori.

to have the capacity to determinately denote water, is to know that whatever it is that counts as water also counts as a compositional substance.[17]

The a posteriori premises which support essentialist conclusions are simply drawn from empirical science, for example, "All actual water is composed of H_2O," and "All actual gold has the atomic number 79." We use empirical investigation to ascertain the truth of a posteriori premises.

A brief recap is in order. Bealer's suggestion for bridging the Modal Gap is to have de re essentialist claims as the conclusions of arguments and to argue for each premise, rather than arguing directly for the de re essentialist claims. The first premise will be a general a priori premise which contains the essentiality. Since we can figure out—via ideal rational reflection—what the essential relations between a priori concepts are, we can tell that the first premise is true. The second premise will be a linking premise which subsumes a naturalistic concept under a category concept. Since we can figure out—via ideal rational reflection—which naturalistic concepts fall under which category concepts, we can tell that the second premise is true. The third premise will be an a posteriori premise which contains the specific empirical facts relevant to the essentialist claim. We use empirical investigation to ascertain the truth of the third premise. Thus, we have three premises—each of which appears to be such that we can tell it is true. Moreover, we are able to construct a valid argument which leads from the three premises to a de re essentialist claim. All of this takes place against a background endorsement of standard Realism. Hence, apparently one can successfully argue both for standard Realism and for de re Essentialism. In other words, if Bealer's strategy is successful, the Kripkean Essentialist can bridge the Modal Gap.

[17] Prima facie, there is a tension between what the Moderate Causal theory says regarding possessing naturalistic concepts and our experience of possessing naturalistic concepts. If the Moderate Causal theory is correct, then determinately possessing a naturalistic concept requires a great deal. Namely, knowing which category concepts mediate the application of the naturalistic concept to objects in the world. Yet, it seems that we can be "deeply ignorant about whether a given naturalistic concept applies to relevant hypothetical cases" (that is, whether something composed of XYZ falls under the naturalistic concept *water*) and yet still have beliefs about water, that is, about an object which we have not yet picked out (Bealer, "The Philosophical Limits of Scientific Essentialism," p. 354.). Bealer just bites the bullet here and says that we can have cognitive commitments to propositions "while being deeply ignorant about the essential properties and relations of the concepts involved" (Bealer, "Modal Epistemology and the Rationalist Renaissance," p. 109). In other words, most of the time, we get by just fine merely possessing naturalistic concepts. It is only when we turn to modalizing that we need to determinately possess them.

VI. BEALER'S THEORY REVISITED

Let us look at how Bealer's strategy works for a specific example:

P1: Compositional substances have their compositions essentially.	(a priori premise)
P2: Object *o* is a portion of water and, hence, is a compositional substance.	(linking premise)
P3: Object *o* is, in fact, composed of H_20.	(a posteriori premise)
C1: Object *o* is essentially composed of H_20.	(de re Essentialist claim)

The structure of Bealer's argument can be seen more clearly if we separate the a priori part of the linking premise from the a posteriori part of the linking premise:

P1: Compositional substances have their compositions essentially.	(a priori premise)
P2a: Portions of water are compositional substances.	(linking premise, a priori)
P2b: Object *o* is a portion of water.	(linking premise, a posteriori)
P3: Object *o* is, in fact, composed of H_20.	(a posteriori premise)
C1: Object *o* is essentially composed of H_20.	(de re Essentialist claim)

Ascertaining the truth-value of premises P1, P2a, and P3 is relatively straightforward. By reflecting on the category concept *compositional substance*, we figure out a priori that P1 is true. Likewise, by reflecting on the naturalistic concept *water* and seeing that it is just part of what it is to be water that one be a compositional substance, we figure out a priori that P2a is true. By getting out our scientific instruments and looking, we figure out a posteriori that P3 is true. The difficulty arises when we try to figure out the truth-value of P2b. In order for P2b to be true, it must be the case that object *o* is a portion of water. This is something that, contra Bealer, we cannot ascertain a posteriori. We can ascertain that object *o* is, in fact, composed of H_20, that it is a clear, odorless liquid, and that it has all of the properties we standardly associate with water. But we cannot tell whether object *o* is water (that is, an object which is essentially composed of H_20) or swater (that is, an object which is only accidentally composed of H_20). Hence, there is no way for us to tell whether or not P2b is true.

