THOUGHT EXPERIMENTS BY STEVIN, MACH AND GOUY: THOUGHT EXPERIMENTS AS IDEAL LIMITS AND AS SEMANTIC DOMAINS

Ronald Laymon

THE principal question to be considered is how thought experiments that postulate counterfactual situations can be informative about the physical laws that govern this world. Sometimes such thought experiments can be understood as being about the ideal limits of real experimentation. I shall investigate what such a claim might mean and how understanding thought experiments this way yields an answer to our question. I shall also discuss the natural relation that exists between the development of counterfactual thought experiments, so conceived, and the development of real experiments. Sometimes thought experiments are better understood as being implicitly semantic arguments whose conclusions deal with the logical properties of theories. So I shall briefly develop such an account and use Mach's criticisms of Newton's bucket experiment as an illustration.

I. Introductory Example: Stevin and the Law of Equilibrium

The presentation of many thought experiments can be understood as containing arguments of the form,

$$\exists x(Tx) \;\&\; P_1 \;\&\; P_2 \;\&\ldots P_n \twoheadrightarrow Q$$

where $\exists x(Tx)$ is a highly idealized experimental description, $P_1, P_2,\ldots P_n$ are laws or principles believed true, and Q is to be demonstrated. The symbol $\twoheadrightarrow$ is to be given the usual handwaving sense it has in ordinary mathematical practice: if I had time and interest I could think of suitable premises which when conjoined with the antecedent of $\twoheadrightarrow$ would logically yield the consequent.[1] Our use of this operator is meant to capture the fact that in scientific contexts the argumentation associated with thought experiments is never very explicit. We shall call the above expression the *initial argumentation* of the thought experiment. A basic problem with such thought experiments is that the initial argumentation is unsound, since $\exists x(Tx)$, being highly idealized, is false. One of our aims is to develop natural transformations of the initial argumentation, so that what results is an acceptable demonstration of Q. It will make for a clearer and more compact style if we formally define our subject of interest.

> A *thought experiment* is an ordered pair $<\Phi,\vartheta>$ where Φ is a set of persons (audience and/or presenter) and ϑ is a set of statements $\{T, P_1, P_2, \ldots P_n, Q\}$ where:
>
> (1) T is a description that is not in fact true (because it is idealized) of any experiment in this world.
>
> (2) Members of Φ believe that P_1, P_2,...P_n are scientific laws or principles.
>
> (3) Members of Φ believe that $\exists x(Tx)$ & P_1 P_2 &...$P_n \rightarrow Q$.

Let me emphasize that I am not trying to specify conditions that can be used to specify ordinary scientific usage of the expression "thought experiment." My aim is to mark off a natural scientific practice that is of scientific importance and philosophical interest. Since abstract definitions by themselves tend to be neither interesting nor helpful, we move to a motivating example, a thought experiment originally conceived by the medieval theorist Stevin and popularized by Mach in his *Science of Mechanics* (1960, pp. 32-41). The problem is to ascertain the weight required to keep in equilibrium some fixed weight situated on an inclined plane. The situation is illustrated in figure 1, where W is the weight to be kept in equilibrium, and W' is the weight to be determined.

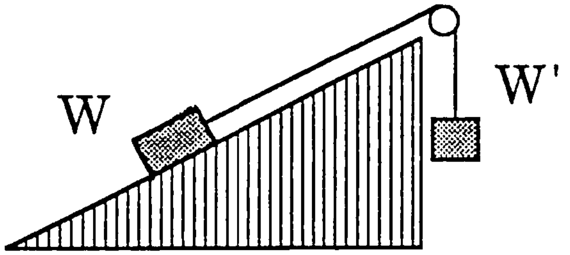


Figure 1

The thought experiment begins by imaging a right-angled triangle ABC oriented so that its hypotenuse AB is parallel to the ground. Actually, we'll have to give our triangle some thickness, turning it into a prism, so that we can hang on it a loop of rope with fourteen or so balls tied so as to be equidistant from one another. (See figure 2.) Since this is a thought experiment we are free to imagine the idealized situation where all impediments to motion are removed. While we are at it, we will want the rope to be totally flexible. Now the system will be in equilibrium or it will not be in equilibrium, i.e., the rope and balls will or will not be in motion. If the latter, the motion will continue forever since there are no reasons for it to stop. But thought experiment or not, this cannot be, or so claims Stevin, since we would have a perpetual motion machine. Therefore, the rope and balls will be in equilibrium. Modern audiences may well hesitate at this step since it requires an Aristotelian concept of motion. To keep the discussion going one should therefore assume such a concept, or modify the experiment so that all frictional forces are eliminated and the apparatus gently brought to rest. Assuming then that our system is at rest, we next consider that part of the rope which hangs underneath the prism. Since it is totally flexible, it will be symmetrically oriented. Therefore the equilibrium of that part of the rope and balls that is on the prism will not be disturbed if we remove the hanging part of the rope. Now nothing we have said or done would have been different if we had arranged our triangle so that side *AC* was parallel to the ground. (Figure 3.) (Remember our rope is totally flexible.) But now we have the answer to our original question since it is obvious that the ratio between *W* and *W′* will be directly as the length *AB* to the length *BC*.

Figure 2

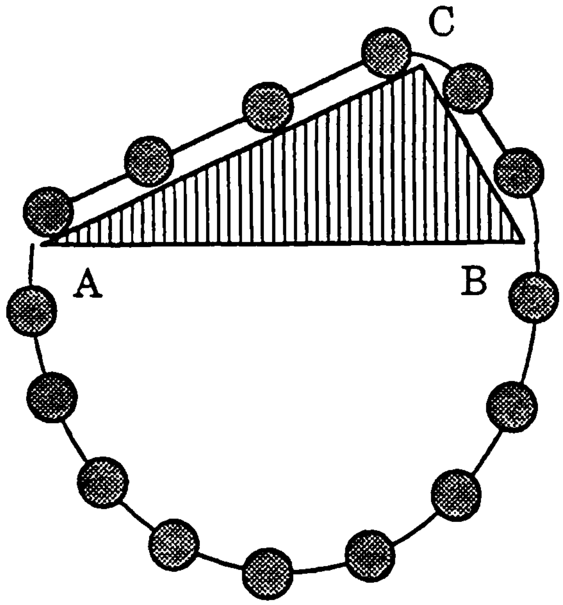
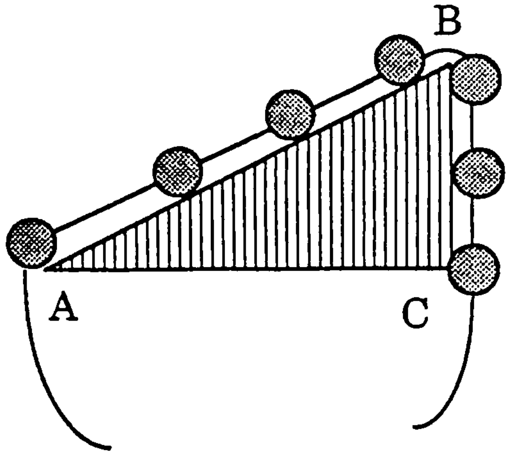


Figure 3



It's worth quoting Mach's appraisal of this thought experiment since he captures very well the immediacy of its point.

> ...we accept the conclusion drawn...without the thought of an objection, although the law if presented as the simple result of experiment...would appear dubious. (Mach 1960, p. 34)

Mach with great perception points to an interesting puzzle, namely, to explain why the conclusion follows with such immediacy from the thought experiment and why, by contrast, the conclusion is so difficult to demonstrate on the basis of actual experimentation. Mach's initial suggestion is certainly on the right track:

> We cannot be surprised at this when we reflect that all results of [actual] experiment are obscured by adventitious circumstances (as friction, etc.), and that every conjecture as to the conditions which are determinative in a given case is liable to error. (Mach 1960, p. 34)

While this explains the difficulty of using real experiments to demonstrate or generate mechanical laws, it does not explain the apparent success and psychologically compelling nature of the thought experiment. As a first step toward such an explanation we shall isolate and make somewhat more explicit what we have called the initial argumentation of Stevin's thought experiment. The following will be the basic components:

$Tx = x$ is a situation of the sort described by Stevin, namely, one where there's a prism with hypotenuse parallel to the ground, with a rope wrapped around it that has equally spaced balls attached and that is friction free, &c.

