Advanced Reasoning Forum

Proceeding of the 2nd Annual Conference on Logic and Reasoning

New Europe College
Bucharest, Romania
July 2000

Bulletin of Advanced Reasoning and Knowledge
Volume 1, 2001
Edited by Richard L. Epstein

New Europe College, Bucharest
Proceedings of the
2nd Annual Conference
on Logic and Reasoning

New Europe College
Bucharest, Romania
July 2000

Bulletin of Advanced Reasoning and Knowledge

Volume 1, 2001
Edited by Richard L. Epstein

New Europe College, Bucharest
Proceedings of the 2nd Annual Conference on Reasoning and Logic
New Europe College
Bucharest, Romania
July 2000

Preface .............................................. 1

Bulletin of Advanced Reasoning and Knowledge .......... 2

Theory
Arguments and explanations ...................... 3
   Richard L. Epstein

Things that aren’t there ......................... 19
   Fred Kroon

Consciousness, materialism, and ideology .......... 33
   William S. Robinson

Modals and quantifiers .......................... 49
   Mircea Dumitru

The problem of unwanted epistemic necessities .... 65
   Célia Teixeira

The new theory of reference: Problems and solutions concerning proper names 69
   Sorin Costreie

Generating necessary a posteriori truths ........... 83
   Desidério Murcho

Ex contradictione non sequitur quodlibet .......... 89
   Walter A. Carnielli and João Marcos

Pedagogy and Examples
Knowing the exact proportions is not enough .... 111
   Stanisław Krajewski

Discussion: On the ethics of killing dogs .......... 117
   Richard L. Epstein, Williams S. Robinson, Fred Kroon

Cartoon .............................................. 121
   Alex Raffi

Prizes and Announcements ........................ 122

Advanced Reasoning Forum members and contributors to this issue ... 123

© 2001 Advanced Reasoning Forum and New Europe College
Preface

In July 2000, the Advanced Reasoning Forum in conjunction with New Europe College held its Second Annual Conference on Reasoning and Logic. Organized by Mircea Dumitru, a Relink Program Fellow at New Europe College, a series of talks were given at New Europe College and in the countryside of Romania. As at the other Conferences on Reasoning and Logic of the Advanced Reasoning Forum, the talks were on works in progress, meant to stimulate discussion and research. Now, a year later, we are pleased to be able to present some of the results of those discussions.

We are also pleased to have this opportunity to thank not only Mircea Dumitru, but also Anca Oroveanu, Scientific Director of the New Europe College, who arranged many of the details of the conference and ensured the publication of these proceedings. We hope that our work here will help New Europe College to further its goals.

New Europe College is a small independent Romanian center of excellence in the humanities and social sciences founded in 1994. Its aims and purposes include:

• Creating an institutional framework with strong international links in order to offer young scholars from Romania and South-Eastern Europe working conditions similar to those in the West, and to provide an environment that stimulates contact between scholars from different countries and fields of research while encouraging critical debate.
• Promoting contacts between Romanian and regional scholars and their peers worldwide.
• Cultivating the receptivity of scholars and academics in Romania to methods and areas of research not yet firmly established in Romania, while preserving what is still good in the work developed, against all odds, in the difficult intellectual, cultural, and political milieu in Romania before 1989.
• Contributing to the formation of a core of promising young scholars who are expected to play a significant role in the renewal of Romania’s academic, scholarly, and intellectual life.
The Advanced Reasoning Forum was founded to promote research in the theory and pedagogy of the art of reasoning well. The founding meeting and first Conference on Reasoning and Logic was held in Cedar City, Utah USA in September 1999; the Second Conference on Reasoning and Logic was held in Bucharest, Romania in July 2000; and the Third Conference on Reasoning and Logic was held in Berkeley, California USA in July 2001. A description of the proceedings of those conferences along with further information about ARF can be found at the website <www.AdvancedReasoningForum.org>.

At the second meeting the members decided to begin a journal on logic and epistemology—in the broadest sense of the art of reasoning well and knowledge as derived from reasoning—that would be of interest to scholars, teachers, and students. This includes work in philosophy, mathematics, computer science, psychology, formal logic, linguistics, and any other discipline in which prescriptive or descriptive aspects of the art of reasoning are studied.

Thanks to New Europe College we have been able to begin this project with this inaugural issue of the Bulletin of Advanced Reasoning and Knowledge. It consists of works that were presented at the meeting at New Europe College and since refined from the discussions there, as well as other works stimulated by discussions at that meeting.

Each issue of the Bulletin of Advanced Reasoning and Knowledge will have two sections:

- Theory for papers concerned with the foundations of logic and reasoning, as well as applications of those concepts to other disciplines.

- Pedagogy and Examples for: (i) Presentations on the teaching of logic and reasoning, with special emphasis on critical thinking, and (ii) Examples that illustrate or call into question theoretical issues in the foundations of logic and reasoning, and examples of argument analysis.

At this time BARK will be limited to papers developed out of presentations at meetings of ARF, along with papers presented to the Forum by invitation of a member.

Come, let us reason together.

© 2001 Advanced Reasoning Forum and New Europe College
Arguments and Explanations

Richard L. Epstein

Reasoning is the art of saying "therefore": How to determine whether the truth of one claim follows from the truth of some others. There are many different ways to say "therefore." In this paper the relation between two of those, arguments and explanations, will be discussed.

To say the truth of one claim follows from the truth of others, we first need to be clear what claims are.

Claim A claim is a declarative sentence that is used in such a way that it is either true or false, but not both.

Claims are the smallest parts of spoken or written language with which we reason that we can call "true" or "false." This is not to deny that there are abstract objects—propositions—that are the real things that are true or false. If there are, then perhaps claims are just our imperfect representations of those. But claims are what we use in reasoning.

In trying to understand what we mean by saying one claim follows from others, it's best to consider first the archetype of saying "therefore."

Argument An argument is a collection of claims; one is called the conclusion, whose truth the argument is intended to establish. The others, called the premises, are meant to lead to, or support, or convince that the conclusion is true.

Arguments are attempts to convince, whether someone tries to convince you, or you try to convince someone else, or you try to convince yourself. But that does not mean the criterion for whether an argument is good is whether the argument actually does convince. If I am drunk, you may give me an excellent argument that my driving home is dangerous; though I remain unconvinced, the argument is no worse. A politician may make a bad argument that you should vote for him, but just because you are convinced does not make the argument good. Perhaps other ways to convince, such as entreaties, exhortations, sermons, and advertisements, can be judged by how well they convince, but attempts to establish the truth of a claim cannot be so judged.

---

1 This is a revised version of a talk given at the second Advanced Reasoning Forum meeting sponsored by New Europe College in Bucharest, Romania. Many of the ideas and the exposition were refined over the last two years through discussions with the members of the Advanced Reasoning Forum. A much fuller account of this work can be found in my Five Ways of Saying 'Therefore'. A textbook version of these ideas can be found in my Critical Thinking, and in the Science Workbook for that text. I am grateful to Alex Raffi for providing the illustrations.

© 2001 Richard L. Epstein.
Good argument  An argument is good if its premises give good reason to believe the conclusion is true.

This seems a very imprecise standard, if any standard at all. What does “good reason” mean?

First, note that from a false premise we can derive a true claim or a false claim. For example,

(1) All authors of books on arguments are women.
    Richard L. Epstein is an author of a book on arguments.
    Therefore, Richard L. Epstein is a woman.

(2) All authors of books on arguments are women.
    Richard L. Epstein is an author of a book on arguments.
    Therefore, Richard L. Epstein is a human being.

The first of these has a false conclusion, the second a true one.

So it would seem that a good argument should have true premises. But that is too strong a condition. Consider:

There are an even number of stars in the sky.
So the number of stars in the sky can be divided by 2.

There are an odd number of stars in the sky.
So the number of stars in the sky cannot be divided by 2.

One of these has a true premise, but we cannot tell which. A standard that gives us no way to evaluate arguments is not part of the art of reasoning. Rather, for an argument to be good, we must have good reason to believe its premises; that is, the premises must be plausible.

What counts as good reason to believe a premise? That is not for the logician to say. It may depend in part on the subject matter: The biologist, the car mechanic, the professional football player, the physicist all have their own standards. It may depend in part on the metaphysics we adopt: Some say we never have good reason to believe claims whose truth cannot be ascertained through empirical tests. At best, we can give some rough standards for when to accept unsupported claims, which I present in Critical Thinking.

So a good argument should have plausible premises. But more is needed. Consider:

(3) Richard L. Epstein speaks English.
    Richard L. Epstein wrote Critical Thinking.
    Therefore, Richard L. Epstein lives in the United States.

Each of these claims is highly plausible. But the conclusion does not follow from the premises. What, though, do we mean by saying that the conclusion follows from the premises? There are two different standards.
**Valid argument**  An argument is valid if it is impossible for the premises to be true and conclusion false (at the same time).

For example,

(4) Maria is a widow.
So Maria was married.

It is not possible for the premise of this argument to be true and conclusion false.

In our daily lives, however, we often cannot employ valid arguments. For example, Dick hears that there are parakeets for sale at the mall. He knows that his neighbor has an old birdcage in her garage. He wonders whether the cage will be big enough for a parakeet if he buys one. He reasons:

(5) Every parakeet I or anyone I know has seen, or read, or heard about is less than 60 cm tall.

Therefore, the parakeets on sale at the mall are less than 60 cm tall.

This argument is not valid. There could be a new kind of parakeet discovered in the Amazon that is 1 meter tall; or a new supergrow bird food has been developed that makes parakeets grow very tall; or aliens have captured some parakeets and shot them with rays to make them very large; or . . . But any possibility that we can think of for the premise to be true and conclusion false is very unlikely—so unlikely that Dick has good reason to believe the conclusion.

**The Scale from Strong to Weak**  We classify invalid arguments on a scale from strong to weak. An argument is strong if it is very unlikely that the premises could be true and conclusion false (at the same time). An argument is weak if it is not unlikely that the premises could be true and conclusion false (at the same time).

Here “very unlikely” means relative to what we know.

In reasoning in our lives, and in almost every area of science, we rely on strong arguments that we cannot replace with valid arguments. For example, replacing the premise of (5) with “All parakeets are less than 60 cm tall” would yield a worse argument, for that claim is less plausible than the premise of (5). Indeed, there is often a trade-off between how plausible the premises of an argument are and how strong the argument is: the less plausible the premises, the stronger the argument.

So for an argument to be good, it should be either valid or strong. Argument (3), for example, is bad because it is weak: It is not unlikely that I could speak English and write *Critical Thinking* while living in Barbados.

The standard of whether an argument is strong is subjective. But it is not unusable, any more than the judgment of whether an auditorium lit at one end by a candle is dark at one end and light at the other. In practice, we almost always agree on the evaluation of the strength of an argument once we exchange the background assumptions on which we base the evaluation.

In any case, it is not just the criterion of whether an argument is strong that
introduces a subjective element into judging whether an argument is good. The question of whether the premises of an argument are plausible is also subjective. Even the classification of a collection of claims as an argument is subjective. It may be that there are collections of abstract propositions that satisfy objective criteria that our notion of strong argument imperfectly attempts to describe. But those abstract objects will be of little use to us in our reasoning.

In order for an argument to be good, it is not enough for it to be valid or strong and have plausible premises. Consider:

Every dog has a soul.
Therefore, dogs should be treated humanely.

Even if you find the premise plausible, it is not more plausible than the conclusion. For an argument to give us good reason to believe the conclusion, its premises should be more plausible than its conclusion.

**Begging the question** Any argument that uses a premise that is not more plausible than the conclusion is said to **beg the question**.

Many valid arguments beg the question, for example (4) above. But not every valid argument begs the question. For example, “George is a duck. If George is a duck, then George quacks. So George quacks.”

We now have three tests that an argument must pass in order to be good.

**Necessary Conditions for an Argument to be Good**
- The premises are plausible.
- The argument is valid or strong.
- The argument does not beg the question.

In our daily lives there are many arguments that we classify as good that do not seem to satisfy these conditions. For example,

Lee: Tom wants to get a dog.
Maria: What kind?
Lee: A dachshund. And that’s really stupid, since he wants one that will chase a frisbee.

Lee has made an argument, if we interpret rightly what he said: Tom wants a dog that will chase a frisbee, so Tom shouldn’t get a dachshund. But on the face of it that argument is not strong or valid. Still, Maria knows very well, as do we, that a dachshund would be a bad choice for someone who wants a dog to chase a frisbee. Dachshunds are too low to the ground, they can’t run fast, they can’t jump, and the frisbee is bigger than they are, so they couldn’t bring it back. Any dog like that is a bad choice for a frisbee partner. Lee just left out these obvious claims, but why should he bother to say them?

We normally leave out so much that if we look only at what is said, we will be missing too much. We can and must rewrite many arguments by adding an unstated premise or even an unstated conclusion.
Arguments and Explanations

When are we justified in adding an unstated premise? How do we know whether we’ve rewritten an argument well or just added our own prejudices? To “repair” arguments that are apparently defective, we must have some standards. Otherwise we will end up putting words in someone’s mouth. Such standards depend on what we can assume about the person with whom we are reasoning or whose work we are reading.

The Principle of Rational Discussion We assume that the other person with whom we are deliberating or whose arguments we are reading:

- Knows about the subject under discussion.
- Is able to reason well.
- Is willing to reason well.
- Is not lying.

Often people with whom we wish to reason do not satisfy these conditions. But then it makes no sense to reason with that person. We should be educating, or consoling, or pointing out the errors in what he or she says.

The Principle of Rational Discussion justifies adopting the following guide for when we can add or delete claims in an argument.

The Guide to Repairing Arguments Given an (implicit) argument that is apparently defective, we are justified in adding a premise or conclusion if and only if all the following hold:

- The argument becomes strong or valid.
- The premise is plausible and would seem plausible to the other person.
- The premise is more plausible than the conclusion.

If the argument is valid or strong, yet one of the original premises is false or dubious, we may delete that premise if the argument becomes no worse.

Given only this Guide, we might try to repair every argument into a good one. That would be wrong, for there are clear standards for when an argument is unrepairable.

Unrepairable Arguments We cannot repair a (purported) argument if any of the following hold:

- There is no argument there.
- The argument is so lacking in coherence there’s nothing obvious to add.
- A premise is false or dubious or several premises are contradictory and cannot be deleted.
- The obvious premise to add would make the argument weak.
- Any obvious premise to add to make the argument strong or valid is false.
- The conclusion is clearly false.

It is not that when we encounter one of these conditions we can be sure the speaker had no good argument in mind. Rather, we are not justified in making that argument for him or her—it would be putting words in the speaker’s mouth.
In addition to these conditions for an argument to be unrepairable, a list of other kinds of arguments, called fallacies, have been deemed to be typically so bad that they, too, are rejected as unrepairable when we encounter them.²

In Critical Thinking I set out how the Principle of Rational Discussion justifies adopting these guides, and I give many examples how these standards are useful in evaluating arguments, from arguments in our daily lives to those we encounter in science journals. But one particular kind of argument that is related to explanations seems to require different standards of evaluation.

Generalization An argument in which we conclude a claim about a group, called the population, from a claim or claims about some part of it, called the sample. Sometimes the conclusion of the argument is called a “generalization.”

For example, the following is a generalization: “All dogs that I have ever met except for one can bark; so almost all dogs bark.” Some standards for a generalization to be good are the following.

Necessary Conditions for a Generalization to be Good A good generalization requires as premises the following three claims (whether stated not):

1. The sample is representative.
2. The sample is big enough.
3. The sample is studied well.

But these are not different standards than the necessary conditions for an argument to be good. They only spell out in more detail what is required for the argument to be strong and have plausible premises. Still, it remains open whether the necessary conditions for an argument to be good are also sufficient. That is a difficult issue to resolve, and takes up much of Five Ways of Saying “Therefore.”

Finally, we need to note one particular mistake in reasoning with arguments, for it is surprisingly common and is related to the use of explanations. Some people, when encountering an argument that is valid or strong which has a clearly true conclusion, conclude that the premises are plausible. But that’s wrong, as argument (2) shows.

Arguing backwards To conclude that the premises of an argument are true because the argument is valid or strong and its conclusion is highly plausible.

What in the study of arguments can we carry over to all other reasoning? It cannot be the requirement that the premises be plausible. Proofs in mathematics are often acceptable even when that is not the case.

It cannot be that the premises are more plausible than the conclusion. A good explanation of why the sky is blue will certainly use claims more dubious than “The sky is blue.”

What must be analyzed in any reasoning is the relationship of the premises to the conclusion.

² See Five Ways of Saying “Therefore” or Critical Thinking.
Arguments and Explanations

Inference A collection of claims, one of which is designated the conclusion and the others the premises, that is intended to be judged as valid/invalid or on the scale from strong to weak.

An inference is valid if it is impossible for the premises to be true and conclusion false (at the same time). An inference is strong if it is very unlikely that the premises could be true and conclusion false (at the same time). An inference is weak if it is not unlikely that the premises could be true and conclusion false (at the same time).

An argument is an inference that is meant to convince that the conclusion is true. This is “therefore argument”.

Many explanations are inferences, too. Consider:

Why is the sky blue? Because sunlight is refracted through the atmosphere so as to absorb other wavelengths of light.

“The sky is blue” is explained in terms of why it is true—what it follows from, the reasons for its truth.

Not every explanation can be understood as an inference. For example, when a traveler asks a policeman to explain how to get to the Post Office, she’s not asking him to show why some claim is true.

For an explanation that can be judged as an inference, the conclusion, what’s being explained, should be highly plausible: We can’t explain why the sun rises in the west. The claims that do the explaining should be plausible, too. We don’t accept “The sky is blue, because there are blue globules high in the atmosphere” as a good explanation, because “There are blue globules high in the atmosphere” is known to be false. But for an explanation the premises shouldn’t be more plausible than the conclusion, for otherwise we’d have an argument.

Moreover, the inference should be valid or strong. We don’t accept “Dogs lick their owners because they aren’t cats” because the inference is neither valid nor strong and there is no obvious way to repair it. As with arguments, we allow that the inference might be repaired: We understand that an explanation “E because of A” may require further premises to supplement A. And as with arguments, we can invoke the Principle of Rational Discussion to motivate the Guide to Repairing Arguments, except for the condition that the premises be more plausible than the conclusion.

Inferential explanations An explanation that can be judged as an inference, “E because of A, B, C, . . .”. For it to be good, all the following must hold:

1. E is highly plausible.
2. Each of A, B, C, . . . is plausible, but at least one of them is not more plausible than E.
3. “A, B, C, . . . therefore E” is a valid or strong inference, possibly with respect to some plausible unstated claims.
4. The explanation is not “E because of D” where D is E itself or a simple rewriting of E.
We call E the *explanandum* and collectively A, B, C, . . . the *explanans*. Sometimes the explanans alone is called “the explanation.”

It is often said that an explanation must answer the right question. For example,

**Mother:** There were two pieces of cake in the cupboard. Why is there only one now?

**Flo:** Because it was dark and I didn’t see the other piece.

Flo thinks she has given a good explanation: Her answer makes it clear why the claim “There is only one piece of cake in the cupboard now” is true (assuming some other fairly obvious claims). But her mother won’t accept it. Flo answered “Why is there only one piece of cake in the cupboard, instead of none?” but her mother meant, “Why is there only one piece of cake in the cupboard, instead of two?”

Questions are often ambiguous, and a good explanation to one reading of a question can often be a bad explanation to another. If the explanandum is ambiguous, then that is a fault of the questioner; we should not be expected to guess correctly what’s meant. Still, when it is clear that a different reading of the explanandum is meant, we can say that an explanation is bad: It’s answered the wrong question.

How are explanations and arguments related? A good explanation is not a good argument. Consider what Zoe said to Dick Sunday morning:

You drank three cocktails before dinner, a bottle and a half of wine with dinner, and then a couple of glasses of brandy. Anyone who drinks that much is going to get a headache. So you have a headache.

Zoe has given a good explanation of why Dick has a headache. But it is a bad argument, because it begs the question: It is much more plausible to Dick that he has a headache than that anyone who drinks that much is going to get a headache. With a little rewriting, this shows that aristotelian syllogisms that are faulted as begging the question are often perfectly good attempts to codify or explain.