Given the full empirical investigation that has been done (that is, given P3), we want to say, "Of course it is water. Look!" But this begs the question. Since P1 (all compositional substances have their composition essentially) and P2a (portions of water are compositional substances) are both true, the bar concerning what is required for an object to be a portion of water (that is, what is required for P2b to be

true) has been raised. To take Bealer as having shown there are true de re Essentialist claims (that object there is essentially H_20), there needs to be some way to tie Bealer's a priori reasoning to actual objects. This job is supposed to be done by the linking premises. But the linking premises only link the a priori with actual objects if we have some antecedent reason to think that some of the actual objects fall under naturalistic concepts such as *water*. Once we understand that falling under such a naturalistic concept requires more than simply actually having certain properties (for example, actually being composed of H_20), we no longer have any reason to think any actual objects fall under any of Bealer's naturalistic concepts. This is not to say, of course, that no actual objects fall under any naturalistic concepts. There may very well be some actual object which is essentially H_20 and which, hence, falls under the naturalistic concept *water*. It is merely to say that Bealer has given us no reason to think there are such objects.[18] Bealer's strategy of arguing for de re Essentialist claims by placing them as conclusions of arguments which contain premises that are either purely a priori or purely a posteriori fails because Bealer is unable to find a satisfactory way to tie his a priori reasoning to objects in the actual world. The Modal Gap remains: how do we make the leap from claiming that object *o* actually has property *p* to claiming that object *o* has property *p* essentially?

VII. GENERALIZING BEALER

My argument that Bealer fails to bridge the Modal Gap would be of little general interest were it not the case that Bealer's failure generalizes. It is the underlying metaphysics, standard Realism plus de re Essentialism, which gives rise to the Modal Gap. If standard Realism for ordinary objects is true, then the existence of ordinary objects is constitutively independent of us. We may play a manipulative role in bringing some of them into existence (as, for example, when we fit pieces of wood together to create a table), but we do not play a constitutive role in their existence: it is no part of what it is for an object to, for example, be a portion of water that we have certain responses. Given this, what can we know about an object which is purportedly a portion of water? We can have empirical knowledge of it, namely, that

[18] One might wonder if the de re Essentialist can just say, "(i) *o* is actually composed of H_20, (ii) any thing which is actually composed of H_20 is essentially composed of H_20, and, hence, (iii) *o* is essentially composed of H_20." He cannot say this, as doing so just raises the very question at issue. The difficulty is with the second clause. Whatever concerns arise regarding the Kripkean Essentialists' grounding of "*o* is essentially composed of H_20," arise equally for the claim that "any thing which is actually composed of H_20 is essentially composed of H_20." What reason have we for thinking that any thing which is actually composed of H_20 is essentially composed of H_20?

it is at location *L*, that it weights *k* kilograms, and that it is not widely scattered. But, in order to assert de re Essentialism of it, we must be able to have more than mere empirical knowledge of it—we must have modal knowledge of it. The standard Realist who wishes to endorse de re Essentialism, thus, must tell some story which leads from knowledge of an object's (empirically accessible) nonmodal properties to knowledge of its (nonempirically accessible) de re essential properties.

Bealer's goal—and the goal of others who, like Bealer, see their project as clarifying, against a background of standard Realism, when our intuitions that *o* has *p* essentially are legitimate—is to provide a bridge from our concepts to the external world. But given (i) that this external world is constitutively independent of our cognitive responses (for example, we cannot make object *o* fall under the naturalistic concept *water* just by responding as though it does) and (ii) that we cannot distinguish an external world which contains an object *o* which has *p* essentially from a world which contains an object *o** which is perceptually identical to *o* but which has *p* only accidentally, such a bridge can never be provided. The fact that the external world is constitutively independent of us ensures that our merely conceiving *o* to have *p* essentially in no way causes it to be the case that *o* has *p* essentially—so we cannot create a world in which *o* has *p* essentially. The fact that we cannot distinguish an external world which contains object *o* from a world which contains object *o** ensures that we cannot discover that *o* has *p* essentially. But if the Kripkean Essentialist can neither create nor discover a link between our conceiving that *o* has *p* essentially and *o*'s having *p* essentially, then he cannot bridge the Modal Gap. Bealer's failure to bridge the Modal Gap, hence, generalizes to all Kripkean Essentialists.