$Ex = x$ is in equilibrium

$Px = x$ is a situation where there is perpetual motion (of Aristotelian type)

$L =$ the equilibrium law, i.e. that $W/W' = AB/BC$

The first premise of our reconstruction then will be simply: $\exists x(Tx)$. For convenience we will introduce a name for Stevin's combination of prism and rope, namely, a. More formally, we make our first premise: $\exists x(Tx \,\&\, x = a)$. The next premise is just an instance of a tautology, namely, that $Ea \vee {\sim} Ea$. Finally, we need the denial of perpetual motion machines: ${\sim}\exists x(Px)$. So our three premises are:

(1) $\exists x(Tx \,\&\, x = a)$

(2) $Ea \vee {\sim} Ea$

(3) ${\sim}\exists x(Px)$

Stevin's argument that the rope is in equilibrium can be represented as:

(4) $\exists x(Tx \,\&\, x = a) \,\&\, {\sim}\exists x(Px) \,\&\, (Ea \vee {\sim} Ea) \rightarrow Ea$ (claimed logical fact)

(5) Ea (by 1, 2, 3, 4 and some logical cousin of *modus ponens*)

To get the equilibrium law:

(6) $\exists x(Tx \,\&\, x = a \,\&\, Ea) \rightarrow L$ (claimed logical fact)

(7) L (by 1, 5, 6 and *modus ponens*)

While not very elegant, our reconstruction has the virtue of not doing great injustice to the surface grammar of the presentation of the thought experiment. It is also sufficiently precise to allow us to place our problem on the table, which is that the above argument, while conceivably valid, is *unsound* because of the falsity of the first premise, i.e., that $\exists x(Tx)$. After all, we began by agreeing that our thought experiment start with the postulation of a highly idealized situation. Therefore, since the argument is unsound, it cannot be taken as demonstrating the truth of its conclusion. This, I submit, is a fairly major problem with the argumentation typically given to justify conclusions drawn from thought experiments. I think also that the unsoundness of the argument goes some way toward explaining our general discomfort with thought experiments.

The argumentation associated with thought experiments frequently can be naturally represented as being a *reductio ad absurdum*. It might be thought that this argument form will somehow allow escape from the unsoundness problem.[2] The middle stage of Stevin's argument, for example, could have been represented as: $\exists x(Tx \,\&\, x = a) \,\&\, {\sim}\exists y(Py) \,\&\, {\sim}Ea \Rightarrow Pa \,\&\, {\sim}Pa$. This reconstruction can also be made to apply with slight adaptation to a modern example we shall be discussing later, Einstein's blackbody radiation thought experiment. Einstein imagines a mirror free to move in an environment consisting of a gas and blackbody radiation. His aim is to show that blackbody radiation pressure fluctuates. Let Bx mean x is an environment consisting of a gas and blackbody radiation, and let Tx mean x contains a mirror free to move (perpendicularly to its surface). Einstein's idealized thought experiment therefore postulates that: $\exists x(Bx \,\&\, Tx)$. Let Py mean that y is a perpetual motion machine of the *second* type, and Ex mean that x is a situation where there is fluctuation of the radiation pressure. Einstein's argument can be represented as a *reductio* of the form:

$$\exists x(Bx \,\&\, Tx) \,\&\, {\sim}\exists y(Py) \,\&\, (x)(Bx \supset {\sim}Ex) \Rightarrow \exists x(Px \,\&\, {\sim}Px)$$

Now what follows is that ${\sim}[\exists x(Bx \,\&\, Tx) \,\&\, {\sim}\exists y(Py) \,\&\, (x)(Bx \supset {\sim}Ex)]$. Since we believe that ${\sim}\exists y(Py)$ (at least on the macroscopic level), the conclusion of the *reductio* simplifies to ${\sim}[\exists x(Bx \,\&\, Tx) \,\&\, (x)(Bx \supset {\sim}Ex)]$. But here the argument stops, since we already know that $\exists x(Bx \,\&\, Tx)$ is false. That is, we cannot, on this reconstruction draw the desired conclusion, namely, that ${\sim}(x)(Bx \supset {\sim}Ex)$. Therefore, the induction needed to get $(x)(Bx \supset Ex)$ cannot even be started.

Perhaps our procedure has been overly syntactic. It might be better to view the first premise of the initial argumentation as asserting the existence of the thought experiment as a model or interpretation for various object language claims and principles. On this account, what Stevin's thought experiment shows is that there exists a model or interpretation such that the sentence ${\sim}\exists yPy \,\&\, {\sim}Ea$ is false on that interpretation. This may be nice to know, but we wonder about its significance.[3] Full blown semantic inconsistency requires falsity in all models. But surely we want our scientific laws and principles to be inconsistent for some interpretations or possible worlds, otherwise they would be logically true. Therefore, it should not be surprising that we can invent counterfactual situations that provide such interpretations. Once again we are stymied in our attempt to render thought experimentation respectable.

Heavy idealization in a thought experiment is both a strength and a weakness. The strength comes from the freedom it provides to eliminate complicating features, thus rendering the argumentation more explicit, precise and complete than it would be otherwise. But because the assertion of the existence of such situations is strictly false, the argumentation will be either unsound or will show only that there exists a model where a conjunction of principles can be interpreted as being false. What is required is some procedure that will render the falsity of the idealizations benign. A natural proposal is that idealizations can be rendered benign if one can show that real experiments can be refined and thereby made to approach the situations postulated. Let me illustrate in terms of Stevin's experiment what I have in mind. An immediate objection one might make to Stevin's argument is that the measurement units afforded by the equally spaced balls are too crude to support the desired conclusion. Since we are dealing with a thought experiment, we are of course free to use finer and more closely spaced balls. But such successive refinements can also be applied to any real version of the experiment. Similarly, if one objects that that part of the rope which hangs under the prism is not in fact symmetrically displaced (and especially not so in the case where one side of the prism is vertical), we can extend the rope in length or increase its flexibility. With respect to friction one can easily imagine a series of refinements which would allow any real experiment to ever more closely approximate the ideal friction-free thought experiment. Our experimental refinements in combination generate a series of real experiments (or imagined but truly possible experiments) where each successive experiment more closely approximates the initial conditions of Stevin's thought experiment. It is important to note that these refinements are quite natural and would occur to virtually any audience contemplating Mach's (though perhaps not Stevin's) presentation of the experiment. On the other hand, it must also be kept in mind that while these refinements are immediate and obvious, a rigorous enumeration was not in fact presented by Stevin or by Mach. The lack of explicit argumentation justifying the use of idealized and hence counterfactual situations is a common feature of thought experiments. This lack indicates that if they are to be convincing, such thought experiments must be presented to audiences that can be expected to imagine *for themselves* the sorts of experimental refinements we propose. Or audiences must believe on the basis of their experience with what they take to be analogous experiments that such refinements can be developed. A thought experiment, we expect, can sometimes be construed as being an *invitation* to construct arguments that show its relevance for claims about this world.