For the relation of particular explanations to arguments, consider the following. Dick, Zoe, and Spot are out for a walk in the countryside. Spot runs off and returns after five minutes. Dick and Zoe notice that Spot has blood on his muzzle. And they both really notice that Spot stinks like a skunk. Dick turns to Zoe and says, “Spot must have killed a skunk. Look at the blood on his muzzle. And he smells like a skunk.” Dick has made a good argument:

Spot has blood on his muzzle. Spot smells like a skunk.

Therefore, *argument* Spot killed a skunk.

He’s left out some premises that he knows are as obvious to Zoe as to him:

Spot isn’t bleeding.

Skunks aren’t able to fight back very well.
Normally when Spot draws a lot of blood from an animal that is
smaller than him, he kills it.

Only skunks give off a characteristic odor, an odor that drenches
whoever or whatever is near if they are attacked.
Dogs kill animals by biting them and typically drawing blood.

Zoe replies, “Oh, that explains why he’s got blood on his muzzle and smells
so bad.” That is, Zoe takes the same claims and uses them to make a good
explanation, relative to the same unstated premises:

Spot killed a skunk.
Therefore_{explanation} Spot has blood on his muzzle and smells like a skunk.

Explanations and associated arguments  Given an explanation:
A therefore_{explanation} E  (relative to some other premises P, Q, R, . . . )
the associated argument is:
E therefore_{argument} A  (again relative to P, Q, R, . . . )

For an explanation with many claims in the explanans, $\Sigma$ therefore_{explanation} E,
reversing the role of E with any one of the claims in $\Sigma$ is an associated argument.

An independent explanation is one in which each premise that is less
plausible than the explanandum can be established by an associated argument.
An explanation is dependent if the reason to believe at least one of the premises
must be established by claims outside the explanation.

Zoe’s explanation of why Spot has blood on his muzzle and smells bad is
independent, because the associated argument that Dick makes is good: It
establishes that Spot killed a skunk.

But many explanations are dependent. Consider what Dick told his
neighbor’s little girl:

Spot chases cats because he sees cats as something good to eat and
because cats are smaller than him.
Are the claims in the explanans plausible? Certainly “Cats are smaller than Spot” is plausible. But how about “Spot sees cats as something good to eat”? What reason do we have to accept this? The associated argument for it is:

Spot chases cats, and cats are smaller than Spot.
Therefore, Spot sees cats as something good to eat.

This is weak. So the explanation is dependent. Without more evidence for “Spot sees cats as something good to eat,” we shouldn’t accept the explanation.

Any explanation with a generalization in the explanans is likely to be dependent, too. For example,

Dick offers an explanation:

The oar appears bent because light is bent where the water meets the air.

The explanans here is a generalization: “Light is bent where the water meets the air.” Zoe hasn’t taken a physics course, so the only reason she has to believe that claim is the associated argument:

The oar appears bent, therefore argument light is bent where the water meets the air.

But this is a weak generalization, needing more examples to convince. So Dick has given a dependent explanation. Dick has other reasons to believe the claim from his studies in physics, which he can offer to Zoe to make this explanation good.

Testing explanations is often how we establish a generalization. For example, consider what the children Flo and Becky were saying last week:

Flo: Spot barks. And Wanda’s dog Ralph barks. And Dr. E’s dogs Anubis and Juney bark. So all dogs bark.
Becky: Yeah. Let’s go over to Maple Street and see if all the dogs there bark, too.

Flo is generalizing. Relative to her experience it’s a pretty good generalization. Becky wants to test the generalization.

Suppose that A, B, C, D are given as inductive evidence for a generalization G. (Some other highly plausible unstated premises may also be needed, but we’ll keep those in the background.) Then we have that G explains A, B, C, D.
But if G is true, we can see that some other claims must be true, instances of the generalization G, say L, M, N. If those are true, then G would explain them, too. For example, Rodolfo barks, Lady barks, Fido barks, \ldots

That is, G explains A, B, C, D and predicts L, M, N, where the difference in this case between the explanation and the prediction is that in the explanation we know the conclusion is true, whereas we don’t know if the predictions are true.

Suppose we find that L, M, N are indeed true. Then the argument “A, B, C, D + L, M, N therefore G” is a better argument for G than we had before. At the very least it has more instances of the generalization as premises.

But how do more instances of a generalization prove the generalization better? They can if (i) they are from different kinds of situations, that is, A, B, C, D + L, M, N cover a more representative sample of possible instances of G than do just A, B, C, D. This is typically what we do: We deduce claims from G for situations that we had not previously considered.

And (ii) because we had not previously considered the kind of instances L, M, N of the generalization G, we have some confidence that we haven’t got G by manipulating the data, selecting situations that would establish just this hypothesis.

The best way to test an hypothesis-generalization, it’s often said, is to try to falsify it. Trying to falsify the generalization just means that we are consciously trying to come up with instances of the generalization to test that are as different as we can imagine from A, B, C, D. Trying to falsify is a good way to ensure (i) and (ii). So we say that an experiment confirms the explanans if it shows that a prediction is true. Confirmation amounts to strengthening the associated argument.

Here is an example of this relation between explanation and prediction:

Consider the explanation offered by Torricelli for a fact that had intrigued his teacher Galileo; namely, that a lift pump drawing water from a well will not raise the water more than about 34 feet above the surface of a well. To account for this, Torricelli advanced the idea that the air above the water has weight and thus exerts pressure on the water in the well, forcing it up the pump barrel when the piston is raised, for there is no air inside to balance the outside pressure. On this assumption the water can rise only to the point where its pressure on the surface of the well equals the pressure of the outside air on that surface, and the latter will therefore equal that of a water column about 34 feet high.\textsuperscript{3}

That is, Torricelli offered an explanation, but the only evidence he had for the explanans, which was a generalization, was the explanandum.

The explanatory force of this account hinges on the conception that the earth is surrounded by a “sea of air” that conforms to the basic laws governing the equilibrium of liquids in communicating vessels. And because Torricelli’s explanation presupposed such general laws it yielded predictions concerning as yet unexamined phenomena. One of these was that if the

water were replaced by mercury, whose specific gravity is about 14 times that of water, the air should counterbalance a column about 34/14 feet, or somewhat less than 2 1/2 feet, in length. This prediction was confirmed by Torricelli in the classic experiment that bears his name. In addition, the proposed explanation implies that at increasing altitudes above sea level, the length of the mercury column supported by air pressure should decrease because the weight of the counterbalancing air decreases. A careful test of this prediction was performed at the suggestion of Pascal only a few years after Torricelli had offered his explanation: Pascal’s brother-in-law carried a mercury barometer (i.e., essentially a mercury column counterbalanced by the air pressure) to the top of the Puy-de-Dôme, measuring the length of the column at various elevations during the ascent and again during the descent; the readings were in splendid accord with the prediction.

Predictions are made of further instances of the generalization or of consequences of the claim in the explanans; those are shown to be true; the explanans thus becomes more plausible because the associated argument for it (adding as premises all the instances of the generalization that have been tested and found to be true) is strengthened. The story is much the same for explanans that aren’t generalizations, too.

This relation of arguments and explanations is often misunderstood and used badly. Some scientists believe that if you have an explanans that could explain a lot, it must be true. For example, Charles Darwin said:

It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified [the geographical distribution of species, the existence of vestigial organs in animals, etc.]. It has recently been objected that this is an unsafe method of arguing; but it is a method used in judging of the common events of life, and has often been used by the greatest natural philosophers.

Darwin was arguing backwards: From the truth of the conclusion(s), we can infer the truth of the premises. The direction of inference is incorrect. Rather, the conclusions that are drawn from the explanans together serve as evidence for the explanans, not because the explanans gives the best explanation of them, but because they are premises for an argument concluding with the explanans.

Gilbert Harman, however, thinks Darwin’s method of arguing is right if the explanation is the best:

In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making an inference. Thus one infers, from the premise that a given

---

4 Ibid.
5 On the Origin of Species, p. 476. Paul R. Thagard presents further examples where scientists reasoned in this way in “The best explanation: criteria for theory choice”
Arguments and Explanations

hypothesis would provide a “better” explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true.6

This method of argument he calls inference to the best explanation.

But even if we had clear criteria for what we mean by “best explanation,” which we certainly don’t, it wouldn’t rescue inference to the best explanation from the charge of arguing backwards. Consider what happened to me a few years ago:

(In the hospital emergency room at 2 a.m.)

Me: Doctor, doctor, why do I have such pain in my back?
It doesn’t feel like a muscle cramp or a pinched nerve.

Doctor: (after examining me) A kidney stone would explain the pain.
Kidney stones give that kind of pain, and it’s in the right place for that. I can’t think of anything else that would give you that much pain there.

The doctor offered me the best explanation he had:

Your back hurts this way because you have a kidney stone.

This would have been a good explanation if we’d had good reason to believe the explanans. But at that point the only evidence for the explanans was the associated argument, and that was not strong.

So the doctor made predictions from the explanans: “A kidney stone would show up on an X-ray,” “You would have an elevated white-blood cell count,” “You would have blood in your urine,” “A kidney stone will show up on a CAT-scan.” He tested each of these and found them false. He reasoned by reductio ad absurdum that if the explanans were true, one or more of these would be true; they are false; therefore, the explanans is very likely false. Hence his original explanation turned out to be bad. Nothing else was found, so by process of elimination it was conjectured that I had a severe sprain or strain, for which exercise and education were the only remedy.

If inference to the best explanation were a good method of argument, there would have been no point in doing tests. The doctor and I should have believed “You have a kidney stone.” But I’m glad we didn’t, or I’d have undergone needless surgery.

Still, you might say that in this example the explanation the doctor offered was not good, since “You have a kidney stone” was not plausible. So we didn’t have reason to believe it. Inference to the best explanation doesn’t fail here.

But if we require in an inference to the best explanation that the explanation be not only better than all the others but good as well, then we don’t need inference to the best explanation. The explanans of a good explanation is plausible, perhaps because of the associated argument, or another argument, or just by inspection. We have good reason to believe it. But the explanans of the best explanation need only be more plausible than any (or all) of the explanans of

6 “The inference to the best explanation”, p. 89.
the other explanations we can think of. Ruling out all but one explanation does not by itself show that that one is plausible. It could equally show how bad our imagination is. We need further evidence.

Finding an explanation that is better than all others does not justify belief in the explanans. It only provides motive for us to investigate whether the proposed explanans is true, which is exactly how the doctor saw it. Used that way, inference to the best explanation is called abduction.\(^7\)

**Fallacy of inference to the best explanation** An argument “These claims give the best explanation, so they are true.”

Instead of “That’s the best explanation we have, so it’s true,” think “That’s the best explanation we have, so let’s investigate it.” Instead of “That’s the only explanation of . . .” say “All the evidence points to . . .”.

Inference to the best explanation is what stands behind many claims that there are numbers and sets, abstract objects of that sort. Mathematics and science “need” them, in the sense that they best explain why our mathematical and scientific theories are true. Certainly if those objects exist, they explain that. But, as in any use of inference to the best explanation, we must ask what other evidence we have to believe the claim “Numbers, as abstract objects, exist,” since the inference from these scientific theories being “true” to numbers and sets existing is weak.\(^8\)

Inference to the best explanation is no better in mathematics than in daily life. The difference, it seems, is that in mathematics there is no other evidence we can cite for “Numbers, as abstract objects, exist.” Mathematics, for the platonist, is built on faith; and the necessity of numbers for mathematics—all numbers, natural, rational, real, in their abstract plentitude—is a guide, a sign towards that faith.

\(^7\) C. S. Peirce saw this clearly (Vol. 5.189 and Vol. 5.171):

The surprising fact, C, is observed.
But if A were true, C would be a matter of course.
Hence, there is reason to suspect that A is true.

Deduction proves that something must be; Induction shows that something actually is operative; Abduction merely suggests that something may be.

\(^8\) W. D. Hart, Introduction to The Philosophy of Mathematics, p. 6:

Sophisticated natural science as it comes is always to be formalized in an extension, in the logician’s sense, of some mathematics, often number theory and analysis. Equations are obvious to anyone reading serious science. So, by abduction [inference to the best explanation], we are justified in believing true at least as much mathematics as we need for the best scientific explanations of what we observe.
Since the truth of that much mathematics requires very abstract objects, Quine thereby began an empiricist justification for belief in the abstract objects required for mathematical truth.
Conclusion
Arguments are different from explanations. By carefully distinguishing the criteria for what counts as a good argument from what counts as a good explanation, we can see that arguments and explanations are related. It is not inference to the best explanation that turns explanations into arguments, but simply reversing the roles of premise and conclusion, which accounts for the notion of confirmation of an hypothesis. To argue well, we must be able to distinguish different ways of saying “therefore.”

Bibliography
Darwin, Charles
Epstein, Richard L.
  2001  Five Ways of Saying “Therefore”, Wadsworth.
Harman, Gilbert
Hart, W. D., ed.
  1996  The Philosophy of Mathematics, Oxford University Press.
Hempel, Carl G.
Peirce, Charles
Thagard, Paul R.
Things That Aren’t There

Fred Kroon

Introduction
Let me begin with a bunch of simple claims, all seemingly true, none very startling in its own right.

(1) 7 is a prime number.
(2) Loyalty is a virtue.
(3) The proposition that snow is white is true.
(4) The golden mountain is a golden mountain.
(5) Sherlock Holmes lived on Baker Street.
(6) Holmes doesn’t exist.
(7) Holmes is a much-admired fictional detective.

To see what the problem is, let the Naive View of Predication be the following claim:

A sentence $a$ is $F$ is true to the extent that the object designated by $a$ has the property designated by $F$, while $a$ is not $F$ is true to the extent that the object designated by $a$ does not have the property designated by $F$. Similarly, $a$ bears [does not bear] $R$ to $b$ is true to the extent that the objects designated by $a$ and $b$ stand [do not stand] in the relation designated by $R$; etc.

The Naive View is a view about how to understand the truth-making underpinnings of predications. It implies (presumably correctly) that a sentence like “Bill Clinton is tall” is true to the extent that a certain individual, Bill Clinton, has the property of being tall. But it also implies, for exactly the same reasons, that “7 is a prime number” is true to the extent that a certain object, the number 7, has the property of being a prime number. Similarly, “Loyalty is a virtue” is true to the extent that a certain attribute, loyalty, has the property of being a virtue, “The proposition that snow is white is true” is true to the extent that a certain object, the proposition that snow is white, has the property of being true, “Sherlock Holmes doesn’t exist” is true to the extent that a certain object, Sherlock Holmes, fails to have the property of existence, and so on.

But these claims about truth-conditions import reference to entities that have in one way or another often been thought problematic. (1)-(3), for example, import reference to numbers, attributes, and propositions—entities that are abstract rather than concrete, and so are rejected by nominalistically-minded

---

1 I am much indebted to participants at the New Europe College conference for their very useful comments. Special thanks, as always, to my fellow ARF members, especially to Mircea and Dick (GA) for making the ARF-in-Romania experience possible.

© 2001 Fred Kroon.

philosophers. In rejecting such entities, nominalists are then required to do something about the offending predications. Some will be error theorists, banishing the offending terms and predications because they are infected with ontological error. Others will blame the language, seeing here an exception to the Naïve View and offering rephrasing. Some, like W. V. O. Quine, are only partly nominalist, accepting numbers on grounds of their indispensability to science, but rejecting an ontology of attributes in favour of a regimented rephrasing of the common sense language of attributes.

Most of us, however, incline to forms of moderate realism about things not in space and time, rather than Quine’s grudging realism. We accept what the Naïve View tells us about ordinary predications concerning physical objects, and also accept what it tells us about number-talk and (here departing from Quine) attribute-talk. We are happy to accept that there are numbers and that their properties make mathematical sentences true or false, and that there are such things as attributes which make claims like (2) true. But even moderate realists tend to balk at some of what the Naïve View implies. As we have seen, that view not only licenses the thought that there are such things as numbers and attributes, but also that there are purely fictional objects like Hamlet and Holmes as well as mythological objects like the golden mountain, which all have properties and stand in relations to other objects—not don’t exist! Moderate realists, who like their realism in moderation, find this hard to swallow. From their point of view the Naïve View implies far too much.

But that leaves moderate realists—most of us—with the problem of how to fix the Naïve View. Should we be error theorists, and recommend that we give up talking the way we do about Holmes and the golden mountain? Or should we say that our apparent ontological commitments are merely apparent, that such talk can be given an innocent construal that doesn’t commit us to a strange ontology of non-existent objects? Alternatively, should we bite the bullet and accept that even moderate realism hasn’t gone far enough? Should we embrace non-existent objects after all, on the grounds that we encounter them in both language and thought, much as the Naïve View suggests?

What I want to do in this paper is to survey a number of options. My purpose is not to argue that one or another view is clearly right, but to show you that there are interesting and sharply conflicting solutions to the problem, each carrying a cost. There is little consensus about who is right, although my survey will certainly show which side I favour.

**Meinong**

No doubt (1) to (7), especially (4)-(7) will have put the reader in mind of Meinong.² Rightly so: Meinong is the supreme realist, inclined, as it seems, to let every meaningful term have an object as its denotation. He famously allowed propositions/states of affairs (objectives) into the class of objects, as well as objects like the golden mountain and the fountain of youth. He even allowed impossible objects such as the round square. Some of these objects are

² See, for example, A. Meinong, “The theory of objects”.
abstractions, by their very nature not to be found in the real world of space and
time, while others are (would-be) spatio-temporal in nature. Propositions,
numbers, relationships, and so on, are of the former kind, and these have the
potential to have a kind of abstract or ideal existence that Meinong calls
subsistence. Objects like Mt. Everest and the golden mountain have the potential
to have concrete existence (Meinong simply calls this “existence”). Being either
abstract or concrete, objects have the potential either to subsist or to exist. But not
all abstract objects subsist; states of affairs that don’t obtain (e.g., Bill Clinton’s
being a virgin) do not, nor could impossible mathematical objects like the round
square. And not all concrete things exist; the fountain of youth does not, nor does
the golden mountain. Meinong famously claimed that such non-subsistent and
non-existent objects don’t have any form of being. That, however, seems merely
a semantic choice on Meinong’s part, since we can surely choose to define a new
weak and general form of being that any object has simply in virtue of being an
object (“O has weak being iff O is an object”).

In any case, Meinong’s ontology is extreme indeed. He thought that
d Judgments like (1)-(7) show that our thought is sometimes directed at objects like
the number 7, the golden mountain, Holmes, and so on, and not just at real spatio-
temporal objects like Clinton, Mt. Everest, or my favourite arm-chair. He thought
that we should therefore count judgments like (1)-(7) as being true in virtue of
objects that in many cases don’t have concrete existence or even ideal, abstract
existence. He also thought that the properties of non-existing objects can be
recovered a priori from the way they are characterised. Thus we can tell a
priori that the golden mountain is golden and the round square is round and
square. Generalised, this becomes the claim that the object which is given as
being F₁, ..., Fₙ ipso facto has all of these properties, what Richard Routley
calls the Characterisation Postulate. Meinong also thought that non-existent
objects of this kind only have properties that are implied by their characterisation.
Thus the golden mountain is neither higher than 10,000 feet nor lower nor the
same height—even though, as a mountain, it presumably has some height or
other! Oddly, therefore, non-existent objects are typically incomplete with respect
to their properties. As for objects like Holmes or Hamlet, it is likely that Meinong
thought their characterising properties were the ones given in the texts that
produced the names.

... then along came Russell
Bertrand Russell was the main early critic of Meinong’s theory, after having
flirted with a Meinongian framework in his The Principles of Mathematics
(1903). By 1904, he had changed his mind, and he launched a seminal attack on
Meinong’s (and his own earlier) views in his reviews of Meinong’s work. His
two most famous arguments against Meinong were as follows: (a) Impossible

Routley, Exploring Meinong’s Jungle and Beyond.

See Russell’s reviews of Untersuchungen zur Gegenstandstheorie und Psychologie and
Über die Stellung der Gegenstandstheorie im system der Wissenschaften in Mind 14, 1905, and
Mind 16, 1907; reprinted in D. Lackey, Essays in Analysis.
objects like the round square infringe the law of contradiction, since they both do and do not have some property P. Hence Meinong’s theory can’t be true. (b) Meinong’s theory licenses clearly unsound ontological arguments. By the Characterisation Postulate, the existent golden mountain is not only golden but also exists, which is absurd.