VIII. WHAT NOW?

Suppose I am right that the Kripkean Essentialist cannot bridge the Modal Gap. What, then, is he to do? He might simply go on endorsing Kripkean Essentialism, being more enamoured by its intuitive plausibility than by any epistemological arguments against it. Alternatively, he might give up either standard Realism or de re Essentialism.

One might think giving up de re Essentialism is not so bad. After all, de re Essentialism is quite a strong position. Perhaps everything the Essentialist ought to want to capture can be captured with less. Nothing I have said prevents the standard Realist from making the hypothetical essentialist claim that if there is any water, then it is essentially H_20. My arguments address only his ability to endorse both (i) if there is any water, then it is essentially H_20, and (ii) there

is, in fact, water. Hypothetical Essentialism, however, is a view which will satisfy few. Those of us who want to endorse Essentialism want to show that there are, in fact, objects which have properties essentially. We thus find Hypothetical Essentialism unsatisfactory.

Alternatively, the Kripkean Essentialist might give up standard Realism and hold on to de re Essentialism. So doing would give him the resources to block the Modal Gap. We are not, as the standard Realist claims, in the business of trying to ascertain the essential natures of independently existing objects. Rather, we are in the business of creating—given certain distributions of matter in space-time—objects whose essential natures match our concepts. This is the option I favor, but that is another paper.

DANA GOSWICK

University of Melbourne

New from

OXFORD

SKETCH FOR A SYSTEMATIC METAPHYSICS

D. M. ARMSTRONG

D.M. Armstrong sets out his metaphysical system in a set of concise and lively chapters each dealing with one aspect of the world. On the basis of the assumption that all that exists is the physical world of space-time, he constructs a coherent metaphysical scheme that gives plausible answers to many of the great problems of metaphysics.

2010 128 pp. $35.00

ART IN THREE DIMENSIONS

NOËL CARROLL

A collection of essays by a leading figure in the philosophy of art. The animating idea behind Carroll's work is that philosophers of art should look beyond aesthetics and refocus their attention on the ways in which art enters the life of culture and influences the moral and emotional experiences of audience members.

2010 536 pp. $74.00

OXFORD STUDIES IN EARLY MODERN PHILOSOPHY

Volume V

Edited by DANIEL GARBER and STEVEN NADLER

A selection of the best current work in the history of early modern philosophy with a focus on the seventeenth and eighteenth centuries.

2010 280 pp. Hardback $99.00 Paperback $35.00

OXFORD STUDIES IN ANCIENT PHILOSOPHY

Volume 38

Edited by BRAD INWOOD

A volume of original articles on all aspects of ancient philosophy.

(Oxford Studies in Ancient Philosophy)

2010 320 pp. Hardback $99.00 Paperback $40.00

New in Paperback

LOT 2

The Language of Thought Revisited

JERRY A. FODOR

"It is a rare contemporary philosopher whom one looks forward to reading. Fodor is such an exception...long may Fodor enlighten and entertain us."—John Collins, *The Philosophers' Magazine*

2010 240 pp. Paperback $24.95

New in Paperback

ROUSSEAU'S THEODICY OF SELF-LOVE

Evil, Rationality, and the Drive for Recognition

FREDERICK NEUHOUSER

Jean-Jacques Rousseau revolutionized our understanding of ourselves with his brilliant investigation of *amour propre*: the passion that drives humans to seek the esteem, approval, admiration, or love — the recognition — of their fellow beings. Frederick Neuhouser traces the development of this key idea in modern thought.