While it is intuitively appealing to require that there be real experiments that approximate thought experiments, it is not immediately apparent how the existence of such experiments legitimizes the use of idealized premises. Perhaps we should *transform* the first premise of the initial argumentation into something modal whose truth is supported by the existence of real experiments that approximate the thought experiment. This carries some cost since modal logic is a rather subtle and refined subject. But fortunately for present purposes, we shall require only some very basic moves. So, for example, using $\Diamond$ to denote physical possibility, we might transform the middle stage of Stevin's argument into:

1. $\Diamond\exists x(Tx \,\&\, {\sim}Ex) \rightarrow \Diamond\exists x(Px)$

2. But since $\sim\Diamond\exists x(Px)$, it follows that $\sim\Diamond\exists x(Tx \,\&\, \sim Ex)$.

3. By experimental refinement we know that $\Diamond\exists x(Tx)$.

4. Assuming that $\supset x(Ex \vee \sim Ex)$, it follows that $\Diamond\exists x(Tx \,\&\, Ex)$.

So far, so good. However, getting from $\Diamond\exists x(Tx \,\&\, Ex)$ to the desired conclusion L will require some appropriately modalized principle of induction.[4] But looking for such a principle now would be a mistake since there are other modifications of the first premise of the initial argumentation (i.e., that $\exists xTx$) that need to be discussed.

While useful as a presentational and heuristic aid, Stevin's thought experiment, because of its medieval origins, does not illustrate well another feature that I believe central for the use of thought experiments in modern science. Above I suggested that the counterfactuality of thought experiments might be rendered benign by developing, or imagining, a series of ever more refined real experiments that can be made to approach the idealized situation postulated by the thought experiment. To this process of refinement should be added the development of theories of interfering causes. Such theories serve to correct real experimentation and when applied to any real situation subtract out the interfering causes and leave the idealized analysis as remainder. The need for and role to be played by such corrective procedures will become clear in the next section.

II. Mach and the Reform of Mechanics

We now examine a thought experiment of Mach's that has remarkably similar structure to that of Stevin's. But first we need some background to better appreciate the role Mach's thought experiment was intended to play. Mach proposed that we accept as the definition of the expression "the mass of body B has φ/φ' times the mass of A" the following:

> The bodies A and B receive respectively as the result of their mutual action the accelerations $-\varphi$ and $+\varphi'$, where the senses of the accelerations are indicated by the signs. (Mach 1960, p. 266)

Mach asserts that it is a *real* experimental fact that coherent mass determinations can be made by means of ordinary laboratory manipulations. Such determinations will be coherent within types of laboratory manipulation as well as between types of laboratory manipulations. So, for example, if we determine by collision experiments that the masses of two bodies are respectively M_1 and M_2 (using some standard body to provide our mass unit), then if we go on to compare the weights of these bodies we will find that $W_1 = (M_1/M_2)W_2$. Mach is quite right to emphasize that it is not logically necessary that such coherence exist; such coherence can only be an empirical fact.[5] And it is because the world yields coherent mass determinations that Mach feels justified in asserting that:

> The concept of mass when reached in the manner just developed renders unnecessary the special enunciation of the principle of reaction. In the concept of mass and the principle of reaction...the same fact is *twice* formulated; which

> is redundant....in [our] concept of mass no theory of any kind is contained, but simply a fact of experience. (Mach 1960, pp. 269-71)

We should not go overboard here and overlook, as Mach evidently would have us, the usual sorts of discrepancies that infect all sorts of experimentation. For example, if we were to make a mass comparison of wooden and metal objects in the presence of a strong magnetic field that we later turned off, we would not get a coherent set of mass determinations. This sort of case is but an extreme version of a general difficulty that affects all experimentation. Theory mediated schemes are required to correct experimental results for both systematic and random interfering causes. In the absence of such correcting procedures, Mach's laboratory manipulations cannot be counted on to generate coherent mass determinations. So in a strictly empirical sense, of the sort intended by Mach, coherent mass determinations unmediated by theory are *not* forthcoming.

While Mach chooses to ignore complications of real experimentation, he does construct an elegant thought experiment to demonstrate the importance of coherent mass determinations. We are to imagine three "elastic" bodies *A*, *B* and *C* placed on "an absolutely smooth and rigid ring." Bodies *A* and *B* have been found to be of equal mass when compared with one another; similarly, bodies *B* and *C* have been found to be of equal mass when compared with one another. Assume now that bodies *A* and *C* do not interact as if their masses were equal; that is, assume an incoherency in the form of intransitivity in relative mass ratios. In particular, assume that mass *C* is greater than that of *A*. We now impart a velocity to mass *A*. It transmits this velocity to body *B*; similarly *B* transmits the velocity to *C*. However, here's the rub. Since *C* has (in its interactions with *A*) greater mass than *A*, *A* will receive from *C* a greater velocity than that initially imparted to *A*. If the process is allowed to continue, the bodies will move faster and faster thus violating conservation of energy.[6] So we have a thought experiment remarkably similar to Stevin's, where a conservation principle is used to generate some desired conclusion, in this case that mass determinations must be transitive. Mach's thought experiment, being frictionless &c., is clearly counterfactual and as such raises the question of its relevance for the claim that mass ratios are transitive in this world. In fact, if we were to consider a real instantiation of Mach's ring experiment we would find that the apparatus runs down. Are we to interpret this running down then as showing that the mass ratios are not transitive? It is now time to bring in the two part strategy introduced earlier for rendering benign the counterfactuality of thought experiments. We are to (1) show that there exists a series of experimental refinements such that real experiments can be made to approach the postulated idealized thought experiment, and (2) show that there exist a series of theoretical corrections that can be made to apply to real experiments such that once corrected real experiments look increasingly like the original thought experiment. But Mach's anti-metaphysical stance prohibits his using theory mediated corrections. If he uses corrective theories, coherent mass determinations will result, but at the expense of his claim that Newton's law of reaction and the assertion of the existence of mass as a property of bodies are redundant expressions of a simple empirical fact. Therefore, Mach can justify his thought experiment only by making implicit appeal to the fact that real experiments through successive refinements can be made to more closely approximate it. One wonders whether this is sufficient.

Let us assume a modal interpretation of Mach's thought experiment where:

Tx = x is Mach's ring experiment

M = mass relations are transitive

V = collisions respect conservation of mv^2

Kx = the kinetic energy of x increases (without loss of potential energy)

Mach's argument then is:

1. $\Diamond\exists x(Tx)$ & $\sim M$ & $V \rightarrow \Diamond\exists x(Kx)$
2. But $\sim\Diamond\exists x(Kx)$.
3. Therefore, $\sim[\Diamond\exists x(Tx)$ & $\sim M$ & $V]$.
4. By experimental refinement we have good warrant to believe that $\Diamond\exists x(Tx)$.
5. Therefore, on the assumption that V, it follows that M.

This seems fine. But consider the following complication. While it is true that real experiments can be made to approximate Mach's thought experiment ever more closely, it may not in fact be physically possible to totally eliminate all friction. There are many processes in physics that have this asymptotic character. (E.g., the approach to absolute zero.) So, we may have good reasons to distrust inductions that go from asymptotic approach to the possible existence (in this world) of the limit. Furthermore, since we desire some generality for our justification methods, it would be good to consider thought experiments of more radical counterfactuality than so far discussed. (E.g., considering a universe with only two bodies.) So let us assume that the approach to zero friction is not sufficient, because of overriding reasons, to inductively entail that $\sim\Diamond\exists x(Tx)$. Step (4) of the reconstruction is to be withdrawn. But with that withdrawal goes our justification of Mach's thought experiment.