Although these objections have often struck analytic philosophers as devastating, Meinong himself remained largely unperturbed. He argued against (a) that the law of contradiction only holds of the actual and the possible. And he argued against (b) that the existent round square is indeed existent, but that nothing untoward follows unless we are also forced to admit—which we are not — that it actually exists. Russell himself confessed that he could not see a difference between “exists” and “is existent,” but Meinong was really buying into a larger package here, one suggested by his student Ernst Mally. According to this way out, there are both nuclear and extranuclear properties. The usual properties we use to characterise things are nuclear (red, round, being a mountain, being a detective, living on Baker St., and so on), while the properties that things have by virtue of their actual impact on the real world are extranuclear: possibility, impossibility, existence, relationships to actual things such as being admired by Meinong, and so on. Meinong is suggesting that the Characterisation Postulate only works in terms of nuclear predicates, and that there is a way of construing “existent,” but not “exists,” as a nuclear predicate. So Meinong thought that he could reformulate his theory to avoid the sorts of difficulties Russell talks about.5

The real question, of course, is whether this way of saving the theory is worth the costs: both the ideological cost—the strangeness of the nuclear/extranuclear distinction—and the ontological cost. And that rather depends on whether there is any other, better, way of accounting for our apparent ability to talk and think and make true judgments about what does not exist. To this question Russell had his own answer, given in terms of his enormously influential theory of definite descriptions.6

Russell’s answer took a very simple form. Definite descriptions—expressions of the form “the F”—were not genuine referring terms purporting to stand for things in the world. A definite description like “the golden mountain” or “the man who won the US Presidential election in 2000” is at bottom a quantificational expression whose meaning can only be given by showing what it contributes to the meaning of an entire sentence. In particular, Russell held that “the F is G” is to be analysed as the conjunction of:

(i) There is at least one thing which is F,
(ii) there is at most one thing which is F,
(iii) this thing is G.

---

5 See T. Parsons Nonexistent Objects for a history of the nuclear/extranuclear distinction.
Existence statements were simpler. "The F exists" was to be analysed as the conjunction of (i) and (ii) (that is, "There is exactly one thing which is F"), so that any statement of the form "the F is G" entailed the corresponding existential statement "the F exists."

The upshot was that Russell had a way of analysing statements like "The golden mountain is a golden mountain" or "The golden mountain does not exist" without these analyses carrying any commitment to objects. On the proposed analysis, the former simply says, falsely, that somewhere in the world there is a unique golden mountain and it is golden and a mountain, while the latter says, truly, that it is not the case that there is a unique golden mountain. The expression "the golden mountain" is no longer seen as the logical subject of these sentences, contrary to the Naïve View which sees surface grammar as a reliable guide to logical grammar and truth-conditions.

On the surface, this doesn't answer Meinong, since Meinong thought it obvious that both sentences were true. But Russell can simply respond that Meinong's intuitions, like many intuitions in philosophy, should not be taken at face value in cases where speakers may find it hard to discern fine distinctions. When Meinong claims that it is a priori that "The golden mountain is golden," Russell will claim that a priority really attaches to the sentence "If it exists, the golden mountain is golden," a close neighbour of the former. On this score, there is no reason to prefer Meinong's account to Russell's.

What about a claim like (6), that Holmes doesn't exist? I said earlier that Meinong would probably have thought Holmes was describable in terms of the descriptions given in the stories. Russell is more explicit. Any name, unless it is one that attaches directly to sense-impressions or universals, is a description. That means that "Holmes doesn't exist" is logically equivalent to a sentence like "The man called 'Holmes' who lived at 221B Baker St, solved such-and-such crimes, had a friend called 'Watson', etc., doesn't exist", and that sentence obviously succumbs to the machinery of Russell's theory. Thus, (6) is true since there is no (unique) person in the world who meets these conditions.7

Sentence (5) is more difficult, since the non-existence of Holmes means that it comes out as false, contrary to our tendency to discern a sense in which it is true. But even here there is much to be said for Russell's theory. Russell might take the view that ordinary intuition again requires slight correction, that the real truth requiring analysis is "It is true in the Holmes stories that Holmes is a detective," where "It is true in X that . . ." is a wide-scope operator (like "N believes that . . ."). This would leave us with a Russelian translation like "It is true in the Holmes stories that there is a unique man called 'Holmes' who . . ., and he is a detective," which is clearly true.

Sentence (7) is far more difficult. Russell can perhaps make sense of the claim "Holmes is a fictional detective"—try "It is true in some fictional stories

---

7 Couldn't someone accidentally fit some story without the story being about him? To deal with this, Russell may need to allow an additional predicate like "is a person with whom the author of this text is acquainted" into any description abbreviating a fictional proper name.
that Holmes is a detective”—but not “Holmes is a much-admired fictional detective.” The latter resists any kind of “in the fiction” operator. It is true not because there is some kind of story in which it is true, but because real people find Holmes appealing and admirable\textsuperscript{8}—even though Holmes is a merely fictional detective and as such doesn’t exist! And as if that is not enough, consider quantified talk about fictional and mythological objects. Just as we can quantify over the objects referred to in (1)–(3) (“There are many qualities that are considered virtues,” for example), so we can quantify over fictional characters. Consider, for example, “There are fictional characters in Charles Dickens’ novels whose drollness and ineffectualness makes them seem both pathetic and intensely likeable.”

**Meinongian realism: another try**

Russell’s theory of descriptions therefore faces serious difficulties when it comes to talk about the non-existent.\textsuperscript{9} And that might suggest a return to forms of Meinongian realism. The most fully worked out version of Meinong’s own approach is found in the neo-Meinongian theories of Richard Routley and Terence Parsons, which accepts the distinction between nuclear and extranuclear properties.\textsuperscript{10} That approach earlier struck us as quite counterintuitive. Now we are saying the costs are perhaps worth incurring, given the difficulties facing Russell’s way out.

But there is also an alternative approach, most fully defended by Ed Zalta.\textsuperscript{11} Zalta denies the intelligibility of the nuclear / extranuclear distinction and suggests instead that there are two different ways in which something might be said to have a property.\textsuperscript{12} Something S might exemplify a property P, or it might encode it. The golden mountain encodes being golden and being a mountain (it is an abstract object somehow built up out of those properties), and to that extent we can truthfully say that the golden mountain is golden and a mountain. But it does not literally exemplify these properties; it only exemplifies

\textsuperscript{8} The claim is true even if the Holmes stories themselves claim that Holmes was cranky and universally disliked; that is, it may be true that Holmes is widely admired, even if in the fiction Holmes is universally disliked.

\textsuperscript{9} The theory also faces other, quite different difficulties. No matter how well it deals with definite descriptions, there are important reasons to think that it doesn’t cope at all well with proper names. See especially Saul Kripke, *Naming and Necessity*. Although Kripke doesn’t explicitly talk about empty, e.g., fictional, names, his arguments apply to those names as well.

\textsuperscript{10} Routley, “Rehabilitating Meinong’s theory of objects” and “Exploring Meinong’s Jungle and Beyond”, and T. Parsons, “Nuclear and extranuclear properties, Meinong, and Leibniz”. The approaches of these two philosophers differ markedly. Routley’s later work favors a radical paraconsistent logic and accepted the idea of true contradictions, while Parsons’ logic is far more conservative.

\textsuperscript{11} Zalta, *Abstract Objects and Intensional Logic and the Metaphysics of Intentionality*. See also P. van Inwagen, “Creatures of fiction” for an earlier, less developed, defence of the idea that fictional objects are abstract objects.

\textsuperscript{12} Like the nuclear / extranuclear distinction, this distinction between ways of possessing a property is also due to Meinong’s student, Ernst Mally.
properties like being abstract (and therefore non-golden), being thought about by Meinong, and so on. Indeed, to the extent that it is an abstract object it does not exist. (For Zalta, as for Routley and Parsons, non-existence is a genuine property.) Similarly, Holmes is an abstract object that encodes properties such as being a detective without exemplifying them. What is true according to the Conan Doyle stories, on the other hand, is that this object Holmes exemplifies being a detective.

Zalta’s theory has a great many theoretical advantages over its neo-Meinongian rivals. It captures our intuitive data well, and, importantly, allows us to continue using something like classical logic. The round square is round as well as non-round, not because the world contains impossibilities but because an object might encode both being round and being non-round. But nothing could ever exemplify both these properties, so the consistency of the natural world is not threatened. Zalta’s theory also allows us to avoid some of the extremely odd features of concrete Meinongian objects of the type favoured by Routley and Parsons. In particular, Zalta’s abstract objects are not incomplete. So long as it is meaningful to predicate F-ness of an abstract object X, we can truthfully say that X exemplifies either F-ness or its complement. What is not true is that such an object encodes either F-ness or its complement.

But Zalta’s theory also comes at a cost. We must not forget its ontological commitments. It accepts a whole new range of objects, the objects that encode properties. Of course these are abstract, not concrete, objects, and so they might strike moderate realists as sufficiently akin to other objects they accept (propositions and properties, say) not to be threatening. But that, I think, would be a mistake. Consider Holmes again. Holmes is admired by a great many of us because of his astuteness, his strength of will, and other admirable qualities. We don’t seem to be admiring an abstract encoding object, whatever that is. Our admiring seems just like the admiring we do in the case of ordinary people, except that the person in this case is fictional! That is a fundamental datum that Russell failed to capture, and that Zalta captures very poorly, despite the added luxury of a more elaborate ontology.

Meinong’s appeal to assumptions
We have seen that the Meinongian program, no matter how pursued, comes at a considerable cost. We have also seen that the contrasting view of Russell has some glaring deficiencies. What is not generally appreciated is that Meinong himself was aware of a radical alternative, one he pursued for a while until he decided it was too radical to do the job by itself.13 The story Meinong tells is found in his important but neglected work On Assumptions.14

Suppose, Meinong asks, that we want to know what some object O is like. To discover O’s properties it is not enough that I have a representation of O, perhaps a mental image. I must in addition apprehend or intend the object, where

---

13 See my “Was Meinong only pretending?”
14 Page references are to the English translation.
intending an object is an active mental act which makes knowledge of an object’s properties possible. The account of intending Meinong provides comes in two parts. If I believe that object O has some form of being (existence or subsistence), then my intending O is the mental act of judging that O has that form of being. By thus intending O I am then able to address the question of O’s properties, perhaps by a further exploitation of the perceptual representation that provided the evidence for the existential judgment, or, if O is a mathematical object, by following proofs whose validity partially rests on my recognition of O’s subsistence.  

Now consider objects that we cannot intend in the way described because we do not think of them as either categorically existent or subsistent. Our lack of commitment may be due to either agnosticism on our part or otherwise positive disbelief. Thus I may say (evincing the first attitude): “The golden mountain may or may not exist,” while another, a disbeliever, may say: “The golden mountain does not exist.” Neither of us intends the golden mountain in the same way as someone who has a perceptual representation “as of a golden mountain” and who apprehends this object by (wrongly) making the appropriate existential judgment.

So how do we intend the difficult cases: the golden mountain, the perpetuum mobile, phlogiston, Hamlet? Meinong’s answer, which he himself calls the Assumption View, is exciting and, I think, close to the truth once we reject the realist commitment to objects in terms of which the question is phrased. Under conditions of agnosticism or disbelief (so runs the answer) we intend objects, and thereby gain a degree of cognitive access to them, by fairly explicitly assuming that they have being. In Meinong’s own crisp summing-up:

In order to give a thing some thought, a person “places himself in the situation in which there is such a thing.”

Note that someone’s assuming such an existential proposition carries no implications about belief, provisional or otherwise. Indeed, Meinong thinks that assuming is a sui generis intellectual attitude towards objectives, different in kind from judging in large part because of such an absence of commitment. The example that perhaps most graphically bears this out is the case of assumption in the case of play-acting, drama and fiction:

The intellectual attitude of the child at play is less than judgment, but it is more than representation; which it is to say, it is an attitude of assuming.  

Similarly “fiction is just assumption.” What Meinong means is that in play-acting, drama and fiction a certain kind of pretense takes place. Thus

The child at play “feigns” properties, situations, and so forth with regard to himself and others,
while

the dramatist will inevitably be confronted with the task of “placing” himself in not just one but, by turns, almost all the roles of the drama.\textsuperscript{20}

Assumption, then, is frequently pretense. We often pretend that things are thus-and-so without ever believing, even provisionally, that things are really this way. While such pretense can in a sense be entirely serious (at one point Meinong mentions war-games), in no sense are participants serious in believing the assertions that occur during the pretense—unless, of course, they forget the point of the activity. There is no delusion here, no matter how provisional,\textsuperscript{21} only a different kind of attitude towards assertions: a pretend attitude in which we place ourselves “in the situation in which there is such a thing.”\textsuperscript{22} Only in this way can one go beyond mere representation to a cognitive grasp of the object. That is how we get to see that the golden mountain is indeed a golden mountain, or that Holmes lived on Baker Street.

6. Pretense

Throughout his discussion, Meinong preserves talk of objects. The assuming or pretending he talks about is part of a quasi-psychological story about how we find out about certain kinds of objects. But there is an analogue of Meinong’s story that does without special objects altogether, and in my view represents one of the most exciting new ways of dealing with the problem of our talk of apparently non-existing things. The 1970s saw the (re-)emergence of the idea of pretense or make-believe, where it was put to work dealing with some difficult problems in aesthetics. One such problem was how to make sense of people’s reactions to the events depicted in movies and books: the way we recoil and scream as the Green Slime approaches Charlie in some B-grade horror-movie, for example, while knowing full well that no one can be in any danger since the Green Slime doesn’t exist.\textsuperscript{23}

In Kendall Walton’s influential development of the idea of make-believe, children as well as adults play games of make-believe on the basis of props that mandate that they imagine certain things.\textsuperscript{24} Thus, a children’s game may require its participants to imagine that a certain oddly shaped stump is a bear, that actions done to the stump are acts against the bear, and so on. Another sort of game, indulged in by children as well as adults, involves reading/listening to a story or watching a movie, a game that requires its participants to imagine that certain events really happened (that there really was a famous detective called “Holmes”)

\textsuperscript{20} Ibid., p. 86.
\textsuperscript{21} Ibid., p. 83.
\textsuperscript{22} Ibid., p. 175.
\textsuperscript{23} See K. Walton, “Fearing fictions”.
\textsuperscript{24} See Walton, \textit{Mimesis as Make-Believe} and G. Evans, \textit{The Varieties of Reference}. For an application of the pretense view to various issues in philosophical logic, see M. Crimmins, “Hesperus and Phosphorus: sense, pretense, and reference”.
who lived on Baker Street, for example, or that there really is a green, slimy creature approaching Charlie). In Walton’s terminology, something is true in a game of make-believe when it is thus mandated to be imagined. Some such claims are truths internal to the story (“Holmes knew Watson” and the like) while others involve the game-players themselves. In children’s games, of course, there is often no real distinction. Truths are largely created by what the participants do.

The case of novel-reading or movie-watching is especially interesting from this point of view. Walton holds that participants create truths to the extent that they are readers or watchers. “I am reading about Holmes” is such a truth; “We are all sad about what happened to Anna Karenina” is another. Truths like the latter are created by our reactions to the events described in the novel or depicted on the screen. This provides an answer to the problem of our reaction to Charlie’s plight as the Green Slime approaches. No doubt our screams and the palpitations of our hearts are real enough, but what we experience is not genuine fear for Charlie. It is only make-believe that we fear for Charlie as the Green Slime approaches. While there is a sense in which we are genuinely scared, what shows that we are not in the grip of genuine fear for Charlie is that we don’t act appropriately—we don’t rush towards the screen, for example, or call in the army (unless, of course, we forget that we are watching a movie).

What is important from the point of view of the larger ontological questions at issue is the anti-realism of this theory. When speakers talk about “The Green Slime,” they do so from the perspective of a game of make-believe in which their use of the term “the Green Slime” refers to some creature depicted on the screen they are watching; in reality, of course, there is nothing for it to refer to. Similarly, when I remark that “Holmes is much-admired,” I am speaking from the perspective of a game of make-believe in which other people are also participants, acquainted with records about a man called “Holmes” whose character and deeds occasion admiration in much the same way as we might admire historical persons for the deeds attributed to them in historical records. What I say is true from the perspective of this game.

But surely I don’t simply utter a pretend-truth? Isn’t there a sense in which what I say is in fact genuinely true? This is something that Meinongians, of course, would insist on. And this is one place where pretense theorists may well seem to be in trouble. Many of them, Walton included, are inclined to accept a direct-reference view of names, a view on which the semantic role of a name is simply to stand for its referent. But then no sentence containing the name “Holmes” expresses a genuine proposition, since the name “Holmes” has no referent. It is only make-believe that the name has a referent, and that the sentence expresses a proposition. Walton’s way out is to invoke a distinction between what is semantically expressed by a sentence and what a speaker asserts or conveys through its use. When you speak “about Holmes” from the perspective of the game of make-believe, you also implicitly assert that you have

---

25 See Recanati, Direct Reference: From Language to Thought for a detailed account of the direct reference view of names and demonstratives.
said something that is true from the point of view of the game; you assert that the sentence’s fictional truth-condition holds. And this reflective claim may well be genuinely true. It is in this sense genuinely true to say that “Holmes is a much admired detective.”

But what about the claim that Holmes doesn’t exist, for example, or that he is a fictional detective? Surely these are literal truths, truths that have nothing to do with make-believe or pretense? Pretense theorists like Walton disagree. To see what a pretense theorist might say about such a case, it is worth mentioning some little-known comments by Meinong on negative existential claims like “Phlogiston does not exist.” When articulating his Assumption View, Meinong asks this:

What if someone intended something like “phlogiston”, in order to make this judgment about it: that there isn’t anything of that sort?  

He continues that only the assumption or pretense that there is such a thing can allow us to think about it and hence make this judgment about it. He seems to think, therefore, that an utterance of the negative existential “Phlogiston does not [really] exist” must involve speakers in a shift from pretense to reality, with the sentence being true just when a certain bit of pretense indulged in by the speaker fails to match reality. The latter is presumably the case when the real world, unlike the world of the speaker’s pretense, fails to contain anything deserving the name “phlogiston.”

Modern pretense theorists believe something rather similar. According to Walton, for example, when we say that some objects exist (e.g., London, Denmark) and others not (e.g., Holmes and Hamlet), we are playing a new kind of game of make-believe. It is a game that contains elements not authorised by the works of fiction in question. In particular, it contains special fictional predicates such as “exists,” whose rules of application—that is, the rules that tell us when it is fictionally true in the game to say something of the form “X [doesn’t] exist”—are governed by facts about the real world. Simply put, “X exists” is true in this kind of unofficial game just when genuine or non-pretended attempts to refer with the term “X” (and any other terms that co-refer in the scope of the work) succeed in securing reference to something in the real world. Now consider a specific negative existential like claim (6) above (“Holmes doesn’t exist”). Given what was said earlier, when a speaker utters (6), he or she implicitly asserts that the fictional truth-condition of this sentence is satisfied. To that extent, he or she implicitly asserts that genuine attempts to refer with the name “Holmes” and kindred terms fail to secure reference to anyone in the real world. 

---

26 Meinong, On Assumptions, p. 175.
27 This is not the only way pretense theorists have tackled negative existentials. My preferred account holds that ‘exists’ is a universal predicate which literally holds of everything, and that negative existentials exploit this logic of ‘exists’ to convey the claim that underlying determinants of reference fail to secure reference to anything in the world. (See my “ ‘Disavowal through commitment’ theories of negative existentials”; Walton, “Existence as metaphor” contains a reply to an earlier version of this idea.)
claim is indeed true, and it is this salient truth we focus on when we decide that the claim “Holmes doesn’t exist” is true. Walton tells a similar story about sentences involving the predicate “is a fictional X” (e.g., “Holmes is a fictional detective.” “There are fictional characters in Charles Dickens’ novels whose drollness and ineffectualness makes them seem both pathetic and intensely likeable,” etc.).

Where to go from here?

Earlier I said that Meinong had a kind of pretense theory—his Assumption View. But unlike the modern form of pretense theory developed by Walton and his school, Meinong’s view was combined with an unyielding commitment to an extreme form of realism. Given that Meinong came so close to a “no objects” form of pretense theory, why did Meinong not finally adopt such a theory? The reasons are instructive. In fact, the first edition of On Assumptions saw Meinong contemplating this move. There he had proposed that what is involved in the objectivity of representations is the fact that we merely assume or pretend to ourselves that there exists a represented object. This is what Meinong now has to say in the second edition of On Assumptions:

If objectivity consists in the “having” of an object, and the “had” object must be one that has being, then a representation simply does not have an object in any of the instances of fiction, and at this point one really cannot see why objectivity is still ascribed to the representation, despite its dependency on a fiction. . . . [T]hose proceeded in a more consistent fashion who flatly termed a representation like that of the golden mountain . . . objectless.