2010 296 pp. Paperback $29.95

MANY WORLDS?

Everett, Quantum Theory, and Reality

Edited by SIMON SAUNDERS, JONATHAN BARRETT, ADRIAN KENT, and DAVID WALLACE

What follows when quantum theory is applied to the whole universe? This is one of the greatest puzzles of modern science. Philosophers and physicists here debate the Everett interpretation of quantum mechanics.

2010 528 pp. $99.00

OXFORD STUDIES IN METAETHICS

Volume 5

Edited by RUSS SHAFER-LANDAU

Oxford Studies in Metaethics is the only publication devoted exclusively to original philosophical work in the foundations of ethics. It provides an annual selection of much of the best new scholarship being done in the field.

2010 320 pp. Hardback $99.00 Paperback $40.00

FREEDOM AND BELIEF

Revised Edition

GALEN STRAWSON

This is a revised and updated edition of Galen Strawson's groundbreaking first book, where he argues that in a fundamental sense there is no such thing as free will or true moral responsibility.

2010 320 pp. Hardback $99.00 Paperback $35.00

DESIRE, PRACTICAL REASON, AND THE GOOD

Edited by SERGIO TENENBAUM

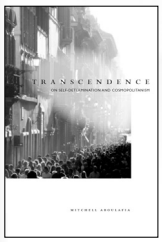
The "Guise of the Good" thesis — the view that desire, intention, or action always aims at the good — has received renewed attention in the last twenty years. The book brings together work on various issues related to this thesis both from contemporary and historical perspectives.

2010 264 pp. $65.00

Prices are subject to change and apply only in the US. To order or for more information, go to **www.oup.com/us**.

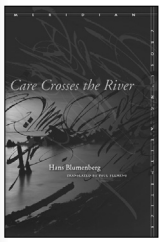
OXFORD UNIVERSITY PRESS

New from Stanford University Press



Transcendence
On Self-Determination and Cosmopolitanism

MITCHELL ABOULAFIA
$21.95 paper $60.00 cloth

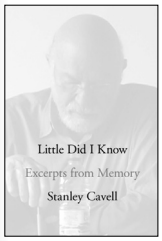


Care Crosses the River

HANS BLUMENBERG
Translated by
PAUL FLEMING
$21.95 paper $60.00 cloth

Little Did I Know
Excerpts from Memory

STANLEY CAVELL
$34.95 cloth

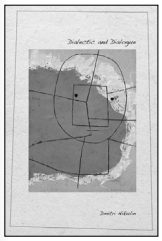


Dialectic and Dialogue

DMITRI NIKULIN
$19.95 paper $55.00 cloth

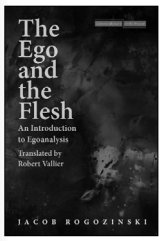
The Ego and the Flesh
An Introduction to Egoanalysis

JACOB ROGOZINSKI
Translated by
ROBERT VALLIER
$29.95 paper $80.00 cloth



Rawls and Habermas
Reason, Pluralism, and the Claims of Political Philosophy

TODD HEDRICK
$24.95 paper $70.00 cloth



Heidegger Among the Sculptors
Body, Space, and the Art of Dwelling

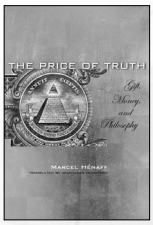
ANDREW J. MITCHELL
$18.95 paper $50.00 cloth

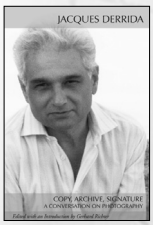
The Price of Truth
Gift, Money, and Philosophy

MARCEL HÉNAFF
Translated by
JEAN-LOUIS MORHANGE
$29.95 paper $80.00 cloth

Copy, Archive, Signature
A Conversation on Photography

JACQUES DERRIDA
Edited and with an Introduction by
GERHARD RICHTER
Translated by
JEFF FORT
$16.95 paper $45.00 cloth





Stanford University Press

800.621.2736 www.sup.org