To see how a theory of interfering causes can be used to justify thought experiments, and Mach's ring experiment in particular, assume a simple theory of rubbing friction between masses and ring where this friction is some positive and well-behaved function $F(v,\alpha_1,\alpha_2,\ldots\alpha_n)$ of velocity and other parameters $\alpha_1,\alpha_2,\ldots\alpha_n$. Assume also that rubbing friction is the only dissipative force that acts on the system. Therefore, work done by dissipative forces in one complete circuit of Mach's ring will be some integral $\int F(v,\alpha_1,\alpha_2,\ldots\alpha_n)dx$. For simplicity we assume that $\int F(v,\alpha_1,\alpha_2,\ldots\alpha_n)dx$ is always computable. (Similar arguments are possible if this assumption is weakened.) Since the kinetic energy of any real version of Mach's ring experiment decreases, it must be the case that (for one circuit),

$$1/2(m_c\text{-}m_a)v_o^2 < \int F(v,\alpha_1,\alpha_2,\ldots\alpha_n)dx$$

where v_o is the initial velocity imparted to A, and m_a and m_c are the masses of A and C with respect to one another. What the expression says then is that the gain in kinetic energy due to mass intrasitivity is less than the work done by friction. Since by experimental refinement the value of $\int F(v,\alpha_1,\alpha_2,\ldots\alpha_n)dx$, can be made increasingly small (our masses rotate for longer and longer periods), it follows that possible violations of mass transitivity (i.e. $m_c– m_a$) can be restricted to a range that is correspondingly small. Therefore, if we can avail ourselves of a theory of friction satisfying the above requirements, the conclusion of Mach's thought experiment can be approached as closely as experimental refinement allows. This, of course, is not quite the same thing as asserting that conclusion *simpliciter*. Saving Mach's conclusion as suggested will be robust if all forces acting on the ring system are dissipative, since in this case more sophisticated analyses (including considerations of, for example, air viscosity and variable friction due to heating) will lead to essentially the same result.

It will be convenient for future purposes to represent our justification of Mach's thought experiment in slightly different fashion. To do this we form the following equation where v_1 is the velocity of body A after one cycle is completed, and v_i is to be solved for:

$$1/2(m_a v_o^2\text{-}m_a v_1^2)– \int F(v,\alpha_1,\alpha_2,\ldots\alpha_n)dx = 1/2(m_a v_o^2–m_a v_i^2)$$

The variable v_i has a nice interpretation, namely, it is what the velocity would have been if friction had not been operating on the system. This is the real output of the experiment once frictional effects have been corrected for. What we would like, of course, is that it turn out that $v_i=v_o$. But this rarely happens; usually $v_i=v_o\pm\varepsilon$. However, we typically do find that as better theories of disturbing causes are used to correct the experimental data, ε approaches zero. In fact, there may exist general arguments showing that this must be the case.

We introduce some notation for the type of corrective procedure just illustrated. Let Ex be an accurate description of some experiment x, and let R be a theory or analysis of disturbing forces. We interpret $R(Ex)$ as being the corrected experimental values that result from applying R to Ex, i.e., $R(Ex)$ is what would have been measured if the disturbing forces had not been present. Returning to the Mach case, our belief that more adequate theories of interfering causes will yield decreasingly small ε values can be represented as: there exists a real experiment e and a series of corrective theories $R_1, R_2, \ldots R_n$ such that as i increases $R_i(Ee)$ approaches T, Mach's original idealized description.

With this notation in hand, we can express two ways that Mach's thought experiment can be justified. First, we construct a series of experiments, $e_1, e_2, \ldots e_n$, with decreasing amounts of friction and produce an analysis of friction R such that $R(Ee_i)$ approaches T as i approaches n. Second, we produce a series of increasingly more accurate analyses, $R_1, R_2, \ldots R_n$, of disturbing forces such that for some real version e of Mach's thought experiment, $R_i(Ee)$ approaches T as i approaches n. Obviously, both procedures can be combined to more thoroughly nail down the conclusion Q, or to reduce any associated range of possible variation ε.

As suggested earlier, Mach would probably have rejected justifications of these sorts on the grounds that they are excessively theoretical and violate the empirical requirements of his reform program. If this is correct, then it seems the only way Mach can demonstrate the relevance of his thought experiment is to insist on the truth of the modal premise $\Diamond\exists x(Tx)$. But I think the establishment of such a modal will be very difficult given Mach's anti-metaphysical program.

III. A Definition of Successful Thought Experiment

The presentation of some thought experiments can be understood as containing arguments of the form,

$$\exists x(Tx)\ \&\ P_1\ \&\ P_2\ \&\ldots P_n \rightarrow Q$$

where $\exists x(Tx)$ is a highly idealized experimental description, $P_1, P_2, \ldots P_n$ are laws or principles believed true, and Q is to be demonstrated. We have called this the *initial argumentation* of the thought experiment. A basic problem with such thought experiments is that the initial argumentation is unsound, since $\exists x(Tx)$, being highly idealized, is false. Our strategy has been to transform this first premise into something true, either $\Diamond\exists x(Tx)$, or $\exists$ series (or there is good reason to think that such series can be constructed) $R_1, R_2, \ldots R_n$ and $e_1, e_2, \ldots e_m$ such that $R_i(e_j)$ approaches T as i and j approach respectively n and m.

Our transformations have been highly reconstructive. It's what we with philosophical and scientific hindsight can do to make sense of some thought experiments. We need to relate such reconstruction to actual historical practice. This we now do. Our focus is on thought experiments that make use of idealization and counterfactuality. A striking feature of historically presented experiments of this type is the absence of clear argumentation explicitly justifying their relevance for the real world. We have suggested that justification for the use of idealized experimental situations could be obtained by showing that there exists a series of real experiments that approaches the postulated situation, or by developing theories that will enable one to analyze away interfering causes so that what results are residual analyses that approach that used in the postulated thought experiment. But such argumentation is at best only implicitly used or encouraged by presenters of thought experiments. Given the lack of explicit justifying argumentation it becomes all the more striking that idealized thought experiments tend to be extremely persuasive. This last feature creates a difficult problem for the psychology of scientific development. Our suggestion for understanding the persuasive efficacy of historically presented thought experiments is that audiences will naturally bring to bear certain experiences that will be seen to be analogous to the problem at hand. In other words, the thought experiment invites and triggers a psychological response (of analogy construction) that makes it seem plausible that real experiments can be made to approach thought experiments or that they can be analyzed, after correcting for interfering causes, as if they were thought experiments. These thoughts are expressed more precisely in terms of a definition of *successful thought experiment.*

> A *thought experiment* (as defined earlier) is *successful* with respect to the set of persons Φ if:
>
> (4) If $\Diamond\exists x(Tx)$ or if $\sim\Diamond\exists x(Tx)$ but T is asymptotically approachable, then members of Φ believe that it is possible to construct a series of real experiments, $e_1, e_2, \ldots e_n$, such that the description of each successive member more closely approximates T.
>
> (5) If $\sim\Diamond\exists x(Tx)$, then members of Φ believe that it is possible to construct a set Ω of real experiments $\{e_1, e_2, \ldots e_n\}$ and a set Ψ of theories or analyses of interfering causes such that members of Ψ can be selectively applied to members of Ω so as to yield a series of residual analyses that converges to T.
>
> (6) $\supset \alpha \in \Phi$, if α did not believe Q before being presented with ϑ then α believes Q after being presented with ϑ.

The conditions given for success are meant to only be sufficient. We shall discuss other roads to success below in section 5. Conditions (4) and (5) are to be understood as meaning that members of Φ believe that there exist some experiments or other that satisfy the stated conditions. They are not to be understood as requiring that members of Φ have specific experiments in mind, although they may have. (Similarly for theories of interfering causes.) My definition of *success* is something of a hybrid notion since the antecedents of (4) and (5) are stated objectively while the consequents of (4) and (5) are relativized to the beliefs of members of Φ. I have chosen this form in the interests of a readable definition that would allow the importance of the processes of conditions (4) and (5) to be clearly evident. Of course, the definition could be consistently relativized to beliefs, and so relativized yield a notion of *convincing thought experiment.* On the other hand, one could make the definition consistently objective by adding the qualification that all beliefs be correct ones. This can be taken to yield a *truly successful thought experiment.* There are obviously many ways to calculate success here, but such subtle distinctions are not required for the purposes of this paper.