Yet even now this attitude strikes me as being completely in conflict with the facts. Just as clearly now as before, the facts tell me instead that when I think of unclouded human happiness or of the perpetual motion machine, my thoughts are directed to “something”, i.e., to an object, just as surely as if it were a matter of the most everyday piece of actuality.28

Meinong here lays down the limits of the use of assumption or pretense. Contrary to a “no objects” view, Meinong now claims that pretense by itself cannot complete the goal of accounting for the full range of phenomena that seemingly involve non-existent objects.

Modern pretense theorists like Walton deny that there is such a limit. They agree with Meinong, contra Russell, that the logical form of sentences like (4)-(7) is indeed that of genuine subject-predicate sentences whose semantics fully accords with the Naive View. What they deny is that more than a pretend-commitment to objects is needed to capture the truth of such sentences, or that more than a pretend-commitment to objects is needed to account for the feeling that our thoughts about Holmes, the perpetual motion machine or phlogiston are directed at actual objects. Like most of us, pretense theorists have little trouble with the commitments apparently required by (1)-(3) (they are mostly moderate realist about things like numbers, properties and propositions), but they think that

---

28 p. 170.
talk of the golden mountain, phlogiston, Holmes, and so on, is of a different order altogether and is not to be understood in realist terms.

Who is right? Here I must express my allegiance to the pretense theory, while agreeing that the jury is still out. Deciding if pretense is enough will require much more research into the nature of pretense and make-believe, and will need to look at a great many different but interconnected kinds of contexts involving our thought and talk about the apparently non-existent. Perhaps in the end only a mixed theory will do, with some, but not all, such contexts succumbing to a pretense analysis. (Kripke holds a view of this kind.) But if the pretense theory is ultimately successful, it will have achieved something rather significant. It will have shown why our talk of the non-existent is so robust (why it can’t, for example, be analysed or explained away in Russellian fashion), without this requiring that there be something non-existent for us to talk about.

**Bibliography**

Crimmins, M.


Evans, G.

1982 *The Varieties of Reference*, Oxford University Press.

Kripke, S.


Kroon, F.


Lackey, D.


Meinong, A.


Ostertag, G., ed.


---

29 The view is presented in his 1976 John Locke Lectures (see also Evans, *The Varieties of Reference*). Kripke thinks that the truth of relational claims like “John admires Holmes” as well as the truth of a claim like “Holmes is a fictional character” require an abstract fictional object, *Holmes*. Ordinary claims about fiction (“Holmes was a detective,” say) can be understood more simply in terms of the pretense that there is a concrete person, Holmes, of whom the Holmes stories are a reliable record. There are a number of other philosophers who laud Walton’s introduction of the notion of pretense, but think that Walton’s attempt to do without any fictional and mythological objects is doomed to failure. See, for example, Amy Thomasson, *Fiction and Metaphysics*. 
Parsons, T.
1980 Nonexistent Objects, Yale University Press.
Recanati, F.
Routley, R.
1980 Exploring Meinong’s Jungle and Beyond, Departmental Monograph # 3, Philosophy Department, Research School of Social Sciences, The Australian National University.
Routley, R. and V. Routley
Russell, B.
Thomasson, A.
1999 Fiction and Metaphysics, Cambridge University Press.
v van Inwagen, P.
Walton, K.
1990 Mimesis as Make-Believe, Harvard University Press.
Zalta, E.
1988 Intensional Logic and the Metaphysics of Intentionality, MIT Press.
Consciousness, Materialism, and Ideology

William S. Robinson

I

I am honored to have the pleasure of presenting a few ideas to the Advanced Reasoning Forum for the first time. In view of the name of this group, I believe it will not be out of order for me to begin by discussing a piece of advanced reasoning—that is to say, a piece of reasoning that was on the forefront of thought when it was put forward about 370 years ago. I should warn you that it is somewhat controversial to say that the argument I am about to discuss was offered by Descartes, for it is not perfectly transparent that he argued in precisely the way I am going to represent him as having argued. Nonetheless, the argument I will ask you to consider is fashioned from clearly Cartesian premises that occur very close together in the Meditations, and reflection on it will take us directly to the heart of the debate between dualism and materialism.

Let us briefly remind ourselves that by the middle of the second of his Meditations, Descartes has convinced himself of two things. One is that he exists; this proposition is safe even if there is a supremely intelligent and powerful Evil Demon who is determined to mislead poor Descartes. The other is that, with respect to the whole of the world of bodies, he is in danger of deception. The Evil Demon might present Descartes with the appearances of bodily things, even if there is nowhere anything that corresponds to those appearances, in the way that Descartes has been accustomed to believe that there are corresponding bodily things.

Having convinced himself of this much, Descartes is in a position to argue as follows.

(D1) I am certain that I exist.
(D2) I am not certain of any bodily thing that it exists.
(D3) I cannot be both certain and uncertain of the (existence of the) same thing at the same time.

Therefore,
(D4) I am not (identical with) any bodily thing.

In view of the fact that the word “certain” occurs in all the premises of this argument, it is somewhat natural to expect to find it in the conclusion. This naturalness may lead you to want to reformulate the conclusion of the argument, for example, to “I am not certain that I am identical with a bodily thing.” I have no objection to anyone’s considering the merits of such revised arguments, but I regard them as less interesting than the one stated. In any case, even if you think

---

The ideas below were originally discussed in a seminar that I presented to the Advanced Reasoning Forum in Humor, Romania, in July, 2000. For the sake of readability, I have written the following as a lecture, but many of the points made were developed cooperatively in the seminar.


some revised argument would be a *better* argument, there can be no objection to considering the merits or demerits of the argument that I have actually stated; and it is this task that I ask you to undertake with me. It will, perhaps, increase the interest of this task if I remind you that Descartes drew the strong conclusion—that he is distinct from any bodily thing—and not merely that he wasn’t certain he was a bodily thing.\(^2\)

The third premise of the foregoing argument (let us call it “the \(D\)-argument”) is perhaps the least transparent sentence in the argument. Let us therefore consider a second version of essentially the same argument, which will make the force of (D3) a little clearer. This “\(D’\)-argument” is stated in terms of a somewhat peculiar property, which I will abbreviate by “\(C\)”. \(C\) is the property that a thing, \(x\), has, just in case \(x\) is known (with certainty) by me to exist.

(D’1) I have property \(C\).

(D’2) No bodily thing has property \(C\).

(D’3) For all \(x\) and \(y\), \(x = y\) if and only if

for all properties \(P\), \(Px\) if and only if \(Py\).

(Leibniz’s Law for identity.)

**Therefore,**

(D’4) I am not identical with any bodily thing.

Well, now, what shall we say? Are we in possession of a good argument for the conclusion that I am not identical with any bodily thing? Few philosophers indeed would be willing to say that either the \(D\)- or the \(D’\)-argument gives us a good argument for its dualistic conclusion. The reason is the standard one: there are plainly invalid arguments that have the same form as the foregoing arguments. We may, for example, imagine a detective who is investigating a theft and who is interrogating a suspect, John Jones. Let us imagine this detective to reason as follows.

(1) John Jones has the property of being certainly known by me to be sitting before me.

(2) The thief (in the case under investigation) does not have the property of being certainly known by me to be sitting before me.

(3) Leibniz’s Law for identity is true.

**Therefore,**

(4) John Jones is not (identical with) the thief.

Here, surely, is a candidate for quick demotion.

Someone might object that this argument concerns only accidental properties, and that Descartes believed he was considering an essential (necessary) property of himself. But necessity makes no difference, as we can see from the following example of a beginner at geometry.

\(^2\) Strictly, Descartes did not draw his conclusion until he thought he was in a position to know that the goodness of God certified the correctness of his reasoning powers. Nonetheless, once he believed he had such assurance, it was the foregoing argument that led him to the claim of distinctness of himself (or, thinking substance) from anything bodily (extended substance).
(1) I am certain that I am thinking of a right angled triangle.
(2) I am not certain that I am thinking of a figure that has the Pythagorean property.\(^3\)
(3) I cannot be certain and uncertain of the same thing at the same time.
Therefore,
(4) The figure of which I am thinking does not have the Pythagorean property.\(^4\)

A little reflection will lead us to a generalization of the difficulty in these arguments. There is nothing very special about “certain”. We will get parallel invalidities with “know” and “believe”, and even with “hope for” and “desire”. The presence of a psychological verb (and certain other contexts) defeats the D- and D’-arguments: We cannot apply Leibniz’s Law (without special provisos) in contexts where psychological verbs are present.\(^5\)

II

It has taken us a few minutes to arrive at the generalized reason for the failure of the D- and D’-arguments, but it was clear almost immediately that there was something wrong with them.\(^6\) The ease of seeing that there is something amiss

\(^3\) The Pythagorean property is the property of having the square of one side equal to the sum of the squares of the remaining sides.

\(^4\) In D’ form:

(1) A right angled triangle has the property of being certainly known by me to be that of which I am thinking.
(2) No figure that has the Pythagorean property has the property of being certainly known by me to be that of which I am thinking.
(3) Leibniz’s Law for identity is true.
Therefore,
(4) The right angled triangle of which I am thinking does not have the Pythagorean property.

\(^5\) Chisholm in Perceiving: A Philosophical Study, tried to use (apparent) failure of Leibniz’s Law as a mark of the mental. However, the problem also arises in contexts containing modality or quotation. Following Quine, Word and Object, we may note that nine is necessarily identical with nine, but nine is not necessarily identical with the number of the Sun’s planets; we cannot validly conclude that nine is not the number of the Sun’s planets. We may further observe that “eerie” has five letters. “The sixth word of the fourth sentence of this note” does not have five letters. We cannot validly conclude that the sixth word of the fourth sentence of this note is not "eerie".—The remainder of this lecture, however, will be concerned only with cases involving psychological verbs.

What do I mean by “without special provisos”? It is open to us to insist as, e.g., Quine does in Word and Object, that we preserve Leibniz’s Law by adopting referential readings. On such a reading, our detective is simply not entitled to premise (2); i.e., regarding the thief, it may well be true that the detective certainly knows that he is sitting before him. (He just doesn’t recognize this truth when the person is described as "the thief" rather than as "John Jones"). But such readings often do not come naturally; hence, it is a "special proviso" that we insist on requiring them.

\(^6\) Several members of the seminar were ready to offer plainly invalid parallels of their own before I had finished stating the D’- version.
ought to make us a little suspicious. Indeed, one plausible reaction would be that Descartes cannot really be supposed to have advanced an argument that can so easily be seen to be invalid. I am inclined to agree, but the argument seems to me so close to what Descartes says that I do not think we can reject it outright. Instead, I suspect that Descartes was tacitly supposing something that made the D-argument look better to him. I am thus led to look for a premise that we might add to the D-argument (or to some slightly revised version of it) that will significantly improve the argument. Such an additional premise should, of course, be one that Descartes himself could reasonably be supposed to have accepted, even if he never made it explicit.

Now there is an idea that comes up from time to time in contemporary discussions of phenomenal consciousness that will provide us with a candidate for Descartes’ missing premise. I would not say that this is a point that is often emphasized, but most philosophers will at least recognize it as something that Descartes’ contemporary descendants are likely to hold. (These descendants are qualia realists, also known as “qualia freaks”.) Not to tantalize you, the idea can be baldly put in this way: There is nothing to qualia (or to phenomena, experiences, raw feels, sensory consciousness, pains, afterimages, etc.) other than the properties that constitute them as the qualia that they are. There is no inner or hidden nature to them; their presented surface is their nature, and their entire nature is to be that presented surface.

Let us approach this key idea a little less baldly but perhaps with more understanding. Consider a plain fact, a fact that is the root of much philosophy, a fact with which Descartes was intimately (some might say excessively) concerned, namely: Things can appear other than they are, and can even appear when they aren’t there at all. What is the logic behind this familiar fact? First, there is a contrast between a thing and the way it appears. (“Appears” may seem a little formal; that is because it is most often replaced by one of its species, “looks”, “sounds”, “tastes”, etc.) An ordinary thing, e.g., a table or a tree, is, in general, a complex object. Even when it appears the way it is, it is understood that a thing can have aspects that do not appear: insides, backsides, microstructures, and relations to other things. And, of course, for any number of reasons, a thing may not really be as it appears to be. Now, in this statement, the way things appear (or, the appearance of a thing) has a certain role, which is very limited. The way a thing appears is just the way that thing presents itself; an appearance has no other job than to be the way a thing appears. Serving that function, so to speak, exhausts its role.

Now, nonassignment of a further role need not imply inability to have a further role. But let us see what it would be like to embrace “thick” appearances, that is, to regard appearances as possessors of aspects other than that of merely being the way a thing appears. The moment we enrich our conception of appearances in this way, we will find it necessary to say that not all of their aspects appear (to their possessors). And the moment we say this, we will be committed to introducing a new level of appearances: we will have to have our
old appearances of things, and then, in addition, the way those old appearances happen to be appearing to their possessors. We will be so committed, because we will have to have some way of enforcing a distinction between the aspects of (our old) appearances that are the way they are appearing, and the aspects of (our old) appearances that are not the way they are appearing.

I hope your suspicions are thoroughly aroused by this talk of appearances of appearances, or the way appearances appear. For one thing, the grammar of things and their appearances, or things and the ways they may appear, is very familiar; but talk of appearances appearing, or the way appearances appear is not. Behind this grammatical observation is a more serious problem. If the new appearances are appearances, that is, appearances in the same sense as the old appearances, there would seem to be no reason why they too may not be “thick”, i.e., no reason why they may not have aspects that are not apparent; and then we will need the way appearances of appearances appear, and so on. While such a stack of appearances is not a vicious regress, the lack of a nonarbitrary place to stop makes it a theoretical monstrosity. The natural response of any theorist will be to cut the regress off at the earliest possible step. Since we cannot get along with no appearances at all, the best we can do is to refuse to allow appearances of appearances. In that case, we must regard appearances as not being possessors of aspects that do not appear. We must say, instead, that there is nothing to an appearance other than its surface; it just is the aspect that some thing appears as. Appearances don’t appear, they are; and their being is just some thing’s appearing.

Although I know of no text in which Descartes says anything like the foregoing, the problem of allowing appearances to be “thick”, that is, to have aspects that do not appear, is so obvious that it is plausible to imagine that Descartes thought of appearances as “thin”, that is, as being exhausted by the property that they present something (some thing) as having. “Thin” appearances also seem to be the outcome of Descartes’ method of doubt. One can doubt that reality is appearing to one as it is, but it is much harder to doubt that the appearances are what they are. Alleged nonapparent aspects of appearances, however, would have to go on the side of realities “behind” the appearances, and would thus fall within the realm of the dubitable. If, as Descartes seems to do, one confines appearances to what is indubitable, then appearances must not have “hidden” aspects, i.e., they must be “thin” in the sense we have been using this term.7

If this point is found agreeable, the next question will be how the thinness of appearances might be of any help with the D-argument we reviewed earlier. Here

---

7 Descartes says in Meditation II, Haldane and Ross (1931), translators:

Finally, I am the same who feels, that is to say, who perceives certain things, as by the organs of sense, since in truth I see light, I hear noise, I feel heat. But it will be said that these phenomena are false and that I am dreaming. Let it be so; still it is at least quite certain that it seems to me that I see light, that I hear noise and that I feel heat. That cannot be false; properly speaking it is what is in me called feeling; and used in this precise sense that is no other thing than thinking.
again, I shall offer a rough and very bald formulation. After that will come some refinement, and then, at the last, an evaluation of the newly-formulated argument.

The rough formulation is this. Let us suppose that Descartes is thinking that what he is certain of is the existence of appearances. Now, appearances are thin: They don’t have aspects that don’t appear. So, if there is some property that is not appearing to Descartes, when he does have an appearance, that property is definitely not a property of the appearance. And if there is some property, $P$, that is never a property in any appearance, then that property is definitely not a property of any appearance, and appearances can be taken to be distinct from anything that has $P$.

Before I set this argument out a little more formally, let us think for a moment about our incompetent detective’s reasoning about the thief. The way this incompetence works is this. The thief is a thing; it is “thick”. It is the kind of thing that can have properties that are “hidden”, i.e., not apparent. Another terminology in which this kind of point is sometimes made derives from Frege. One may say that a thief may be presented under one mode of presentation (e.g., as the person sitting across the table from me) without being presented under another mode of presentation (e.g., as taker of jewels). Our detective’s mistake may be described as an invalid inference from “is not presented under mode of presentation $M$” to “is not $M$”. But, just as appearances do not plausibly appear, so modes of presentation do not plausibly have modes of presentation. Thus, if Descartes were thinking of his argumentation as applying only to modes of presentation, then the detective counterexample would violate a restriction on the universe of discourse. It would thus be irrelevant to Descartes’ argument, and his reasoning might not be incorrect after all. The same point can be made about our geometrical example. Abstract though they may be, geometrical figures (and other mathematical entities) can be presented in more than one way. They are not mere modes of presentation; so, if Descartes is restricting his argument to modes of presentation, the geometrical example is also irrelevant to his case.

Let us now see if we can incorporate these suggestive remarks into a formal statement of a neo-Cartesian argument.

(ND1) I am certain I am having an appearance that is constituted by the property $F$.

(ND2) I am not certain that anything has the property $G$. In particular, $G$ is not a property that is constitutive of any appearance I (now) have.

(ND3) Appearances have no properties other than those that constitute them. Therefore,

(ND4) The appearance I am having does not have the property $G$.

Several comments are in order. “Constituted by” is a new term, and so needs some explication. Let us start with the basic forms of appearance

---

8 Likewise, a figure may be presented under the mode of presentation "right angled triangle" without being presented under the mode of presentation "bearer of the Pythagorean property". A number may be presented as odd without being presented as prime, or as the eighth term of the Fibonacci series, etc.
statements: “Object O appears to subject S as F” and “It appears to S that there is an F object.” In these forms, “F” tells us what kind of appearance S is having. To say that an appearance is constituted by a certain property is just to identify that property as the one that substitutes for “F” in these forms.

On this understanding, the ND-argument has a strong claim to be regarded as valid. This claim should be greeted with initial suspicion, because the psychological predicate “certain” occurs in the first two premises. I have included this term so as to make the connection to Descartes’ text visible, but now we can note that its contribution to the ND-argument is quite different from the contribution that “certain” makes to the D- and D’-arguments. In (ND1), all that “certain” does is to emphasize the epistemic status of an assertion that does the real work in the argument, namely, “I am having an appearance that is constituted by the property F.” Likewise, the argumentative force of (ND2) lies in the claim that I am not having an appearance that is constituted by the property G. To emphasize the point being made here, let us reformulate the argument a little.

(RND1) I am certain I am having an appearance that is constituted by the property F.
(RND2) I am not certain that I am having an appearance that is constituted by the property G.

So,
(RND3) I have an appearance that is constituted by the property F.
(RND4) I do not have an appearance that is constituted by the property G.
(RND5)Appearances have no properties other than those that constitute them.

Therefore,
(RND6) The appearance I am having does not have property G.

Whatever force this argument has remains (or, perhaps, increases) if we lop off the first two premises. These premises somewhat redundantly say that I am warranted in asserting (RND3) and (RND4).

(RND6) is considerably weaker than Descartes’ intended conclusion. However, suppose there is a property G such that

(RND7) I can never have an appearance that is constituted by the property G.

Then we may derive the stronger conclusion:

(RND8) No appearance has the property G.

Descartes’ candidate for a property G that would make (RND7) true is extension. There are problems with this candidate. On the one hand, visual experience, tactile experience, and bodily sensations all have a “spread out” character to them. Sounds also have a locational aspect. If “extension” is understood in such a way as to conflict with these points, then its use simply falsifies the phenomenological facts. And on the other hand, “extension”, as

---

9 This point is further explained and supported in my book in progress, Understanding Phenomenal Consciousness; and in Chapter 6 of M. Velmans' Understanding Consciousness.
Descartes understood it, does not include mass. It is therefore inadequate to capture what contemporaries would understand to be the physical world.

The failure of Descartes’ candidate for a property that could serve as $G$ in (RND7) does not, however, show that there is no good candidate, and in fact there is a better one. None of our appearances appear to us as neurophysiological events (e.g., ratios of neural firing rates, or rises of activations in particular brain parts). Thus, if the RND-argument is accepted, and neural event properties are substituted for “$G$”, we will have an argument that our appearances are not neural events.\footnote{There is nothing in this conclusion that conflicts with the view that neural events cause appearances, and this is, indeed, the view of qualia realists.}

This result is a strong neo-Cartesian result, but still not as strong as Descartes’ own which, let us recall, was that I am not any bodily thing. I hope no one will be disappointed to find out that I am not going to defend Descartes’ putting his conclusion in this way. Despite Descartes’ “modernism”, that is, his break with Scholasticism, he was so habituated to thinking in terms of substances and accidents that he was unable to avoid an unsupported substantialist outlook. However, disavowing Descartes’ substantialism should not be regarded as a reason for rejecting the view that the RND-argument is genuinely neo-Cartesian. There are many passages in which Descartes limits what should be thought to “pertain” to “himself”, by excluding everything of which he can doubt, that is, everything that is not the way he is thinking of something or the way something seems to be appearing to him. To put the matter in an intentionally paradoxical way, Descartes argues repeatedly that

I am nothing but the ways I think and the ways it seems things appear to me.

In contrast, the upshot of accepting the RND-argument is only that appearances are not identical with neural events, or with any physical event of type $G$ such that nothing appears as $G$.

Substantialism about selves is not the only kind of ontological worry one might have about the RND-argument. “Appearance(s)” functions as a noun in this argument or, as one may also say, the argument treats appearances as things (albeit in a very inclusive sense of “thing”). But perhaps appearances are better thought of as events; and perhaps it is not innocent to apply a logic of things to events.\footnote{This objection reflects my understanding of a point that was raised in the Humor seminar by Richard Epstein.} So, it may be that a better understanding of event-logic would bring to the surface some now-hidden reason for doubting the validity of the RND-argument.

I have no decisive reply to this objection, but I am not very worried by it. While I am no specialist in the logic of adverbs, I am under the impression that there is no generally accepted account of how to treat adverbial inferences. This fact does not make me doubt that “Juney barks loudly, therefore, Juney barks” is valid. I suspect that there are countless similar arguments that seem valid, and
that will continue to be regarded as valid whenever a satisfactory account of inferences involving adverbial constructions is developed. The RND-argument is fairly simple, and the key concept, appearance, is well understood. Thus, I suspect that the apparent validity of the RND-argument will survive its recasting into a properly articulated event logic, whenever such a logic is developed.

III

I have tried to explain the RND-argument in such a way that it will recommend itself as an important argument for dualism (or, to use more contemporary language, anti-physicalism) with respect to phenomenal qualities (appearances, qualia, sensory experience, phenomenal consciousness, etc.). But I know that contemporary materialists will not be moved by it. One might suppose that the reason is that they will doubt the new premise,

(RND5) Appearances have no properties other than those that constitute them.

But doubting this premise would not serve the purpose of materialists. Their view is not merely that all things are physical, that is, that all things have at least one physical property, whatever other properties they may have. Their view is that all properties are physical properties. What is necessary in order to maintain this view is to hold that the RND-argument is invalid, and that it is so because

(M) The property $F$ and the property $G$ can be identical properties, even if $F$ is constitutive of an appearance and $G$ is not constitutive of that appearance.

If (M) is true, then the premises of the RND-argument can be true while its conclusion is false.

In this lecture, I cannot survey all of the recent literature produced by materialists and their opponents. The most I can do is to illustrate some of the points I will make by showing how they occur in the work of one materialist author, and I will do this shortly. But the conclusion I have drawn from an extensive involvement with this issue is that, after many layers of surface issues have been resolved or dissolved, the irreducible and irremovable core of the contemporary materialist/anti-materialist debate is one's acceptance or rejection of (M). My further remarks in this lecture will thus concern the prospects for establishing or refuting this key proposition.

I will begin by noting that (M) is certainly not obvious on its face, and seems to cry out for an explanation. "How," we may wonder, "could it happen that $F$ is the way something appears, and $G$ is not the way anything appears, and yet $F$ and $G$ are the very same property?" The matter becomes even more mysterious when we consider what materialists routinely offer as the substituends for "$F$" and "$G". There are a number of proposals here, and we cannot go into the reasons for their differences. But a typical idea is that $F$ may be a color—e.g., red may be a way something is appearing to us—and that very property may be the same property as a neural property, e.g., a set of ratios of activation potential rates in a collection of neurons.
It would seem that if such an identity is possible, we ought to understand how it is possible. But here we come upon one way of expressing the problem summarized in J. Levine's famous phrase, "the explanatory gap". Namely, we have no explanation of how such property identities are possible, and we have no idea of what sort of research program might lead to such an explanation.

It is true that we have models for establishing property identities in some kinds of cases. Thus, for example, it is a result of chemistry that the property of being water is identical with the property of being $\text{H}_2\text{O}$. Perhaps not everyone will agree with this result, on the ground that ice is $\text{H}_2\text{O}$, but not water, or on the ground that a molecule of $\text{H}_2\text{O}$ isolated in interstellar space is not water. I think we should opt for theoretical simplicity and accept what the chemists say in this case, but we need not insist on this resolution. For, suppose we say that the property of being water is identical with the property of being a sizeable collection of $\text{H}_2\text{O}$ molecules within a certain temperature range. Then we will still be recognizing a case of property identity that is not immediately obvious, i.e., one that had to be discovered by scientific methods; and this case can serve as an illustration of how we can sometimes establish property identities.

Cases of property identity of this kind work by construction. That is, chemistry shows us how we can construct a body of water, namely, by assembling molecules of a certain kind. If anyone doubts this, there are many further explanations available to back it up. For example, water is transparent. It is so, because it does not absorb light in the visible frequency range; and this is true, in turn, because the resonant frequency of $\text{H}_2\text{O}$ molecules lies outside the visible frequency range. Again, water dissolves many substances, and does not dissolve others; these facts can be accounted for by the ways in which $\text{H}_2\text{O}$ molecules distribute themselves as a result of the forces exerted on them by molecules of other substances.

Now, the problem of the explanatory gap is that no explanations of this kind are available for understanding how, e.g., red could be "constructed" from neural events (ratios of neural event rates, etc.). Association, correlation, and causation present no problem. You can prick yourself with a pin, which is plainly a physical event, and produce a pain every time. Medical science shows us that pinpricks cause brain events, and that if we use anesthetics to prevent the usual brain events, we forestall the pains. So the correlation of pains and brain events is not in question and, since the pinpricks come first, the causation of pains by physical events seems questionable only on grounds that are common to causal cases in general. But neither correlation nor causation is identity, and the firmness of the correlational and causative claims contributes not one iota to explanation of how the property of being painful could be the same property as a property of a set of neural events.\(^\text{13}\)

---

\(^{12}\) "Materialism and qualia: the explanatory gap."

\(^{13}\) Some philosophers, e.g., W. G. Lycan, *Consciousness*, will insist that the point be put as a claim of identity of having a pain with having a brain event of a certain kind. I will not go into the niceties of proper reformulation here. When these are carried out, however, the explanatory gap remains. See my book in progress, *Understanding Phenomenal Consciousness* for details.
The upshot of these remarks on explanation is that those who assert (M) are asserting a substantive claim that is nonobvious and requires explanation, but for which they have no explanation, nor even a plausible program for discovering the required explanation.

Now, it might be thought that those who deny (M) can turn these ruminations into a disproof of (M). Let us try to imagine how such an Anti-Materialist argument might go. I will put the idea in a general form, but for concreteness we may imagine that $F$ is some color property, while $G$ is some property of a set of neural events.

(AM1) $F$ has the (second level) property of constituting an appearance.
(AM2) $G$ does not have the property of constituting an appearance
(for any appearance whatsoever).
(AM3) Two properties are identical if and only if they have all their second level properties in common.
(Leibniz’s Law for property identity.)

Therefore,

(AM4) $F$ is not identical with $G$.

Now, if our discussion in the first part of this lecture has been successful, you will be immediately suspicious of this argument. In fact, you will see just what a proponent of (M) will be likely to say. Namely, “constituting an appearance” is a psychological context, and we know that we are not entitled to make use of Leibniz’s Law in psychological contexts. The AM-argument is thus invalid and is powerless to overthrow (M).

Actually, this formulation somewhat overstates what the materialist is entitled to. The contexts we can confidently affirm to vitiate applications of Leibniz’s Law are contexts expressing propositional attitudes (e.g., beliefs or desires). What we are sure of is that we can believe or know about a thing under one mode of presentation, without knowing or believing anything about it under a different mode of presentation; and we can desire something under one mode of presentation without desiring it under another mode of presentation. But constituting an appearance is not constituting a propositional attitude. So, it may be that the context in the AM-argument is not one that vitiates applications of Leibniz’s Law. Certainly, qualia realists will think of appearances as realities that do not count as propositional attitudes and that do not present the same kind of difficulty as the arguments that involve certainty, knowledge, belief or desire.

But materialists will not be contradicting themselves or begging the question if they say that the context in the AM-argument is one that may vitiate the application of Leibniz’s Law. They cannot explain how it could do so, but they know that some contexts vitiate Leibniz’s Law, and that some of these are psychological. Thus, they are entitled to say that the AM-argument may use a context that vitiates applications of Leibniz’s Law, and that if it does, it is not a refutation of the key proposition, (M).

Putting together the points of this section leads to the following conclusion.
Materialism has not been refuted; we have found no way in which (M) can be disproved. But the way it which it can be rescued from refutation requires that materialism be empty. Materialists cannot explain why the AM-argument doesn’t work, they can only make the negative point that a certain possibility has not been ruled out. They do not actually have an argument to show that the possibility (that AM is invalid) is a genuine possibility; they have only the right to say that the genuineness of this possibility has not been disproved. They have no explanation as to how (M) could be true (this lacuna is the explanatory gap); the only comfort they may find is the fact that (M) fails to be refuted.

The result that a defensible materialism is empty in the sense just explained is the most important conclusion of this lecture. Rather than just leave it to your consideration, however, I want to make it a little more memorable by giving a concrete illustration of some of the points that have led to it. Then, I want to use our conclusion and our illustration to raise a general, and somewhat disturbing, question about philosophical reasoning.

IV

Just so you won’t think I have merely invented the dialectic of the previous sections, let us quote a little from a well known contemporary materialist, Patricia Churchland.\textsuperscript{14} I will be considering her response to a view that I hope you will remember from our foregoing discussion: the view, namely, that “in the case of consciousness, the appearance is the reality. . . . Feeling the pain is all the reality there is to pain.”\textsuperscript{15} Her response to this view is:

What is troublesome is the idea that all the reality there is to a sensation is available through sheerly having it. How could you possibly know that? I suggest, instead, a rather simple alternative: A sensation of pain is real, but not everything about the nature of pain is revealed in introspection – its neural substrate, for example, is not so revealed.\textsuperscript{16}

This passage introduces the idea that there is something about the “nature of pain”—not merely what is correlated with or causative of it—that is not revealed in introspection (or, in my terminology, is not constitutive of a kind of appearance). It is natural to ask what this hidden aspect of the nature of pain is. “Neural substrate” gives some hint of an answer, but in the following paragraph Churchland makes explicit what she has in mind.

Commonly, science discovers ways of taking us beyond manifest or superficial properties of some phenomenon. Light is refracted in water – that is observable. Light turns out to be electromagnetic radiation, a property not straightforwardly observable.\textsuperscript{17}

The property of being electromagnetic radiation and the property of being refracted in water are here treated as distinct properties. But light is held to be

\textsuperscript{14} P. S. Churchland, “Brainsy: nonneural theories of conscious experience”.
\textsuperscript{15} Ibid., p. 117.
\textsuperscript{16} Ibid., p. 117.
\textsuperscript{17} Ibid., pp. 117–118.
electromagnetic radiation, in which case the property of being light and the property of being electromagnetic radiation are the same property. So, Churchland’s model really is one of discovering identities. The same point holds of the other reductions she mentions as examples, namely, temperature to mean kinetic energy, burning to rapid oxidation, and electricity to movement of electrons.

One might now expect Churchland to tell us what property it is to which sensations, or some particular sensation, e.g., pain, reduces, or how we are to “explain consciousness in neurobiological terms”. But she offers no such account; she says only that attempts in this direction are “worth trying”.

Our remarks about water, H₂O, and the explanatory gap show, however, that there is no reason to think such a project is even worth trying. We certainly expect to be able to find out more and more about the neural causes of pains, and of other sensations. But, as we have seen, this causal or correlational investigation offers no help in explaining how the property of being a pain could be identical to the property of being a set of neural activations of some kind. The net force of Churchland’s remarks is thus no more than an expression of a hope that some future development will change the present situation, and lead us to the ability to explain the property identity that materialism implies.

At this point, it is often suggested that science is known for making progress, often in surprising directions, and that we should, therefore, trust in the progress of science to eventually provide a genuine explanation of a kind we cannot now even imagine. But here, I think, we should take our cue from Hume’s (1748) essay “Of a Particular Providence and Of a Future State”. To summarize very briefly, this essay is about the argument from design. A watch found on a beach would lead us to suppose an unseen watchmaker; how much more, then, should the exquisite organization of nature lead us to suppose an unseen designer of the world. Hume proceeds by conceding the point (at least for a while); but he insists that we are in no way entitled to infer any more perfection in the designer than what we observe in the design. If, therefore, we observe injustice and death in the actual world, the argument from design will fail to license an inference to an afterlife in which perfect justice is achieved. Matching the measure of the designer to the observed measure of perfection in the creation can lead us at most to suppose there is a designer that is subject to imperfections or limitations.

The point of this parable is to persuade you that we are not entitled to presume on the future of science beyond what we have observed in its actual development. And what we actually observe is that, while the term “explanatory gap” is of recent vintage, the problem it labels has been very well understood for over 300 years. Examples of the kinds of reductions Churchland points to have been available for over 200 years. Yet no progress at all has been made in understanding how it could be the case that sensations are nothing but neural events. Moreover, we know that science makes major advances in unexpected

---

18 Ibid., p. 118.
19 Ibid., p. 118.
directions. Physicists of 1850 could hardly have imagined what quantum mechanics or relativity theory would lead their successors to believe. So, even if we appeal to the idea that we will eventually make progress in science, we still have no reason to think that that progress will be in the direction Churchland, or any living materialist, imagines. It is equally compatible with belief in the progress of science to suppose that science will find a way to incorporate anti-materialism (denial of (M)) within an explanatorily satisfying worldview.

V

There are many forms of materialism that we have not considered, and cannot consider in a single lecture. Thus, it is a large promissory note that I issue when I say that the essential points we have made, the centrality of (M), its inexplicability, along with its irrefutability, all hold up when variants of materialism are examined. What I want to do in closing is to consider what we should think if we agree in supposing that this promissory note can indeed be redeemed.

One might think that, in view of the points I have mentioned, materialism would be regarded with either agnosticism or doubt. As I have portrayed the matter, all that materialism has going in its favor is the fact that anti-materialists cannot disprove (M). So, one might be excused for suspending judgment about (M). Or, one might reasonably take the following attitude. The AM-argument is a strong, albeit not demonstrative, consideration against (M), and we cannot even understand how (M) could be true; thus, while agreeing that the falsity of (M) is not certain, we should think that (M) is probably false, and should try to find an overall worldview that incorporates both our sciences and the falsity of (M).

In actual fact, however, materialism is affirmed by the great majority of the best-known philosophers of our time. It is affirmed with passion and without qualification by any suspicion of doubt. The philosophers who affirm it are not only well known, they are, in fact, both knowledgeable and acute. They have expended a prodigious amount of genuine intellectual inventiveness and hard work in defense of the materialist point of view.

For all that, however, it seems clear to me that the degree of attachment to materialism is far in excess of what is warranted by the argumentation that is offered in support of it. The fact that the more defensible materialisms are the emptier materialisms goes unappreciated, and the attachment to the view remains strong even in those who believe that human beings will never understand how materialism could be true. In short, in today’s philosophical milieu, materialism has the status of an ideology—a view whose acceptance has become disconnected from the force of the reasoning that can be adduced in its favor.

I indicated above that I found this conclusion disturbing. The reason is, I hope, obvious. It cannot be any comfort to think that the best philosophers of one’s day are in the grip of an ideology. The suspicion that this is so leads to the unpleasant thought that ideology is well-nigh unavoidable, at least in philosophy.

---

Mere doggedness is not a virtue; but, perhaps, its avoidance is not a virtue to which we can honestly aspire. From the point of view of critical thinking, we must ask whether there is always a limitation, that is, whether there are always fundamental views that resist being brought under the scrutiny of critical thought. Since I am an optimistic person, I harbor the hope that these pessimistic suggestions are overdrawn. But I do think that the place of ideology in philosophical enterprises (especially those involving many contributors) and the question of avoidability of ideology are matters that deserve our serious consideration.

**Bibliography**

Chisholm, R.
1957 *Perceiving: A Philosophical Study*, Cornell University Press.

Churchland, P. S.

Descartes, R.

Hume, D.
1748 *An Inquiry Into Human Understanding*.

Levine, J.

Lycan, W. G.

McGinn, C.

Quine, W. V. O.
1960 *Word and Object*, Technology Press of MIT.

Robinson, W. S.
forthcoming *Understanding Phenomenal Consciousness*.

Velmins, M.
Modals and Quantifiers

Mircea Dumitru

This paper is part of a work in progress in which I explore the prospects of a view about possibilia and matters modal for which I want to argue. The view in question is known as Modal Actualism. Roughly, the view consists of the following two theses, the first of which is primarily concerned with the semantics of modal idioms, and the second of which is motivated by a metaphysical (ontological) stance with respect to objects.

Thesis I (Modalism) The natural language idioms of “necessarily” and “possibly” are primitive with respect to elaborate semantical renderings via, e.g., “possible worlds.”

Thesis II (Actualism) Everything that exists is actual.

In this first part of my research I want to set the stage for giving a direct argument in favor of modal actualism. To this purpose I discuss some more foundational issues. I am particularly interested in disposing of major alternative views about possibilia and the semantics of modal talk. This is why I address certain ramifications of the main issue that are brought about by influential views which do not share with modal actualism either the semantic doctrine of the priority of modal idioms or the ontological stance about actual objects. Toward the end of this paper I will give my own sketch of the modal actualist solution that I will work out in more detail in another paper.

The challenge that a modal actualist has to meet is serious. In order to make sense of the modal language in a logical setting, i.e., to interpret various modal systems and to provide clear criteria for validity of modal arguments, the possible-worlds semantics for a modal language seems to be a solution which imposes itself naturally. But then the question arises of how to make sense of the possible-worlds language unless we are ready to take at face value quantification over possible worlds, and consequently to accommodate into our ontology those alleged entities that our quantifiers range over, namely, “possible worlds.”

If the modal actualist meets the challenge, then one is supposed to seek a strategy for justifying an affirmative answer to the following question: Could we have both the possible-worlds semantics which works so well with respect to validity of modal arguments, and more generally with matters metalogical for modal systems, and have, nevertheless, a way out from an ontology in which dubious entities like “possible worlds” are endorsed?

To be sure, to get the benefit of an ontology free of possible worlds one needn’t be a modal actualist. One can very well be a quantificational actualist

1 I would like to express my gratitude to Richard L. Epstein and Kit Fine for careful comments and useful suggestions on an earlier version of this work.

© 2001 Mircea Dumitru.
who can take possible worlds as a proxy for what in reality are but sets of sentences or propositions, but who, nevertheless, in contradistinction to a *modal* actualist will find quantifiers over possible worlds as giving in a natural way the meaning of modal operators.

For certain reasons to be addressed later, the quantificational approach doesn’t seem to cope with the meaning of modals, for it introduces entities which the modal language doesn’t seem to require in order to be able to make sense of it. So the quantification-over-possible-worlds solution, if taken as giving the meaning of modal operators, has the undesirable feature of being ontologically radical.

The main issue to be addressed here is that of a double reduction. First, there is a reduction of a modal language to a corresponding possible-worlds language according to certain translation schemata. The main reason for this reduction is that by doing possible-worlds semantics for modal systems and arguments we can see the logical mechanisms at stake in metalogical issues and validity. Second, when we come to the issue of the “real” meaning of sentences of modal language, what has been taken before as the class to be reduced, namely, the class of sentences of modal language, reverse roles with the previously reductive class, that is, the class of sentences of possible-worlds language. Thus, the problem of a reverse translation occurs as part of a general strategy through which we, as modal actualists, seek to show that the rendering for technical reasons of modal operators as quantifiers over possible worlds does not impute any meaning to sentences of a modal language, or at least any meaning which could be accounted for only by an ontologically radical solution. In a nutshell, how can a modal actualist use quantificational possibilist discourse in order to explain the validity of modal arguments, and at the same time be an anti-realist with respect to the existence of possible worlds?