In the interests of accuracy to actual practice, I have not built into the definition of *success* that members of Φ understand or know the role that the fourth and fifth conditions play in justifying the use of counterfactual situations. This is important since we wish to clearly distinguish between actual practice and reconstructed justifications of that practice. The practice, I contend, contains the seeds of justification but not the justification *per se.*

The definition perhaps should be expanded to include some conditions on the causal processes by which members of Φ are brought to believe the various things required by the analysis. For example, a clause such as "on the basis of perceived analogies" might be added at appropriate places. The belief conditions could also be modified by assuming some measure of strength of belief, in which case, it might be possible to define some derivative measure of the success of the thought experiment. In this connection, we note that conditions (4) and (5) represent a minimum standard since the definition allows members of Φ to have beliefs as described in the consequent of (5) even if $\Diamond\exists x(Tx)$. The fourth and fifth conditions are too strong as stated and should be weakened to require only a partial ordering or perhaps something slightly stronger.[7]

We next apply our definitions to some thought experiments dealing with the relation between thermodynamics and statistical mechanics.

IV. Gouy's Perpetual Motion Machine of the Second Type

Explaining Brownian motion posed a longstanding puzzle for physicists. Explanations in terms of external disturbances seemed unlikely given a long series of experiments showing a lack of concomitant variation between suspected causes and particle motion. Furthermore, many elegant experimental refinements were introduced to more completely isolate the Brownian particle system from external disturbances. The result was that the more isolated the system, the longer the motion continued.[8] Given this experimental background it seemed safe to describe the Brownian particle system as if it were in a state of perfect thermal equilibrium. For whatever unavoidable disturbances that might exist would certainly be quite small as well as essentially irrelevant for Brownian motion. So even if a system in perfect thermal equilibrium were a physical impossibility, one could in this world approach such perfection with arbitrary closeness modulo the phenomenon of Brownian motion. Therefore in the case of Brownian motion one had explicitly the sorts of justification that were only reconstructive possibilities for Mach's ring experiment.

To convert Brownian motion into a thought experiment the nineteenth century French physicist Gouy added an ideal mechanical energy collector, thus producing a perpetual motion machine of the second type.

> Whatever idea one may have as to the cause that produces [the movement], it is no less certain that work is expended on these particles, and one can conceive a mechanism by which a portion of this work might become available. Imagine,

> for example, that one of these solid particles is suspended by a thread of diameter very small compared to its own, from a rachet wheel; impulses in a certain direction make the wheel turn, and we can recover the work. This mechanism is clearly unrealisable, but there is no theoretical reason to prevent it from functioning. Work could be produced at the expense of the heat of the surrounding medium, in opposition to Carnot's principle.[9]

As we come to expect, no explicit justification is given by Gouy to take his counterfactual thought experiment as being demonstrative. The supposition of perfect thermal equilibrium is no problem since Gouy had himself contributed to the experimental tradition described above, and had reviewed for his readers that tradition before giving the thought experiment. The rachet mechanism though presents a problem. Gouy's modal ambivalence indicates great uncertainty about what to do here: "Ce mécanisme est évidemment irréalisable, mais on ne voit pas de raison théorique qui put l'empêcher de fonctionner" (Gouy, 1888, p. 564). What can this mean? Poincaré came up with an ingenious solution: to find some functional but realizable equivalent of Gouy's rachet. This equivalent was to be simply the work done by a resisting fluid against the motion of the Brownian particle.

> If, then, these movements [of Brownian particles] never cease, or rather are reborn without ceasing, without borrowing anything from an external source of energy, what ought we to believe? To be sure, we should not renounce our belief in the conservation of energy, but we see under our eyes now motion transformed into heat by friction, now heat changed inversely into motion, and that without loss since the movements lasts forever. This is the contrary of the principle of Carnot.

Poincaré also goes on to observe that Brownian motion itself is something of a functional equivalent for another very famous thought experiment, namely, Maxwell's demon.

> ...to see the world return backward, we no longer have need of the infinitely subtle eye of Maxwell's demon; our microscope suffices us. Bodies too large, those, for example, which are a tenth of a millimeter, are hit from all sides by moving atoms, but they do not budge, because these shocks are very numerous and the law of chance makes them compensate each other; but the smaller particles receive too few shocks for this compensation to take place with certainty and are incessantly knocked about. (Poincaré 1906, p. 610)

I do not know of any historical surveys of the specific reaction to Gouy's thought experiment or Poincaré's modification. Gouy's thought experiment should have raised questions about exactly what the second law should be taken to mean. Surprisingly, neither Boltzmann nor Maxwell discussed Brownian motion. Einstein was unaware of the phenomenon and invented it *a priori* as a means of establishing kinetic theory. Poincaré's adaptation raises the specific question of whether it is really sensible to ascribe to the fluid the double function of both generating and resisting motion. Historically, the situation remained obscure until Einstein and Smoluckowski independently developed a statistical explanation of the phenomenon.

Einstein in 1909 gave a thought experiment remarkably like that of Gouy but with a somewhat different purpose in mind (Einstein 1909a, pp. 189-90; 1909b, p. 823). Like Gouy, Einstein began with an environment in thermal equilibrium, but unlike Gouy, Einstein's environment consists of two substances: an ideal gas and blackbody radiation. Einstein's analogue for the Brownian particle is a suspended mirror free to move perpendicularly to its surface. Finally, like Gouy, Einstein separates the driving and retarding functions. The ideal gas serves only to drive the mirror by means of random collisions while the blackbody radiation serves only to damp the motion. But as Einstein shows, such assumptions cannot be consistently maintained without macroscopic violation of the second law of thermodynamics.[10] Since the macroscopic version of the law was not suspect, Einstein concluded that the blackbody radiation like the ideal gas was subject to local fluctuations in pressure. Klein, in an excellent review, reports that Einstein's 1909 calculations of the energy fluctuations in blackbody radiation were not well received. Presumably this negative reception was also true of the thought experiment.

> Convincing as these results were to Einstein, they left his colleagues unpersuaded. While he saw the analysis of fluctuations as a powerful instrument for exploring the structure of radiation, most other physicists were barely convinced of the existence of such fluctuations.[11]

If this appraisal is correct, then our analysis suggests that the negative response to Einstein's thought experiment can be explained in terms of the absence of perceived analogies with existing experimental and analytical work.

V. The Bucket Experiments of Newton and Mach

In this section I want to reconsider the suggestion made earlier to consider thought experiments as being implicitly semantical arguments. The idea to be briefly developed is that thought experiments provide models or semantic domains for theories, and as such can be used to discover some of the logical properties of scientific theories. That is, thought experiments function in much the same way that ordinary semantic structures (e.g. the truths of arithmetic or plane geometry) function. We shall utilize Mach's criticisms of Newton's bucket experiment as a means of introducing this sort of semantic function for thought experiments. Mach has been greatly praised (by, for example, Hans Reichenbach, Ernest Nagel, Max Jammer, Ian Hacking, and Richard Westfall) for having debunked Newton's bucket experiment. What Newton is claimed to have done is represented in the accompanying table:

The Received View of Newton's Bucket Experiment

	Water Surface	Relative Motion of Bucket and Water	Absolute Motion of Water
initial state of rest	flat	none	none
bucket begins rotating	flat	yes	none
bucket continues rotating	curved	none	yes
bucket suddenly stopped	curved	yes	yes
final state rest	flat	none	none

By simple lack of concomitant variation it follows that the shape of the water is independent of its state of relative motion with respect to the sides of the bucket. I do not know who is responsible for having originated this historical fabrication, but as even a cursory reading of the *Principia* reveals, Newton in fact only reports having done what corresponds to the first three lines of the table. This should cast some suspicion on the correctness of the received view. Since my interests in this paper are primarily analytical and not historical, I shall accept for the sake of argument the received view of Newton's bucket experiment.[12] After all Newton could have performed the experiment attributed to him.