To begin with, let’s see what’s the main rationale for asking for a reduction of the modal discourse to a nonmodal one in which modal operators are assimilated to (restricted) quantifiers over possible worlds?

A clear motivation for such a maneuver has been given by David Lewis.

The standards of validity for modal reasoning have long been unclear; they become clear only when we provide a semantic analysis of modal logic by reference to possible worlds and to possible things therein. Thus insofar as we understand modal reasoning at all, we understand it as disguised reasoning about possible beings.2

What is then to be done for achieving such standards? In Lewis’ “Anselm and actuality” we find a strategy which is supposed to bring about the desired standards of clarity for modal discourse, the main principle of which is

Given any statement about what may be conceived to be the case, we translate it into a statement about what is the case.3

So what is required is to replace modal language and reasoning with reasoning in

---

2 “Anselm and actuality,” p. 10.
3 Ibid., p. 11–12.
an extensional first-order language. This boils down to an argument in favor of substitution of talk about the ways in which things could have been otherwise than they actually are for talk about things which could have been in many ways otherwise than they actually are. Taken at its face value this idiomatic rendering of a modal locution is an existential quantification over certain entities described as "ways things could have been." And since the existential quantifier is the logical device of expressing existence, it seems to follow that to make sense of that idiomatic modal rendering of possible nonactual properties one has to countenance the existence of those entities which can be called "ways things could have been" or for short "possible worlds."

This is tantamount further to a requirement to the effect that for any sentence $s_m$ of the modal language there is a corresponding sentence $s_w$ of the language of quantifying over possible worlds (possible-worlds language, for short) such that there is a relation $R$ that holds between them and its holding between the two sentences guarantee the same logical behavior both for $s_m$ and for $s_w$. What counts here for logical behavior is, roughly speaking, the truth-value of $s_m$ and $s_w$, respectively, in corresponding interpretations, and the validity of arguments made up from sentences of modal language, and possible-worlds language, respectively. Thus, what $R$ is bound to guarantee is that whenever $s_m$ is true under one intended interpretation, $s_w$ is also true under a corresponding interpretation, and that an argument $A_m$ carried through within modal language is valid if and only if a corresponding argument $A_w$ in the language of possible worlds which is obtained by some translation schemata from $A_m$ is also valid.

If one asks further what type of relation is required to do this job, we find that the most plausible candidate for $R$ is a synonymy relation between modal operators and quantifiers over possible worlds. For if $s_m$ and $s_w$ mean the same, and provided that $s_w$ and the whole formal apparatus of possible-worlds semantics show the promise of more intelligibility than $s_m$ and the modal reasoning show, then it makes sense to use $s_w$ and that apparatus as a guide for the logical behavior of $s_m$ and of any modal argument of which that $s_m$ is a part thereof.

So, it seems that if we want clear standards for the validity of modal arguments then we need to buy into the view which construes modal operators as quantifiers over possible worlds. And this view comprises two essential claims: that a modal sentence and its corresponding rendering through restricted quantification over possible worlds are synonymous, and that the meaning of a modal sentence is given through its synonymous possible-worlds sentence.

The problem with this proposal is that it seems to force us to acknowledge in our ontology a special sort of entity, "possible worlds," which the quantifiers corresponding to modal operators are supposed to range over. Since in the modal discourse no appeal to possible worlds is apparent, it follows that this solution is ontologically radical.

The need for those special entities is notably emphasized by David Lewis in

---

4 Compare G. Forbes, The Metaphysics of Modality, pp. 70–76.
On the Plurality of Worlds by drawing our attention to the fact that it is not the mathematics of modalities, i.e., the metalogical investigation of various modal systems and of their relations, that requires the help of possible-worlds discourse, but the metaphysics of modalities, i.e., the application of modal concepts and systems to matters philosophical.\(^5\)

It seems then that one can’t have the benefits of the metalogical investigations of modal systems and the application of its results and concepts to metaphysical issues without paying a certain ontological price.

The real challenge now for somebody who can’t be persuaded to be a full-blown realist regarding possibilia is to keep to a minimum those commitments, and to see whether it is possible to stick to the mathematical policy toward possible worlds, as described by Lewis, even when we are doing metaphysics. Can we afford to use, for philosophical purposes, any sets of entities whatsoever, which for heuristic guidance “may be regarded as” possible worlds? More to this point, can one develop a full-blown anti-realist position concerning possible worlds, while preserving all the technical advantages of the quantifier approach of the modal operators?\(^6\)

But except for qualms about ontological paucity, why not endorse “possible worlds” as genuine entities that our quantifiers range over? Moreover, what reasons could there be for resisting the overall strategy of reducing modal language to possible worlds language, regardless of the view one might endorse about what a possible world could be?

I think that in the literature one can find two main reasons for resisting this approach, and accordingly, for insisting upon the primacy of modal language with respect to possible-worlds language. First, one may have worries concerning the

---

\(^5\) Lewis in On the Plurality of Worlds, p. 17 says

For that job [the metalogical semantical analysis of modal logic], we need no possible worlds. We need sets of entities which, for heuristic guidance, ‘may be regarded as’ possible worlds, but which in turn may be anything you please. We are doing mathematics, not metaphysics. Where we need possible worlds, rather, is in applying the results of these metalogical investigations. Metalogical results, by themselves, answer no questions about the logic of modality.

And again, acknowledging the seminal results about the relations between modal systems obtained through the technical apparatus of modal frames and models based on different frames, Lewis stresses the limits of the technical approach when coping with metaphysical questions, ibid., p. 19:

But in truth the metalogical results, just by themselves, cast no light at all. If the modal operators can be correctly interpreted as quantifiers over the indices of some or other frame, restricted by the relation of that frame, then we have found out where to look for illumination about controversial axioms. If not, not.

---

\(^6\) In The Metaphysics of Modality, p. 76, Forbes acknowledges that

The most interesting philosophical question about the semantics of modal logic is whether it is possible to develop an anti-realist view that is consistent with our intuition of naturalness in the quantifier treatment of the modal operators, and which can deal with the thought that the invalidity of [a modal argument] is somehow explained by the invalidity of [its corresponding rendering in sentences belonging to the possible-worlds language].
ontological status of that to which the reduction is carried over, i.e., of possible worlds. Second, as we’ll see soon, there is a more subtle motive for being suspicious about the reduction as such. For as Kit Fine shows the main problem with the attempt to reduce modal discourse to possible worlds language, and then further, from a quantificational actualist perspective, to reduce possible worlds to other entities which might appear more respectable from an ontological point of view, such as sets of propositions, and essences, is not mainly the nature of the things by which the reduction is effected, but the question-begging character of the reduction itself.7

First, why not to be a modal realist?

In general, you may not want to be a realist about worlds and possible individuals because within various sorts of realism advocated nowadays you don’t find conclusive grounds for endorsing the view that modal locutions are synonymous with possible-worlds sentences, and because to make sense of possible-worlds language one has to introduce into one’s own ontology entities which don’t show up when perfectly meaningful modal locutions are used as part of the natural language.8

So, although you may prefer to do possible-worlds semantics for technical purposes, when you come to foundational philosophical questions, such as the meaning of modals, and the status of modal operators with respect to quantifiers over possible worlds, you may want to insist that ultimately quantifiers over possible worlds are not to be taken as primitive with respect to modal operators and that modal operators should be given meaning via another approach.

To this general attitude concerning the reduction of modal discourse, one should add, furthermore, the internal drawbacks and difficulties that the few distinct main types of realism about possible worlds face.

Then, the answers to the question “Why not be a realist concerning possible worlds?” will split according to what kind of realism one has in mind.

For the purposes of our discussion a useful distinction concerning realism about possible worlds is that between an absolute realist and a reductive realist.9

The best example, and maybe the single example of an absolute realist is D. Lewis. His realism could be spelled out in principle as being made from two basic ingredients, namely, quantificationalism and possibilism. Together they work to give credit to the main tenet of absolute realism that possible worlds are real entities which cannot be further analyzed. Roughly speaking, Lewis’ absolute realism concerning possible worlds springs from his taking seriously existential quantifiers in ordinary language, i.e., to understand them as expressing the existence of a special sort of entities: “ways things could have been.” Although he provides no decisive arguments in favor of his contention, Lewis

7 “Plantinga on the reduction of possibilist discourse.”
8 Thus, you don’t want to endorse what has been called the ontological radical character of such a solution which introduces for explanatory reasons new entities which don’t show up in the ontology which is required to make sense of the language to be reduced.
9 Compare Forbes, The Metaphysics of Modality, p. 75.
believes that "possible worlds" can do for philosophers the same kind of job as
sets do for mathematicians. Thus, if possible worlds are useful for coping more
easily and elegantly with a host of philosophical problems which otherwise would
be dealt with in a clumsier way, then by way of an inference to the best
explanation, and by analogy with the philosophical attitude of the mathematicians
toward sets, one can legitimately entertain the existence of possible worlds.

There are in the main four tenets which Lewis defends: (i) possible worlds
exist and they are as real as the actual world in which we exist; (ii) other possible
worlds are things of the same sort as the actual world (characterized as "I and all
my surroundings"); (iii) the indexical analysis of the adjective "actual" is the
correct analysis; (iv) possible worlds cannot be reduced to something more
basic. 10

For the purposes of our discussion, the main divide between Lewis’ absolute
realism and other brands of reductive realism can be located at the level of the
second and the fourth theses on Lewis’ agenda. Thus, what makes Lewis’ realism
a very implausible doctrine which flies in the face of our common intuitions is his
systematic defense of the idea of homogeneity of all worlds which constitute the
logical space. That means that according to Lewis other possible worlds differ not
in kind from the actual world but only in what is going on at them. And our world
is actual not because it has attached to it a special ontological status which makes
it different in kind from other possible worlds but because it is the world which
we are parts of. Actuality is not some ontological feature which opposes
absolutely one particular world to any other possible world. Rather, it is a relation
which obtains between the denizens of any world and that particular world in
which they inhabit.

Lewis himself is ready to argue in favor of his implausible position on
grounds of what might be called parsimony. Actually, from several other critical
encounters which he deals with in “Possible worlds” this contention concerning
the unparsimony of his ontology seems to be the only one he grants. 11 However,

10 See R. C. Stalnaker, “Possible worlds.”
11 In particular, one piece of criticism directed against Lewis’ view regarding the homogeneity
of worlds is due to Stalnaker. The way in which that criticism seems to me to miss its target is
paradigmatic for the internal consistency and resourcefulness of Lewis’ ontology of possibilia.

Thus, Stalnaker claims that if a possible world is, according to Lewis, a way things might
have been, then the actual world should be construed as the way things are, and not as "I and all
my surroundings." Then, since “the ways things are” expresses a property of a state of the world,
and not the world itself, it follows that what makes the sentence “The actual world is the way
things are” true is not the construal of is as the is of identity, but rather as the is of an attribution
of a property, viz. the property of being “the ways things are,” to a thing, viz. the world. And then,
if we grant that properties could exist uninstantiated and that “the way the world is” is such a
property, “then the way the world is could exist even if a world that is that way did not.”
(Stalnaker, “Possible worlds”, p. 228.)

Hence, Stalnaker’s argument concludes, while one can accept that there are many ways that
things could have been (which is Lewis’ claim (i)), one isn’t forced to make the further step to
accepting that “there exists anything else that is like the actual world.” (ibid.) i. e., we are not
supposed to accept Lewis’ claim (ii) by which the homogeneity of worlds is endorsed.
by way of drawing a distinction between qualitative and quantitative parsimony, Lewis claims that only the former is to be praised because “it keeps down the number of fundamentally different kinds of entity.”\textsuperscript{12} Whereas his realism, being only quantitatively unparsimonious, is not at fault so damagingly with respect to that stringent and legitimate demand of parsimony. “You believe,” says Lewis, “in your actual world already. I ask you to believe in more things of that kind, not in things of some new kind.”\textsuperscript{13}

The various brands of reductive realism could be interpreted as dispensing in the main with the thesis of homogeneity of worlds and with the unanalyzability demand. For reductive realists, possible worlds are made out of other entities, which are held to be in better ontological standing than worlds themselves.

Some prominent options which have been worked out are that worlds are maximal consistent sets of propositions, or maximal states of affairs, or maximal possibilities.\textsuperscript{14} These reductive positions are not at variance with the main tenet of realism. Within any such doctrine it is perfectly meaningful to say that there are worlds which ultimately means, after analyzing worlds into their basic and genuine constituents, that there are sets of propositions or of states of affairs, and so on, which meet some specific constraints. What is different now is that worlds are not “respectable entities in their own right.”\textsuperscript{15} They are identified by way of reduction with abstract entities, and hence they are taken to be abstract things themselves.

It is worth noticing, though, that the absolute-reductive distinction doesn’t coincide with the possibilism-actualism distinction. Although many prominent positions which are opposed to the former distinction are also opposed to the latter, not all reductive realists are actualists. Thus, a reductive realist who identifies worlds with possibilities is not an actualist. He or she is a reductive-possibilist.\textsuperscript{16}

These reductive realist positions, however, face their own problems. That a realist such as Lewis points out that they can be faulted on scores such as circularity and incorrectness, probably shows that the divide between actualism

---

So the upshot of Stalnaker’s argument seems to be that it can block Lewis’ inference from “Things might be otherwise than they are” to “There are many ways things could have been besides the way they actually are.” Particularly if “ways things could have been” is an entity which the existential quantifier ranges over and which is of the same kind as the actual world.

But Lewis’ answer could be typically articulated as follows: I agree with the premises of your argument. And in particular with your point of view that “the way things are” is a property of the actual world and not the world itself. However, what I demand for the sake of doing systematic metaphysics more easily, and as an outcome of an inference to the best explanation, is to grant me that the world is a huge concrete system (“I and all my surroundings”) and that there are many other such concrete objects as the one we happen to be a part thereof.

\textsuperscript{12} Lewis, “Possible worlds”, p. 185.

\textsuperscript{13} Ibid.


\textsuperscript{15} Lewis, “Possible worlds”, p. 183.

\textsuperscript{16} Compare Forbes, The Metaphysics of Modality, p. 75.
and possibilism means more than that both Lewis and the actualists he criticizes share the same quantificationalist perspective on modal operators. Lewis’ reproach against any position which interprets modal idioms as quantifiers over so-called “possible worlds” (ersatz possible worlds, which in reality would be some sort of respectable linguistic entities) is in the main that “the theory would be either circular or incorrect, according as we explain consistency [of maximal sets of sentences or propositions] in modal terms or in deductive (or purely model-theoretical) terms.”

Likewise, as R. C. Stalnaker points out in relation to Robert M. Adams’ reduction of possible worlds to propositions (world-stories), by taking sets of propositions as primitives we are left with three undefined notions, namely, proposition, possibility, and contradictoriness, which require an account not yet told in a satisfactory way. Whereas, within a possible-worlds based framework, a very familiar analysis of intensions in terms of possible worlds is available; e.g., a proposition is a function from a set of possible worlds to truth values. And the advantage of this analysis seems to be acknowledged by Adams himself who defends, however, an ersatz program defined in terms of intensional concepts, and not in terms of nonactual possibles.

In addition to the ontological problems with possible worlds there is an epistemological nominalist-actualist objection against both absolute and reductive realism. If knowledge of properties of an object – the argument goes – requires an experience of the object or of its effects which should be confined within the powers of human sensibility, then only objects which are actual and concrete could possibly be known, since only they, or their effects, are confined within the range of our sensibility. But then neither the absolute realist nor the reductive realist makes possible our knowledge of that property which either of them, respectively, attributes to an object or world by the expression “◊A.” For the former holds that “◊A” attributes the property expressed by “A” to a non-actual thing, whereas the latter claims that the property is attributed to an abstract object. Thus, the upshot of this epistemological nominalist-actualist objection is that we cannot have knowledge of whether or not “◊A” is true.

We come now to a crucial objection whose target is the very possibility of reducing modal language to possible-worlds language, even if a possible world is taken to be an ersatz entity. The objection is due to Kit Fine, and its target is A. Plantinga’s reduction of possible worlds and nonactual individuals to sets of propositions (or in Plantinga’s preferred version, states of affairs), and to individual essences, respectively.

17 Lewis, “Possible worlds”, p. 183.
19 Adams admits that “there is a not unfamiliar trade-off here, between nonactual possibles and intensions (such as propositions).” And he adds optimistically and quite naturally for someone who favors a program based on an intensional concept that “given either [i.e., nonactual possibles or intensions], we may be able to construct the object or to do the work that was supposed to be done by talking about the other. (“Theories of actuality”, p. 207.)
20 Compare Forbes, The Metaphysics of Modality, p. 79.
The point that Fine makes is that the reduction of possible worlds to propositions and individual essences via a quantificationalist approach over ersatz worlds is bound to be circular. Let's take a quick glimpse at this. The challenge for the modal actualist is to translate back into his/her own modal language the possibilist language, and in particular three of its key locutions: (i) the predicate for the actual world, (ii) the atomic predications used by the possibilist such as statements of identity between possible individuals, ordinary world-relative predications, and the special world-relative predication “x exists in w,” and (iii) the possibilist quantifier over worlds and possible individuals.\(^{21}\)

Plantinga’s reduction proceeds in the following way: Possible worlds are identified with world-propositions, i.e., propositions true in one world alone; possible individuals are identified with individual essences, i.e., properties true of a single possible individual in each possible world; and then properties of possible worlds and individuals are taken to be corresponding properties of world-propositions and essences.

The circularity of this reduction becomes apparent when one comes to realize that in order to specify an individual essence for each possible individual which is to be reduced to that unexemplified individual essence, one has to presuppose the existence of that nonexistent individual whose reduction is supposed to be accomplished through that program. The difficulty cannot be removed unless one can specify the individual essence for each possible individual in actualist terms only. But one can’t have such a specification in actualist terms, unless a version of the following principle of the identity of indiscernibles is also forthcoming: “For any two distinct possible individuals there is an actualist formula A(x), true of the one in each world in which it exists, but not true of the other in each world in which it exists.”\(^{22}\) Fine calls this the “Discernibility Doctrine.” But the problem now is that this Discernibility Doctrine is far from being unobjectionable. Even were the Discernibility Doctrine unobjectionable, we are ill-advised methodologically if we make the reduction depend on such a metaphysical assumption. “For one thing, as Fine puts it, the reduction becomes more vulnerable to criticism. But also, more importantly, we feel that the reduction of possibilist discourse should not depend upon any particular modal views, that if the reduction is possible then that possibility should be written into the very nature of the discourse itself.”\(^{23}\) Thus, if the Discernibility Doctrine is rejected then the specification of individual essences in actualist terms only is rendered impossible. And what this boils down to is as Fine says,

> It will be impossible, *even in principle*, to specify an individual essence for each possible individual without referring to some merely possible individuals. The circularity of the reduction is then apparent; an adequate supply of surrogates for the merely possible individuals requires that we already

---

\(^{21}\) Compare Fine, “Plantinga on the reduction of possibilist discourse”, pp. 146–147.

\(^{22}\) Ibid., p. 148.

\(^{23}\) Ibid., p. 149.
presuppose some of those individuals.\textsuperscript{24}

The relevance of Fine’s point for the main issue of this paper is apparent. Part of an affirmative answer to the question of whether modal operators should be interpreted as quantifiers over possible worlds is the feasibility of a reduction of modal sentences to possible-worlds sentences, regardless of the view which that reductionist has about the nature of possible worlds. However, in view of the question-begging character of the reduction itself, that is, of the fallacious reducibility of merely possible individuals to actual entities that are in better metaphysical standing, the prospects of equating modals with quantifiers are poor.

Let’s sum up:

Reduction of modal language to possible-worlds language which goes along with the interpretation of modal operators as quantifiers over possible worlds is objectionable on three main grounds: (i) the ontologically radical character of the solution; (ii) the dubious nature of the entities referred to as “possible worlds” to which the quantifiers of the possible-worlds language are relativized; (iii) the question-begging character of Plantinga’s reduction of the possibilist discourse to the quantificational actualist discourse in which the quantifiers are supposed to range over ersatz possible worlds and ersatz possible individuals.

Now, let’s move more toward the constructive part. I want to investigate the prospects and ramifications of modal actualism, and against this background I want to develop a view which is anti-realist concerning the existence of possible worlds but which is free to use possible-worlds semantics.

The challenge now is to show that possible worlds sentences \textit{do not mean}, as it were, what they say or what they seem to imply in order to make sense of them. Simply put the problem lies here: the meaning, and, more urgently, the truth of possible-worlds sentences seem to require the existence of some entities which might be called “possible worlds.” To avoid all the problems discussed before we may want: (i) to adopt a metaphysical view which dispenses with the claim that \textit{there are} possible worlds; and (ii) to work on a semantic view which separates the meaning from the truth-conditions, i.e., more specifically the meaning of modal sentences from the truth-conditions of possible-worlds sentences, whatever else the relation between the modal language and the possible-worlds language might be. Is it possible, then, and if so, how to give a truth-definition for possible-worlds sentences, which is required by the exercise of possible-worlds semantics for modal language on the one hand, and to eschew a committal metaphysical view with respect to the existence of possible worlds on the other hand? What would be, then, the anti-realist stance in the ongoing discussion?

There is, first, a very natural intuition that possible-worlds semantics captures very well our intuitions concerning what is valid and what is invalid in

\textsuperscript{24} Ibid.
modal logic. So, the anti-realist is supposed to give an account of the fact that possible-worlds semantics is successful even though there are no such things as "possible worlds." Second, in order that the systematic correlation between the semantics and our modal intuitions not be something miraculous the anti-realist has to accept that a relation of synonymy holds between modal language and possible-worlds language. Nevertheless, the modal actualist who is an anti-realist about possible worlds will shun away from saying that what gives meaning to modal sentences are their truth-conditions when spelled out in terms of possible-worlds via the apparatus of possible-worlds semantics. What he/she must deny is that the meaning of modal sentences is given through their corresponding possible-worlds sentences according to the translation schemata. Hence, he/she must deny the asymmetry part of the synonymy relation as it is endorsed by a realist, and he/she must emphasize that a possible-worlds sentence gets its meaning by its reverse translation into a corresponding sentence of the modal language. Therefore, since "◊A" is the sentence which imputes the meaning of "(∃w) P(w)" , the latter has to have the meaning of the former, and since the former doesn't assert the existence of worlds, neither does the latter.25

To be sure, this approach faces some objections:

(i) The validity problem. Since what modal sentences mean is not what possible-worlds sentences say, another more direct account for the validity of modal arguments is needed, apart from the account in terms of possible worlds, because this latter account doesn't give the meaning of modal sentences and arguments but is only a heuristic device for the purposes of an easier understanding of them. One way out that can be tried is a proof-theoretic account for validity of modal arguments. However, the prospects for such a maneuver do not seem prima facie encouraging. For it seems that a counterpart of the proof-theoretic approach (Gentzen-style) to the meanings of truth-functional sentential connectives via elimination and introduction rules does not work in the modal case. For in the modal case, an argument can be put forward to the effect that the meanings of modal connectives cannot be fixed "from within"—through the elimination and introduction rules of the modal systems for those modal connectives—but only "from outside," through the means of model theory clauses for those connectives. Thus, the sketch for such an argument is the following: Given two arbitrary modal connectives \( c_m \) and \( c'_m \), it is not the case in general that if both \( c_m \) and \( c'_m \) are governed by the same elimination-rule (E-rule) and introduction-rule (I-rule), any sentence that can be inferred by that E-rule from a sentence having \( c_m \) as its main connective is logically equivalent to the sentence that can be inferred by the same E-rule from the corresponding sentence having \( c'_m \) as its main connective. Likewise, the same holds if instead we apply the I-rule to two sentences to get new sentences having \( c_m \) and \( c'_m \) respectively as their main connectives. Simply put, even though the E-rule and the I-rule for both \( c_m \) and \( c'_m \) are the same, the semantics of those two connectives could differ. This question is still open.

(ii) *The reverse translation problem.* The proposal made by the anti-realist works to the extent that any possible-worlds sentence corresponds to a modal sentence through certain translation schemata. For if we want every possible-worlds sentence to have a definite truth-value on an appropriate interpretation—and surely we want that for the purposes of doing possible-worlds semantics—then every possible-worlds sentence should have a meaning. But as modal actualists we want to say that the meaning of a possible-worlds sentence is given by its corresponding modal sentence, and consequently for every possible-worlds sentence there should be a modal sentence which gives to the former its meaning. There is, though, the well-known phenomenon of the expressive weakness of the modal language which needs an account from the anti-realist perspective: there are possible-worlds sentences which can be evaluated on interpretations but which do not seem to be the translation of any modal sentence.

Against this background, the overall structure of my argument for anti-realism with respect to possible worlds embodies, roughly, Michael Dummett’s anti-realist reductive strategy as developed in his paper “Realism.” In its broad structure it goes like this:

(a) Modal language sentences and possible-worlds sentences are synonymous.

(b) The modal discourse is not genuinely factual. It is not a discourse about facts. It is about quasi-facts and quasi-objects (possible states of affairs or merely possible objects), whose possible obtaining and existence, respectively, depend on the actual obtaining and existence of states of affairs and objects, respectively, or it is about ‘abnormal’ facts or things (necessary facts, which obtain no matter what, or necessary objects, which do not come into existence, and whose existence is unpreventable or unperishable). But then in order to be consistent with this non-factualness assumption about the modal discourse, one does not want to give a realist semantic account for its meaningfulness. So, to comply with the constraints of giving an anti-realist account of the meaning of the modal language there are a couple of options that one can consider. One either wants to give an intuitionistic account and divorce the meaning of the modal discourse from the truth-conditions of that discourse—for a truth-condition based theory of meaning invites easily a realist construal of the semantics of the discourse—, or one can insist that if the meaning is to be accounted for in terms of the truth-conditions for the sentences of the modal language, then the construal of those truth-conditions should not be factual. And what I mean by this latter constraint is that when one comes to giving the truth-conditions for sentences governed by modal operators, the modal actualist will state them in terms of ersatz possible objects and states of affairs which are made from actual objects and states of affairs. In other words, in speaking about possible states of affairs or possible objects we, as modal actualists, are speaking modally only about actual states of affairs or actual objects. And since we do not want that the meaning of the modal discourse be given by the truth-conditions of the possible-worlds discourse, in the
reduction that is sought here, ultimately, the meaning of any quantification-over-
possible-worlds sentence will be given by a corresponding modal sentence
through reverse-translation.

Now, if one opts for the intuitionistic approach toward the meaning of the
modal discourse, then what we should work out is an anti-realist non-truth-
conditional semantics, along roughly the following (Dummettian) lines:

(1) The meaning of a sentence is not necessarily given by its truth-
conditions.

(2) The notion of reference (semantic values) of the component parts
of a sentence in a given language does not play a crucial role in the
account of how the truth-conditions for that sentence get determined,
and we do not necessarily have to identify the semantic values of
different parts of a sentence in order to grasp its truth-conditions and
thereby give it a meaning.\(^{26}\)

(3) On a given anti-realist semantics, to grasp the meaning of a
sentence and to know its truth-conditions require that that sentence
belong to a language which as a given class is reducible to another
language (the reductive class).

The meaning of any sentence in the reductive class should be
grasped in advance, and we may need to fix the reference of terms
which occur in those sentences of the reductive class in order to figure
out their truth-conditions and to grasp their meanings. But typically
we don’t need that for the understanding of the meanings and the
truth-conditions of the sentences of the given class (the class to be
reduced). For, basically, the grasp of what a sentence in the class to be
reduced means and of its truth-conditions is made possible by the
grasp of the reductive relation itself and by our prior knowledge of the
meanings of the sentences which belong to the reductive class. Thus,
one does not need a truth-conditional approach for the sentences in the
given class in order to be able grasp their meanings.

(4) The feasibility of this program depends on the existence of
translation schemata which give the conditions for reducing the given
class to a reductive class. Their understanding gives us the possibility
of grasping in a non-truth-conditional way the meanings of the
sentences which belong in the given class. As Dummett’s claim reads:

\(^{26}\) As Dummett spells out this sort of anti-realism in “Realism”, pp. 67–68:

Realism is abandoned, not because a truth-conditional account of the meanings of
statements is impossible, nor, necessarily, because there is any reason to repudiate
the principle of bivalence as applied to them, but because the notion of reference no
longer plays any role in the account of their meanings. ... We do not need to invoke
the notion of reference ...[for a term] ... in order to explain how a sentence
containing such a term is determined as true or false: the determination of the truth-
value of the sentence does not proceed via the identification of an object as the
referent of the term.
This translation is proposed, not merely as preserving truth-values, but as part of an account of the meanings of statements of the given class: it is integral to the reductionist thesis that it is by an implicit grasp of the scheme of translation that we understand those statements.\textsuperscript{27}

(c) However, I do not want to leave open the possibility for a realist construal for the possible-worlds language. For, after all, even if one insists that the meaning of any quantification-over-possible-worlds sentence will be given by a corresponding modal sentence through reverse-translation, there is a coherent way of construing existential quantification over possible worlds which will certainly leave room for countenancing the existence of certain genuine entities that for better or worse are called “possible worlds.”

Moreover, since we really need the possible-worlds semantics for doing the metalogic of modal logic and for having clear standards of valid modal arguments, I cannot see how we can avoid giving the usual model-theory account for intensional languages. But this approach tells us that in order for the modal sentence “$\forall A$” to be true on a given interpretation for the language of that sentence there should exist a “possible world” at which “A” holds. One option which is available here is to say that the existential quantifier which occurs in the metalanguage in which the semantic evaluation clause is given does not have its normal (standard) meaning. In other words, that the English phrase “there exists” is ambiguous, and that its occurrence in the above clause does not carry the usual implication of the existence of things to which the quantifier phrase is relativized in its application. But I don’t think that this solution will do. For one thing, I don’t see any ambiguity in the usual phrase “there exists.” And then, in the usual first order semantic clause for the existential quantifier, the required sense of the English phrase which occurs in the metalanguage should be the standard sense which conveys the existence of some objects which satisfy the given condition on a given interpretation.

Another solution would be to say that the quantifier-phrases in possible-worlds discourse are given a substitutional-instance Fregean interpretation. To form a substitution-instance of a quantified possible-worlds sentence we need a name for that possible world over which the quantifier is supposed to range: “$(\exists w) P(w)$” will typically have as a substitution-instance “$P(w_f)$”\textsuperscript{28}. The individual constant “$w_f$” will stand for a “possible world.” But this latter entity will have nothing metaphysically dubious about it, and it will be perfectly innocuous by actualist standards, for it is but the “reference” of a convenient way of speaking about a complete set of sentences (concrete actual inscriptions) which describe a re-arrangement of the actual world or a way of looking modally at actual states of affairs and existing objects. The name of a “possible world” is a

\textsuperscript{27} Dummett, “Realism”, p. 66.

\textsuperscript{28} If in an actual language there are not enough names for every object in the domain, Frege’s style approach helps us to handle the problem: We construct an interpretation not for the given language but for an expanded language obtained by enriching the original language with at least as many individual constants as the number of objects in the domain of our interpretation such that for each object in the domain there is an individual constant which refers to it.
shorthand for this package of sentences. So, the domain to which our quantifiers
are relativized in an interpretation of the possible-worlds language is a set of
inscriptions as sentence-types which purportedly designate actual objects and
states of affairs which comprise them. Hence, everything which the quantifiers
range over is something actual, after all.

(d) In particular, for the issue of the relation between modal discourse and
possible-worlds discourse, the anti-realist semantics sketched before has the
following bearing:

(1) For technical purposes such as the validity of modal arguments I
shall take the modal language as the class to be reduced and the
possible-worlds language as the reductive class.

(2) For metaphysical issues, however, the reverse relation is to be
sought. The possible-worlds language is the class to be reduced, and
the modal language is the one which gives the meanings to possible-
worlds sentences.

(3) A true substitution-instance of "(\exists w)P(w)" whose meaning is
given by a reverse-translation into "◊A" requires a singular term which
refers rigidly to that entity that is called the "possible world" at which
"P" holds.

(4) Since "(\exists w)P(w)" stands in a reductive relation to "◊A", with
respect to its meaning, the grasp of the meaning of "(\exists w)P(w)" only
requires the grasp of the reductive relation of "(\exists w)P(w)" to "◊A",
and of the meaning of "◊A" itself. It does not require the existence of
an object called "a possible world" which, unless it exists and satisfies
the open sentence "P(w)", "(\exists w)P(w)" could not be rendered
meaningful and true. Thus, it is not a strong requirement whose
fulfillment is indispensable that the variable "w" which is bound by
"(\exists w)" should be assigned a possible world, which therefore must
exist, in order that "(\exists w)P(w)" be meaningful and true. Of course, I
don't claim that this anti-realist position proves the metaphysical claim
that there are no possible worlds. However, there is at least no need
for us to assume their existence in order to make sense of the language
of possible worlds, and thus to be entitled to make use of it when we
do possible-worlds semantics.

Bibliography

Dummett, Michael

Fine, Kit
1985 Plantinga on the reduction of probabilist discourse. In Alvin Plantinga,
J. E. Tomberlin and P. van Inwagen, eds., D. Reidel Publishing Company,
pp. 145-186.
Forbes, G.

Humberstone, L.

Lewis, David
1986 On the plurality of worlds, Basil Blackwell.

Loux, M. ed.
1979 The Possible and the Actual, Cornell University Press.

Plantinga, A.

Prior, A. N., and K. Fine
1977 Worlds, Times, and Selves, University of Massachusetts Press.

Stalnaker, R. C.
The Problem of Unwanted Epistemic Necessities

Célia Teixeira

In this paper I will discuss some of the descriptivist arguments against Kripke’s modal arguments. Of all Kripke’s arguments against descriptivism, the modal ones are the most contested, in the sense that they are arguably weak or not decisive. One way to block modal arguments is by way of rigidifying the descriptions that give the sense of the names. I will show that although this strategy solves the problems associated with the modal arguments, it brings others. The aim of this paper is to prove once again the weakness of the descriptivist theories of names.

1. Anti-descriptivist Modal Arguments

We can formulate the classic descriptivist theory as follows:

(D) For every name and for every use of that name by a competent speaker in a given context, there is a singular definite description such that:
   i. The speaker associates that singular definite description with the name in the context of use.
   ii. The sense of the name is the sense of the singular definite description that the speaker associates with the name in the context of use.
   iii. The reference of the name is determined by that description, as being that object and just that object that the definite description denotes in the context of use.

If, for instance, a competent English speaker associates with the name “Aristotle” the singular definite description “the disciple of Plato,” then the sense of the name is the sense of the description. That is, the singular terms “Aristotle” and “the disciple of Plato” have the same content, they are synonymous. And the name designates that same object that the description denotes, namely, Aristotle. An important corollary of (D), from which we can formulate Kripke’s arguments, is as follows:

(C) If the sense of the name NN is the sense of the definite description The F, then NN and The F are synonymous. Therefore, sentences of the form “NN is The F” are analytic truths, in the sense that they are obtained from a logical truth by the substitution of synonymous by synonymous — NN and The F.

1I would like to thank to all the ARF members, especially Desidério Murcho and Richard L. Epstein. I would also like to thank Paulo Ruas and especially João Branquinho for extensive and helpful written comments on an earlier version of this paper.

© 2001 Célia Teixeira.

the name “Aristotle” in the language of the speaker who associates the name with the description, then, by (C), the following sentence isn’t just true, but analytically true:

(1) Aristotle (if exists) is the disciple of Plato

This is because the terms “Aristotle” and “the disciple of Plato” are, by (C), synonymous terms, and then (1) is an instance of a logical truth of the type \( a = a \).

Two of the best-known modal arguments, which I will discuss, are the so-called argument of unwanted necessities and argument of lost rigidity.²

The argument of unwanted necessities can be formulated as follows. If names are simply abbreviated definite descriptions, then the following sentences have the same meaning:

(1) Aristotle (if exists) is the disciple of Plato
(2) Aristotle (if exists) is Aristotle

This is because, according to (C), “Aristotle” and “the disciple of Plato” are synonymous. (2) is an analytical truth, and by (C), (1) is also an analytical truth. If every analytical truth is a necessary truth, then (1) and (2) are necessary truths. Yet (1) is not a necessary truth but a contingent one: Aristotle is the disciple of Plato, but he could have not been such. Therefore, (D) is false. We have an unwanted necessity.

The argument of lost rigidity is as follows:

*Premise 1* Names are rigid designators
*Premise 2* Descriptions are not rigid designators
*Conclusion* Names and descriptions have different semantic values, and therefore (D) is false.

Kripke set out the idea that names are rigid designators.³ A name is a rigid designator if it designates the same object in every possible world in which that object exists. This is a discovery made by looking to our modal intuitions about how names work. If, for instance, we say that Aristotle is the disciple of Plato, we are in fact referring to Aristotle. But if we say that Aristotle is not the disciple of Plato, to whom are we referring? To Aristotle, of course. “Aristotle” designates Aristotle in all the possible worlds where Aristotle exists, even in those cases where all we know about Aristotle is wrong. But a definite description just refers to those objects that satisfy its predicate. It is for this reason that (2) is a necessary truth and (1) is only contingently true.

An immediate consequence of the thesis of rigid designation is that if names are rigid designators, then they don’t have a descriptive content. Therefore (D) is false.

---

² Devitt & Sterelny, 1999: 51-4
³ Kripke, 1972.
2. Actualised Descriptivism

The descriptivist reply to the modal arguments is also of modal nature. It consists in rigidifying the descriptions that give the sense of the names, by way of blocking the modal anti-descriptivist arguments.

We rigidify a description using the modal operator of actuality. The definite descriptions that result from this process of indexing their predicates to the actual world we call “actualised descriptions.” Actualised descriptions, like names, designate the same object in all possible worlds in which that object exists. With the operator of actuality we can evaluate the term that it applies to in relation to counterfactual situations or possible worlds, in the sense that the semantic value of the term in the actual world is taken for a possible world. For instance, when we say that all we know about Aristotle could be wrong, we are still speaking about the object that the name “Aristotle” designates in the actual world, but in relation to a possible world. The same happens in the case of the actualised descriptions. For instance, the descriptions “the disciple of Plato” and “the actual disciple of Plato,” both denote the same object that satisfies the predicate “is the disciple of Plato” in the actual world, that is, Aristotle. Although the simple description may denote somebody else in relation to a possible world, the actualised description still denotes Aristotle, if it denotes Aristotle in the actual world.

Like the argument of lost rigidity, the argument of unwanted necessities is also blocked with this strategy. That is, the replacement of a name in an expression by the actualised description it abbreviates blocks any difference in the modal status of the expression that results from this process. That is, sentences like the following are both necessary truths:

(2) Aristotle (if exists) is Aristotle
(3) Aristotle (if exists) is the actual disciple of Plato

Therefore, we don’t have unwanted necessities anymore.

3. The Problem of Unwanted Epistemic Necessities

Although actualised descriptivism survives the modal arguments, it is not a satisfactory theory. Although there are modal replies to actualised descriptivism, based in the obstinate rigidity of names versus the persistent rigidity of descriptions, the one I present is an epistemic one.

Although it is true that we could not discover (2) to be false (it’s an instance of a logical truth), (3) is epistemically contingent, since we could discover it to be false. That is, we could come to discover that the description “the actual disciple of Plato” does not denote Aristotle after all. Thus, in the case of (3), but not (2), (3) could be false (in which case it would be a metaphysical necessary falsity). The status of (3) as a metaphysical necessity is compatible with its status of epistemic contingency. But the status of (2) as a logical necessity is not compatible with its status of epistemic contingency. Therefore, names and the

---

4 See J. Branquinho, “Sobrevive o Descritivismo Actualizado aos Argumentos Modais?”
actualised descriptions they abbreviate are not synonymous. So actualised descriptivism is false. But if the actualised description "the actual disciple of Plato" denotes Aristotle, (3) is a necessary truth. But we could be wrong about Aristotle, and in that case, the description denotes another object. Then there exists an asymmetry between (2) and (3): (3) is epistemically contingent but not (2). Thus, (2) and (3) are not synonymous and we have an unwanted epistemic necessity. And if they are not synonymous, they don’t have the same meaning. Then we have a difference in semantic value between names and the actualised descriptions they abbreviated. Thus, descriptivism is, once again, false.

We can put the argument of unwanted epistemic necessities this way:

Premise 1  Sentences of the form, “NN is NN” express epistemic necessities.