Accepting then the received view, as Mach also seems to have done, the experiment is clearly something of a counterfactual thought experiment because, given naked eye observation, the shape of the water becomes curved (if only at the edges) *as soon as* the bucket is rotated.[13] Now, of course, one can *imagine* repeating the experiment with ever larger buckets so that this initial curvature becomes an increasingly less significant variation from the flatness assumed in the first, second and fifth lines of the experimental summary. Furthermore, one can also imagine developing a theory of surface tension such that these annoying variations from flatness can be analyzed away and the surface treated *as if* it were flat. That is, one can engage in exactly those activities that I claim justify the use of counterfactual thought experiments. In fact, all of this is so natural that one cannot seriously believe that anyone, even in Newton's day, would have bothered to have performed the experiment. Or if so, not to have performed it in other than the most perfunctory way.[14]

Mach has no complaints about using the lack of concomitant variation to show the independence of the shape of the water from relative motion with respect to the bucket. His criticism is that the experiment does not show that the surface of the water is independent of relative motion with respect to all bodies.

> Newton's experiment with the rotating vessel of water simply informs us, that the relative rotation of the water with respect to the sides of the vessal produces no noticeable centrifugal forces, but that such forces are produced by its relative rotation with respect to the mass of the earth and the other celestial bodies. No one is competent to say how the experiment would turn out if the sides of the vessel increased in thickness and mass till they were ultimately several leagues thick. (Mach 1960, p. 284)[15]

On the face of it, it seems quite incredible that Newton possibly could have thought that the bucket experiment shows that water shape does not appropriately vary with respect to any objects whatsoever. This implausibility alone should have alerted Mach's supporters that something was seriously amiss in their understanding of Newton. In fact, Newton quite clearly does not make or even suggest the widely extravagant claim commonly attributed to him. His aims were quite modest and were only (1) to give an illustration of how assuming absolute space *ab initio* one can explain the variation of water surface shape; and (2) to show that Descartes' theory of relative motion and its dynamic effects could be refuted. For the latter it is sufficient to note that the degree of water curvature is inversely proportional to its relative motion with respect to the sides of the bucket.[16] Of course, someone might think that Newton's bucket experiment could be co-opted for more general purposes. A charitable reading of Mach would have his aim being the preemptive one of blocking such persons. Say one were to think that Newton's bucket experiment could be reconstructed as $\exists x(Tx \,\&\, {\sim}Rx) \Rightarrow \supset y(Ty \supset {\sim} R'y)$, where Tx means x is the bucket experiment of the received view (but with unspecified thickness of sides), Rx means the shape of the water of x is a function of relative motion with respect to the sides of the bucket used in x, and $R'x$ means the shape of the water of x is a function of relative motion with respect to some other bodies. Mach can be conceived or reconstructed as trying to convince such a person that the argument cannot be completed (i.e., made explicit) with only logical and perhaps other non-controversial truths. Mach's suggested variation of Newton's bucket experiment, that we increase the thickness of the sides of the bucket, is to play some therapeutic role here. There are several ways this could be done. One is to conceive the bucket experiments of Newton and Mach as forming a semantic domain, that is, as together providing a possible interpretation of some (object) language claim. Mach's possibility claim—that the shape of the water become, given sufficiently thick vessel sides, a function of the relative motion of the water and the sides of the vessel—is interpreted as being a claim about acceptable semantic or interpretational domains. In particular, that Newton's original experiment and Mach's experiment, now assumed to show shape dependency on the very thick bucket sides, form an acceptable combination of semantic objects. (Cf. the claims of ordinary arithmetic which can be conceived as forming an acceptable semantic domain.) Understood this way, we see that Newton's experiment models $\exists x(Tx \,\&\, {\sim}Rx)$ while Mach's experiment provides a counterexample to $\supseteq y(Ty \supset {\sim}R'y)$. Therefore, it cannot be the case that the conclusion is logically entailed by the premises. All of this suggests a somewhat different role for thought experiments than that suggested earlier in this paper. But having introduced the possibility of other criteria for success, I shall say no more here about the use of thought experiments as semantic models.[17]

VI. Thought Experiments and Mainstream Science

Inventing thought experiments seems an atypical and rather specialized form of scientific activity. One wonders whether it relates in a natural way to more mainstream forms of scientific activity. In the case of counterfactual thought experiments there is first and foremost the initial argumentation which we have represented as: $\exists x(Tx) \;\&\; P_1 \;\&\; P_2 \;\&\ldots P_n \rightarrow Q$. Consider now the contrasting situation where there is, so to speak, first and foremost the real experiment. In such a situation one casts about for an analysis that will enable the experiment to be attached logically to some theory or other.[18] Given the usual sorts of analytical intractibility, and the usual sorts of shortages of necessary auxiliary theory and data, idealized analyses will have to be used in order to generate practically computable predictions. If this is true, then some justification will be required to support the relevance of the idealized analysis used.

A simple hypothetico-deductive model of theory testing will be sufficient to indicate the sort of relevance problem I have in mind.[19] Let T represent some underlying theory such as Newton's laws or the Relativistic Field Equations. Let I represent the idealizing assumptions made. Include in I the required parameter or initial condition values. Finally, let P be the practically derivable prediction: for example that Kepler's second law will hold true for planetary orbits, or that light rays will deflect according to a hyperbolic law with an ordinate intercept value of 1.75″ at the solar radius. Now P will be true or false, i.e., P will or will not be correct to within calculated or estimated experimental error. Philosophical and scientific common sense has it that in the first case there is confirmation, or at least the satisfaction of a necessary condition for confirmation, and in the second case disconfirmation. Consider the case of disconfirmation:

$$\begin{array}{l} T \,\&\, I \rightarrow P \\ {\sim}P \\ \hline {\sim}T \vee {\sim}I \end{array}$$

Simple inspection reveals an immediate problem. Even assuming as unproblematic the truth of the premises (i.e., that the theory and idealizations have P as a logical consequence, and that the experimental result is correctly described as being inconsistent with the truth of P), nothing logically follows about the truth or falsity of the theory. What follows is only that either or both the theory and idealizations are false. But we *already* know that the idealizations are false; so nothing is gained! In other words, the falsity of *I* protects the theory against refutation. None of this should be surprising if we consider that the idealizations (because they are false) introduce *bias* or distortion into our computations. Hence, given a true theory, the prediction cannot be true (unless there are canceling biases). Laboratory students who attempt to fudge data by distributing bogus experimental values "normally" about a predicted value deservedly fail because they have not assimilated this basic truth. Given this perspective, we can see why the standard hypothetico-deductive account also fails to yield confirmation or confirmatory value. This is because (in the absence of fortuitously canceling biases) only a false theory, when conjoined with a biased idealization, can lead to a correct prediction.

The problem created by the use of idealizations for science then is to determine whether failures to achieve experimental fit to within experimental error are due to the falsity of theory or of idealization. In other words, the problem is to determine when we can *praise* theories for achieving as close a fit as is achieved and *blame* the idealizations for the failure to achieve experimental fit to within experimental error. In rare cases where experimental fit to within experimental error is achieved, it must be determined whether this is due to the truth of theory and fortuitously canceling idealizations, or to a fortuitous combination of false theory and false idealizations.

In other published work I have shown that a consideration of many important historical cases suggests the following theses.[20]

> A scientific theory is confirmed (or receives confirmation) if it can be shown that using more realistic idealizations will lead to more accurate predictions.
>
> A scientific theory is disconfirmed if it can be shown that using more realistic idealizations will not lead to more accurate predictions.

The essential idea behind these proposals is to give up any metrical idea of closeness to the truth of idealized initial or boundary conditions and substitute the *partial ordering* of relative realism based on existing background standards. This means, of course, that as background standards change our judgments of relative realism may also change; hence, judgments of confirmation may have to be modified. What is being proposed is that acceptable scientific theories be *monotonic* toward the truth in the sense that more accurate and less idealized initial condition descriptions lead to more accurate predictions.[21] For example, consider the calculation using kinetic theory of the pressure of a gas. The most elementary calculation of pressure proceeds on the assumption that gas molecules are infinitesimal in size and that they exert forces on one another only in collision. On the basis of these simplifying assumptions, the ideal-gas law can be derived. However, as Jeans notes "neither of these assumptions is true for an actual gas, and so [we] must proceed to calculate the pressure for a real gas in which the molecules are of finite size, and exert forces of cohesion on one another even when they are not in contact" (Jeans 1940, p. 63). What is impressive about the kinetic theory is that a *more accurate* equation of state (i.e., a more accurate prediction) can be generated when the simplifying assumptions are made *more realistic.* The standard example of this sort of improvability is the Van der Waals calculation of the equation of state on the basis of molecules finitely sized that exert forces of cohesion. The kinetic theory or program thus receives confirmation because the use of more realistic descriptions or initial conditions leads to more accurate predictions.

One aspect of mainstream science then is the construction of more realistic and less idealized analyses of real experiments, where the anticipation is that more realistic analyses will yield closer fit with existing data (if the underlying theory is true). A closely related activity is the refinement of real experiments so as to make them more closely approximate existing analyses. Again the anticipation is that experimental fit will undergo improvement. Given our discussion of thought experiments, all of this should sound remotely familiar. Both thought experiments and real experiments are characterized by the fact that idealized descriptions are used that are strictly speaking not true of situations in this world. Both thought experiments and real experiments can be conceived as testing some theory or law. Furthermore, the testing process is at a deep level the same, namely, real experiments are made to approach via theory mediated corrections ideal counterfactual analyses. So where's the difference?

Let us return to the initial argumentation of thought experiments: $\exists x(Tx)$ & P_1 & P_2 &...$P_n \Rightarrow Q$. Such argumentation occurs in the presentations of both real and thought experiments. In thought experiments the premises are believed true; Q is in question. For real experiments, only some of the premises are initially treated as if they were true; others, let us say $P_1,P_2,\ldots P_j$, are to be tested. Also, Q, for real experiments, tends not to be consistent with real experimental results. So for thought experiments the problem is to justify Q given that $\exists x(Tx)$ is false. For real experiments the problem is to justify $P_1,P_2,\ldots P_j$ given that $\exists x(Tx)$ are Q are false. For thought experiments the method used to justify Q is to substitute in the argumentation a justifiable premise for the false $\exists x(Tx)$. The substitute premise and the methods used to justify its truth are as previously discussed. For real experiments the method, or so I claim, is essentially to construct a series of arguments, $\exists x(T_ix)$ & P_1 & P_2 &...$P_n \Rightarrow Q_i$ (or argue that such a series can be constructed), such that $<T_i, Q_i>$ approaches $<T,Q>$ as i increases. On the basis of such approach (and other restrictions), one concludes by induction that the set $\{P_1,P_2,\ldots P_j\}$ produces monotonic graphs. Finally, this production is seen as confirming or lending confirmatory value to $\{P_1,P_2,\ldots P_j\}$.[22]

We shall close this paper with an example that illustrates our distinction between real and thought experimentation.

VII. Newton's Experimentum Crucis

The point of Newton's well-known *experimentum crucis* is to establish the claim that light consists of rays with different degrees of "refrangibility," i.e., refractive index. The experiment begins with the generation of a spectrum by passing sunlight through an aperture and an ordinary prism. A small amount of the spectrum then is separated off by means of a second aperture. But sending this separated light through another prism does not, as in the case of the first prism, generate another spectrum. According to Newton one should interpret this result as follows: the first prism serves to separate out the different rays; the second aperture serves to isolate rays of unique refrangibility; finally, since rays of unique refrangibility have been isolated, the second prism naturally will not produce further separation.

Newton originally asserted that there is no color separation or dispersion after the second prism. But, as his critics correctly noted, this is not experimentally correct; color separation and dispersion were within the capabilities of then-ordinary methods of observation. What Newton had done therefore was to present an idealized description of both the experimental situation *and* its result. Newton's Euclidean ray analysis of the experiment was simplified or idealized since the actual size of the second aperture was not taken into account. It was simply assumed to be infinitely small so as to collect only a single ray. But this assumption is in difficulty in at least two ways. First, experimentally it is not in fact possible to continually reduce aperture size so as to collect light of unique refrangibility. Diffraction effects take over before the second prism ceases to produce dispersion. Second, Newton's attempt to make the notion of rays phenomenological and non-hypothetical precludes any account as to how the second aperture physically could collect rays of unique refrangibility.[23] The reader will have by now developed the suspicion that Newton *started* with the principle that white light consists of rays of different refrangibility and then realized that an idealized *experimentum crucis* would not show dispersion or further separation after the second prism. If so, then the *experimentum crucis* could have been presented, *à la* Stevin, as a thought experiment. What this suggests is that real experimentation in its conceptual or developmental stage will take on the characteristics we have associated with thought experimentation. The developer suspends critical judgment and assumes as a working hypothesis that the theory to be tested is true. An ideal apparatus is then invented that will generate some result that can be measured by available techniques. Once such an apparatus has been found, an attempt is made to explicitly justify the argumentation of the thought experiment in the ways discussed in this paper. Finally, if this is successful, the working hypothesis can be discharged and testing of the theory can begin. In Newton's case such a procedure would have meant realizing that the counterfactual assumption of an infinitely small second aperture (that collected only parallel rays) could be relaxed to a finite aperture, but at the cost of an experimental result less sharp than that of the original thought experiment.

In the light of our proposal, Newton's published report can be seen to be something of a hodgepodge. There is an idealized counterfactual analysis as well as an accurate description, with one exception, of the parameters of a real experiment. The exception, of course, is the reported result. (Remember Newton reported no separation or diffusion after the second prism.) Newton's response to those critical of his not entirely true experimental report was to assert but not to prove the truth of this conditional: If the finite sizes of the apertures are taken into account, the result will be an improved prediction that allows for some color separation and image dispersion. If true, the experiment is confirmatory of Newton's optical theory since a more realistic analysis leads to a more accurate prediction. As we have noted, it is a feature of counterfactual thought experiments that their presenters do not explicitly give the argumentation necessary to show the real relevance of the assumed but counterfactual situation. Newton's response to his critics has a somewhat similar feature since he did not actually construct the more realistic calculation for the experiment. He merely asserted that it could be constructed and that a correspondingly more accurate prediction would result. Newton could have made the calculation but he apparently deemed the computational cost too high. Besides, a mere assertion was sufficient since anyone with experience with Euclidean ray optics would see immediately (based on analogy with other problems) that Newton's claim was correct. So there was in fact no need to explicitly demonstrate the truth of the claim.

VIII. Conclusion

My thesis has been that thought experiments that postulate counterfactual situations can be profitably reconstructed as being invitations to imagine or construct experiments and associated analyses of interfering causes which when taken together yield corrected experimental descriptions that approach the postulated situation of the thought experiment. The specifics of this thesis have been summarized in a definition of *successful thought experiment.* I have also indicated a semantic or interpretative role that thought experiments can be conceived to sometimes play. Finally, the argumentation that we associate with idealized thought experiments has been shown to be closely related to that used in real experimentation.[24]

NOTES

1. The expression "logically yield" should be understood as being with respect to

some concept of logical rigor, whatever that might be. So when we attribute use of → to someone, we should indicate the intended standards of logical rigor.

2. Norton in his 1987 seems to suggest this. However, in his defense, he does sometimes speak as if he has a less narrow sense of *reductio* in mind than here discussed.

3. If we wondered about the logical independence of two principles, the existence of a model where not both are true shows them to be logically independent. We give below in section 5 an example of a thought experiment that can be naturally interpreted in semantic terms.

4. All that follows from (4), and the additional plausible premise $\Diamond\exists x(Tx \mathbin{\&} Ex) \rightarrow \Diamond L$, is that $\Diamond L$, which is not as desired. If we could assert $\Diamond\exists x(Tx \mathbin{\&} Ex) \rightarrow L$, then, of course, L would follow. But it is unclear what the justification is to be for this stronger premise. We might introduce Lx to mean x satisfies the law of equilibrium. Then we assert that $\Diamond\exists x(Tx \mathbin{\&} Ex) \rightarrow \Diamond\exists x(Tx \mathbin{\&} Ex \mathbin{\&} Lx)$, i.e., that Lx is logically contained in or follows from Tx and Ex. Reminding ourselves that $\Diamond$ represents physical or nomic possibility, we then introduce the induction principle: $\Diamond\exists x Lx \mathbin{\&} R \rightarrow L$, where R is some set of reasonable restrictions. Exactly what R should be though is a standard problem in induction and is not one I shall pursue here. See Norton 1987, for additional remarks about the role induction plays in thought experiments.

5. See Buchdahl 1951 for a modern analysis of the type of definitional coherence Mach has in mind.

6. "But a constant increase of *vis viva* of this kind is at decided variance with our *experience*" (Mach 1960, p. 269).

7. I have in mind here a continuous lattice structure; see Laymon 1987.

8. For a review of the experimental literature see Smoluckowski 1905, and for an excellent overall review of the subject see Brush 1968.

9. Gouy 1888, p. 564. The translation is from Brush 1968, p. 12.

10. See Norton 1987 for a very clear version of Einstein's thought experiment that gets to the heart of the argumentation.

11. Klein 1980, p. 178. Klein gives as partial evidence Planck's published views on the need to *exclude* fluctuations from physics.

12. For a review of the received view and a possible explanation of its origins see Laymon 1978b.

13. In his *Science of Mechanics*, Mach gives a quotation from Newton's *Principia* but does not redescribe the experiment; he only refers to it (1960, p. 250, 277 and 284).

14. Newton in *Principia* only parenthetically reports having performed the experiment. But even this should not be accepted at face value.

15. One would have expected Mach's positive claim to have been more circumspect and to have been: "[and not] that such forces are [not] produced by its relative rotation with respect to the mass of the earth and the other celestial bodies," or "that such forces are [possibly] produced." But the translation is correct. ("...dass dieselben aber durch die Relativdrehung gegen die Masse der Erde und dir ubrigen Himmelskörper geweckt werden," Mach 1976, p. 226). The weaker claim would be expected since Mach's argument that absolute space is not required to support the principle of inertia is not given until several pages later. We also note that in the paragraph immediately preceding the quoted passage, Mach says that the "principles of mechanics can...be so conceived, that even for relative rotations centrifugal forces arise." So the weaker claim would have been more in keeping with Mach's immediately expressed aims as well.

16. See Laymon 1978b for an account of Newton's intentions and understanding of the experiment.

17. Kuhn's 1964 is best understood as being about the semantic role that thought experiments can play. For an elementary introduction to the virtues of semantic theory see Van Frassen 1980.

[18] As a description of actual scientific practice all of this is highly idealized. But I believe nothing I shall say depends on the idealization. Everything could be said without it though much more awkwardly.

[19] The problem presented can also be generated on the basis of Bayesian or Glymourian boostrap theories of confirmation.

[20] E.g. Laymon, 1978a, 1983, 1985. Essentially, theories are to be conceived as consisting of laws (such as Newton's laws or the Relativistic Field Equations) which are unconditionally stated, i.e., which do not contain *ceteris paribus* clauses. The theses can be adapted to cases where the theories to be tested contain *ceteris paribus* clauses, but this requires some analysis of the specific computational structures at issue.

[21] Strictly by monotonic we mean that the predictions will be no less accurate. See Laymon 1987 for more details.

[22] This is just a rough sketch of our position; for more details see Laymon 1987.

[23] For a discussion of Newton's problems with *rays* and for the historical details of the *experimentum crucis* see Laymon 1978a.

[24] Thanks go first and foremost to John Norton for having introduced me to the problems of thought experiments with his insightful 1987. He also served as a gracious and critical sounding board for my developing views on the subject. Thanks also go to Ken Lucey and Ulrich Majer, colleagues during my stay at the Center for the Philosophy of Science, at the University of Pittsburgh. I am also indebted to Nicholas Rescher, the Director of the Center, for having provided me with a stimulating research environment as well as financial support. Financial support was also provided by the NSF (SES-8608167) and by The Ohio State University, my home institution. This paper is a small expression of my gratitude for that support.

REFERENCES

Buchdahl, G. (1951), "Science and Logic: Some Thoughts on Newton's Second Law of Motion," *British Journal for Philosophy of Science* vol. 2, pp. 217-35.

Brush, Stephen G. (1968), "A History of Random Processes: I. Brownian Movement from Brown to Perrin," *Archive for History of Exact Sciences* vol. 5, pp. 1-36.

Einstein, Albert (1909a), "Zum gegenwärtigen Stand des Strahlungsproblems," *Physikalische Zeitschrift* vol. 10, pp. 185-93, 323-24.

Einstein, Albert (1909b), "Über die Entwicklung unserer Anschauungen über das Wesen und die Konstitution des Strahlung," *Physikalische Zeitschrift* vol. 10, pp. 817-26.

van Fraassen, Bas (1980), *The Scientific Image* (Oxford: Clarendon Press).

Gouy, M. (1888), "Note sur le mouvement brownien," *Journal de Physique, Theorique et Appliquée* vol. 7 (ser. 2), pp. 561-64.

Jeans, Sir James (1940), *An Introduction to the Kinetic Theory of Gases* (Cambridge: Cambridge University Press).

Klein, M. (1980), "No Firm Foundation: Einstein and the Early Quantum Theory," in *Some Strangeness in the Proportion: A Centennial Sympo-*

sium to Celebrate the Achievements of Albert Einstein. Edited by H. Woolf (New York: Addison-Wesley), pp. 161-85.

Kuhn, Thomas (1964), "A Function for Thought Experiments," L'aventure de la science, Mélanges Alexandre Koyré. Volume II. Paris: Hermann, pp. 307-34. Reprinted in Thomas Kuhn, *The Essential Tension* (Chicago: University of Chicago Press, 1977), pp. 240-65.

Laymon, R. (1978a), "Newton's *Experimentum Crucis* and the Logic of Idealization and Theory Refutation," *Studies in History and Philosophy of Science* vol. 9, pp. 51-77.

Laymon, R. (1978b), "Newton's Bucket Experiment," *History of Philosophy* vol. 16, pp. 399-413.

Laymon, R. (1983), "Newton's Demonstration of Universal Gravitation and Philosophical Theories of Confirmation," in *Minnesota Studies in the Philosophy of Science X*, J. Earman (ed.) (Minneapolis: University of Minnesota Press), pp. 179-99.

Laymon, R. (1985), "Idealizations and the Testing of Theories by Experimentation," *Experiment and Observation in Modern Science,* P. Achinstein and O. Hannaway (eds.) (Boston: MIT Press and Bradford Books), pp. 147-73.

Laymon, R. (1987), "Using Scott Domains to Explicate the Notions of Approximate and Idealized Data," *Philosophy of Science,* vol. 54, pp. 194-221.

Mach, Ernst (1976), *Die Mechanik: Historisch-Kritisch Dargestellt* (Darmstadt: Wissenschaftliche Buchgesellschaft).

Mach, Ernst (1960), *The Science of Mechanics,* translated by Thomas J. McCormack (La Salle: Open Court).

Norton, John (1987), "Thought Experiments in Einstein's Work," *this volume.*

Poincaré, Jules H. (1905), "The Principles of Mathematical Physics," in *Congress of Arts and Science, Universal Exposition, St. Louis, 1904,* Volume I (Boston: Houghton, Mifflin), pp. 604-22.

Smoluchowski, M. R. von Smolan (1906), "Molekular-kinetische Theorie der Opaleszenz von Gasen im kritischen Zustande, sowie einiger verwandter Erscheinungen," *Annalen der Physik* 21 (ser. 4), pp. 756-80.