Premise 2  The substitution of NN by the actualised description that the name abbreviates makes the sentence that results from this process epistemically contingent.

Premise 3  The semantic value of a sentence depends on its epistemic modal value.

Conclusion  Names and actualised descriptions have different semantic values. Therefore, actualised descriptivism is false.

Therefore, although actualised descriptivism evades the classical modal arguments against classical descriptivism, it still has the problem of unwanted epistemic necessities.

Bibliography

Branquinho, J.

Devitt, M. & K. Sterelny

Kripke, S.
1972  Naming and Necessity, Basil Blackwell.

Stanley, J.
The New Theory of Reference

Problems and solutions concerning proper names

Sorin Costreie

The problems are solved, not by giving new information, but by arranging what we have already known.

Wittgenstein

Introduction

Since Naming and Necessity was published, most philosophers are inclined to see it as a decisive refutation of the so-called “Frege-Russell’s account” on proper names and, consequently, to reject any such Fregean view which maintains that proper names have senses. In opposition to this line, our chief goal is to show that, in spite of this impressive assault, Frege’s view is still tenable and may offer us interesting suggestions with regard to the name-bearer relation. Even though Saul Kripke’s work is highly valuable and very important for the field, he misses the target. He rejects conclusively not Frege’s conception, but a view that maintains that proper names are concealed definite descriptions; we shall call this view canonical. Kripke refers to this position as the “Frege-Russell” theory and characterizes it as comprising two main assumptions: that names are in fact disguised definite descriptions (or clusters of such descriptions) and that the reference is determined via such descriptions.

In order to show that Frege’s view on proper names is not the actual target of Kripke’s attacks, we should make clear at least three points: (i) what Frege’s conception of names is, (ii) what position is attacked by Kripke, and, generally, by The Theory of Direct Reference, and (iii) that Frege’s view and the canonical attacked position are different and separate accounts of proper names. To each point one of the following sections corresponds.

Besides trying to shed light on such a controversial point in contemporary philosophy of language, the paper may be also seen as an attempt to provide a

1 I would like to express my thanks to Derek Brown and Richard Epstein for careful comments and useful suggestions on earlier versions of my paper. This work is a revised version of the talk given at the Advanced Reasoning Forum meeting, Bucharest, Romania, 2000. I am indebted to all the participants involved in the discussions concerning proper names.

2 Nathan Salmon in the Introduction to Harnish, Basic Topics in the Philosophy of Language, p. 114, named this “The Orthodox Theory.”

3 In the course of presentation I shall also touch on some other points of view concerning proper names belonging to authors like John Stuart Mill, Bertrand Russell, Keith Donnellan and others, but the core of the analysis will remain the Frege-Kripke debate. At the end of the paper I shall sketch my own view with regard to Fregean senses, showing that we can draw an illuminating distinction among these.

© 2001 Sorin Costreie.

satisfactory characterization of the relation between proper names and their bearers. Such a problem should be divided in two subsequent questions. Firstly, *Do proper names have semantic content or are they only linguistic labels for the designated objects?* Secondly, *Provided that proper names have “senses,” what are these and how can we express them?* Thus, we shall try to provide an adequate answer to the question: *Does a name refer to its bearer directly or indirectly?* Consequently, after investigating the two main competing positions concerning whether or not a name denotes an object via a sense, we intend to offer a new proposal concerning the contemporary positions in this debate, namely, that proper names are *rigid* but not *direct* designators.

**Mill vs. Frege**

Concerning whether or not proper names have (cognitive) content, we can choose between two main alternatives. One option is to say that there is a direct connection between proper names and their designations, whereas the second is to deny this and to endorse the opposite view, namely that the connection between a name and its bearer is mediated by something. On the former account, originated by John Stuart Mill, proper names have no other semantic function than being a linguistic label, simply attached to its bearer. On the latter approach, mainly due to Gottlob Frege, names have senses, and their chief role is exactly to secure the relation between names and objects.

For John Stuart Mill, *proper names*, such as “Dartmouth,” denote objects without connoting any properties.\(^4\) The Millian terms “connotation” and “denotation” correspond with what in contemporary discussions in philosophy of language is called sense and reference. One important consequence of Mill’s view is that propositions of identity, which contain only non-connotative names such as “Mark Twain is Samuel Langhorne Clemens,” do not have semantic content. This means that they are not “informative,” in the sense that someone might learn something new upon reading it. Further, in a similar manner, “descriptive” means that this information could be captured by a definite description.

---

\(^4\) Mill, *A System of Logic*, p. 20:

Proper names are not connotative; they denote the individuals who are called by them; but they do not indicate or imply any attributes as belonging to those individuals. When we name a child by the name Paul or a dog by the name Caesar, these names are simply marks used to enable those individuals to be made subjects of discourse. It may be said, indeed, that we must have had some reason for giving them those names rather than any others, and this is true, but the name, once given, is independent of the reason. A man may have been named John because that was the name of his father; a town may have been named Dartmouth because it is situated at the mouth of the Dart. But it is not part of the signification of the word John that the father of the person so called bore the same name, nor even of the word Dartmouth to be situated at the mouth of the Dart. If sand should choke up the mouth of the river or an earthquake change its course and remove it to a distance from the town, the name of the town would not necessarily be changed. . . . Proper names are attached to the objects themselves and are not dependent on the continuance of any attribute of the object.
But this leads Mill’s theory into serious troubles, namely in what is often called Frege’s Puzzle (FP). The puzzle goes as follows. Let us consider two sentences:

(1) Mark Twain is Mark Twain.
(2) Mark Twain is Samuel Langhorne Clemens.

Although they are formally different (having distinct linguistic forms), they should express by Mill’s account the same thing. But, as Frege pointed out, it is not the same to say or to think either of the two propositions. The identity of reference doesn’t imply identity of meaning. They express different identities, and this is so because they have different semantic contents. The simple fact that such identity sentences like (2) are informative constitutes a strong objection against all (Millian) conceptions that endorse the idea of an unmediated relation between names and objects. There is a clear difference between sentence like:

(3) John thinks that Mark Twain is Mark Twain.
(4) John thinks that Mark Twain is Samuel Langhorne Clemens.

Such identity sentences have different cognitive contents, in the sense that they express different propositions or thoughts. Verbs like “to think,” “to believe,” “to know,” “to consider” and such, are essential ingredients of what is called “propositional attitude contexts,” where we cannot substitute salve veritate a name by another name. We cannot replace in a sentence like (3) or (4) the name “Mark Twain” with “Samuel Langhorne Clemens,” or vice versa, in the virtue of an identity like (2), because the replacement will change the meaning of the whole sentence. For example, the following inference is not truth preserving:

(5) John thinks that Samuel Langhorne Clemens had a pretty funny middle name.
(2) Mark Twain is Samuel Langhorne Clemens.
Therefore
(6) John thinks that Mark Twain had a pretty funny middle name.

There is a difference in meaning between sentences like (5) and (6). Sentence (5) is true and perfectly meaningful, whereas (6) is meaningless since Mark Twain doesn’t have any middle name. Thus, we are not entitled in such contexts to make such substitutions in virtue of a sentence like (2), due to a difference in meaning.

In “Über Sinn und Bedeutung” Frege argues that our understanding of a proper name cannot consist in just knowing its reference, as in Mill’s account. Frege’s proposal is that besides possessing a reference, a proper name has a sense, the sense being a “mode of presentation” (“Art des Gegebenseins”) of the reference. The proposition (thought) expressed by a sentence is determined by the senses of the words composing it, in the sense that the meaning of the whole is a function of the meaning of its components. Thus (1) and (2) are different because they express distinct propositional senses. That is because “Mark Twain” and “Samuel Langhorne Clemens” express different senses, senses which are different
"modes of presentation" of the same reference.

If sense is not reference, as in Mill's case, then what is sense? To this Frege does not provide a very clear answer. He writes that the sense of an expression is "the mode of presentation of what is designated by this expression," but he does not offer any clear elaboration regarding the nature of these modes, leaving the place for various interpretations. Frege does, however, say that the sense of an expression determines the identity of its reference, but not vice versa. For instance, though "Mark Twain" and "Samuel Clemens" designate the same person (they have the same reference), the two singular terms have distinct senses. Maybe because for Frege the problem of what sense is seems to be a clear and unproblematic point, he didn't offer us an explicit account of the issue. He says that "the sense of a proper name is grasped by everybody who is sufficiently familiar with the language or totality of designations to which it belongs." However, a sort of hint may be found in following footnote Frege gave:

(A) In the case of an actual proper name such as 'Aristotle' opinions as to the sense may differ. It might, for instance, be taken to be the following: the pupil of Plato and teacher of Alexander the Great. Anybody who does this will attach another sense to the sentence "Aristotle was born in Stagira" than will a man who takes as the sense of the name: the teacher of Alexander the Great who was born in Stagira. So long as the reference remains the same, such variations of sense may be tolerated, although they are to be avoided in the theoretical structure of a demonstrative science and ought not to occur in a perfect language.

This footnote was used intensively to show that we are entitled to understand the sense of a proper name as a definite description attached to that name. With regard to the footnote, two aspects should be stressed: (*) It is not clear what each individual using a proper name in a natural language understands as being the sense of the name; consequently, descriptions will vary from person to person, from context to context, (**) Due to pragmatic reasons ("so long as the reference remains the same"), such practices in the case of natural languages may be tolerated. "Tolerated" means here that in such determined cases we may understand the sense as being a definite description. But from that it is a long way to claim further with Kripke that:

(B) The sense of a proper name is (determined by) a definite description.

---

5 Frege, "On sense and reference", p. 143.
6 Ibid., p. 159.
7 Michael Beane in Frege: Making sense, p. 172, seems to agree with this interpretation:
[I]t might be suggested that Frege's use of this example was not so much to illustrate a 'description theory' of proper names, as merely to highlight how far short ordinary language falls of the ideal logical language that Frege was primarily concerned to develop (...) The point of the footnote is to make clear that in the case of an ordinary proper name, there is typically no unique definite description that supplies the sense of the name. Only in an ideal language can the demand for uniqueness be satisfied.
8 Cases when the context is determined, namely when we know somehow the speaker and his or her intentions.
Kripke says explicitly in the beginning of *Naming and Necessity*:

Frege and Russell both thought, and seemed to arrive at these conclusions independently of each other, that Mill was wrong in a very strong sense: really a proper name, properly used, simply was a definite description abbreviated or disguised. Frege specifically said that such a description gave the sense of the name.\(^9\)

The quoted passage is supplemented by a note:

Strictly speaking, of course, Russell says that the names don’t abbreviate descriptions and don’t have any sense; but then he also says that, just because the things that we call ‘names’ do abbreviate descriptions, they’re not really names. . . . Though we won’t put things the way Russell does, we could describe Russell as saying that names, as they are ordinarily called, do have sense. They have sense in a strong way, namely, we should be able to give a definite description such that the referent of the name, by definition is the object satisfying the description.\(^10\)

But although Kripke agrees that he somehow misses the target with regard to Russell’s view, he didn’t say anything similar with regard to Frege’s view, where in fact he totally misses the target. (B) is in fact “Russell’s Name Claim,”\(^11\) an idea that supplements his (well known) Theory of Descriptions. It may be worth noting that the two are different and separate issues, and should be treated as such. This theory of descriptions is more a theory of meaning,\(^12\) and shows us how we should deal and analyze expressions like “The present King of France,” whereas the “Name Claim,” works in the line of a theory of reference, stating that what we call proper names are in fact concealed definite descriptions, and that they should be treated as such. Thus, for example, consider:

(7) Mark Twain is the author of *The Adventures of Tom Sawyer*.

From this true sentence we can also conclude that the following identity is true:

(8) Mark Twain = the author of *The Adventures of Tom Sawyer*.

But even if (8) expresses a true proposition, it is a long way to maintain that (B) is true as well. Frege keeps saying, as in the case of (what are) numbers, that sense of a proper name is something objective and unique. “Objective” means here “that might be grasped and shared by many different people.” This means that although we can attach occasionally different definite descriptions to a proper name, the sense of that name is not entirely revealed by a particular description,

\(^9\) p. 159.
\(^10\) Ibid., footnote 4.
\(^12\) David Kaplan in “What is Russell’s theory of description?” says:

Russell’s article ‘On Denoting’ is not about a theory of descriptions comparable to Frege-Carnap or Frege-Strawson. Russell’s article is about logical form, and is in the tradition of those philosophers who have warned us of the dangers of confusing the grammatical form of a sentence in ordinary language with its logical form.
and moreover, it cannot be replaced by any means with this description. To attach something to a thing doesn’t imply that you are entitled to replace the thing with this “something.”

Harold Noonan in a new book on Frege interprets Frege’s note in a similar way. I shall quote the entire passage, because we believe that it marks a step further in understanding Frege’s view of proper names:

This way of identifying a sense - as a way of thinking of something as the satisfier of a certain condition - fits well with Frege’s own infrequent specifications. Thus, in a footnote in “On Sense and Reference”, in which he is illustrating the imperfections of ordinary language, he writes:

[Frege’s footnote (A) above]

In this passage Frege specifies by description the different conditions corresponding to the different senses associated with the name ‘Aristotle’ by different users. However, as has been stressed by Dummett and Evans there is no need to assume that every way of thinking of an object must be via some descriptive condition, and there is not the slightest reason to think that Frege thought otherwise (as sometimes suggested, see Perry). It is obvious that if anything can be thought of descriptively some things must be thought of non-descriptively if we are to have any reason to believe that any of our descriptions are uniquely satisfied . . . .

Thus, despite the fact that when he attempts to specify a sense Frege invariably does so via a descriptively identified condition, we must not suppose that he thought that sense must be descriptive, and in so far as modern critics assume this they are attacking a straw man.13

It may be important to point out the particular significance of understanding Fregean senses as ways of thinking of different objects as being designated by the proper names. Noonan continues:

The easiest approach to the notion of sense (suggested and developed by Evans) is to think of it as a way of thinking of something, a way of thinking of something as something. Thus I can think of the Evening Star as the Evening Star, or as the Morning Star, or as the planet Venus, or as the heavenly body most often referred to by philosophers writing about Frege. All these are different ways of thinking of one and the same object. In each case it is a matter of thinking of the object as the unique one satisfying a certain condition.14

“Condition” seems to suggest here the existence of a definite description that needs to be fulfilled. If so, it seems that the assumption (B) again comes into the picture.

As we have already mentioned, Russell treats ordinary proper names as disguised or truncated descriptions. His theory of descriptions provides interesting solutions to the problems raised by the Millian theory of proper names. According to this “description theory” a name abbreviates the definite description

---

14 Ibid., p. 178.
that the speakers associate with the proper name, e.g., “Shakespeare” might abbreviate the definite description: “the greatest English playwright.” Thus, it is easy to understand why the above mentioned sentences (1) and (2) have different meanings: “Mark Twain” might be associated with the description “the author of the book *The Adventures of Tom Sawyer*” and Samuel Clemens with “the American writer born in Florida, Missouri, on November 30, 1835.”

Russell’s thesis that proper names are ordinarily used as disguised definite descriptions provides an answer to the question whether or not proper names have meanings. A name designates a certain object because this object is denoted by the definite description associated with the name; “Shakespeare” designates Shakespeare because “the greatest English playwright” denotes him.\(^{15}\)

Keith Donnellan holds the opposing view that ordinary proper names do not refer mediated by a sense or a definite description.\(^{16}\) He extends the thesis even to what is probably the most common use of definite descriptions, the so-called referential, in contrast with the attributive use of definite descriptions. Donnellan consider the following example: A detective discovers a mutilated corpse that he identifies as Smith. Though he does not know who the murderer is, he says “Smith’s murderer is insane.” (a) Later, when the killer is identified as Jones and testifies in court, Smith’s wife, who hears the testimony, says, “Smith’s murderer is insane.” (b). On Donnellan’s account, in the first circumstance (a) the description is being used attributively, while in the second case (b) it is being used referentially. Donnellan shows that in our daily life the most common way of using a description is generally referential, namely to pick out an object. On the other hand, Donnellan is pointing out that this usage of the descriptions could not be accommodated in Russell’s theory, which is designated to provide an account only for the attributive role of the descriptions.

**Kripke and Direct Reference Theory**

Donnellan’s idea is fully developed by a whole group of philosophers whose views are usually referred to by the term: “The Theory of Direct Reference.” They hold that proper names (and by extension natural kinds as well . . . ) are direct rigid designators. “Direct” means here that the relation between a proper name and its bearer is unmediated, whereas “rigid” captures the idea that the

\(^{15}\) But along with solutions, arise new difficulties and problems. One difficulty of this ‘classical’ Russelian theory of descriptions is that speakers can differ in the description they associate with a name. This led some contemporary philosophers, especially John Searle, to modify the initial theory. In this new version of the theory the name is not tied to one description but to a ‘cluster’ of definite descriptions. A name can designate its bearer despite the failure of some descriptions in the cluster to denote that object. It designates the object most of the descriptions in the cluster denote. On the other hand, as Stephen Neale pointed out in “Descriptions”: “[I]t is true that Russell paid little attention to the distinction between the linguistic meaning of a sentence type and the proposition expressed by a particular dated utterance of that sentence type.” This is exactly why Peter Strawson in “On referring” criticizes Russell’s theory because the same sentence (“The present king of France is bald”) might be true or false, depending on the circumstances of its uttering.

\(^{16}\) See Donnellan, “Reference and definite descriptions".
name-reference connection is secured by the fact that a name cannot fail to designate its reference. This Millian picture is supplemented by an additional claim that the name-bearer relation is ensured by the existence of a socio-historical chain which ties the name to its reference in the virtue of an initial ceremony of “baptizing” the new object or person.

Let us consider now the arguments of the Theory of Direct Reference against the canonical theory, as they appeared especially in Kripke’s work *Naming and Necessity*. The arguments can be seen as being of three main types: modal, epistemological and semantic.

**MA** The modal argument is directed against the view that proper names are disguised definite descriptions (B), and that they refer through these descriptions. For instance, the name “Mark Twain” is commonly used to refer to the well-known American writer. Take into account now the properties that someone might think of in association with this name. These could be seen as constituting its semantic content in a particular circumstance, e.g., “being the most popular American humorous writer of the 19th Century; the author of several classic books like *The Adventures of Tom Sawyer*, *The Adventures of Huckleberry Finn*, *A Connecticut Yankee in King Arthur’s Court* and so on.” Consider now that if the canonical theory were correct, then the association of the name with all the characteristics mentioned above, and possibly many others, would express a logical truth. In this case we would have an identity relation between the name and the description associated to it. The relation should be characterized as “analytical” in the traditional sense, and therefore should constitute a necessary truth, namely, a proposition true with regard to all possible worlds. But at this point it is clear that I could very well talk about ‘a’ Mark Twain, who in another possible situation/world did not write *The Adventures of Tom Sawyer*. But, even without writing this book, or generally, doing something different from what he actually did, he would remain the same Mark Twain. It follows that the name “Mark Twain” is not descriptive in terms of the properties mentioned above. Consequently, the name “Mark Twain” continues to refer to the same person even with respect to counterfactual situations in which this individual lacks all of the properties that we actually use to identify him. Thus, one important tenet of the new theory of reference is that such expressions as proper names are *rigid designators*.17

**EA** The epistemological argument is also due chiefly to Kripke. Assuming that the canonical theory is correct, the connections between names and their associated definite description(s) should express propositions that are knowable *a priori*, namely without making appeal to any experience or empirical judgments. But, it is not difficult to imagine (counterfactual) situations in which it could be discovered that, for instance, Mark Twain did not write *The Adventures of Tom Sawyer* or *The Adventures of Huckleberry Finn*. Kripke’s

17 This terminology comes from Kripke and means that the expression designates the same thing with respect to the every possible world in which that thing exist.
own version of this argument raises the following question: If any description like ‘the discoverer of the incompleteness of arithmetic’ is associated with the name “Gödel,” does this mean that Gödel refers to that person? What if the theorem was actually proved by an Austrian mathematician, say Schmidt, who died in strange circumstances, and Gödel stole Schmidt’s work and published it as his own?\(^{18}\)

From this argument we conclude that the connection between a proper name and its associated description is a posteriori relation, and, again, it could not serve to secure the referential relation between the name and the designated person.

\((SA)\) The semantic argument goes as follows. Let us assume that the meaning or conceptual content of the name “Leonardo da Vinci” is determined by the description “the Italian painter who painted Mona Lisa.” Suppose now that, due to some mistakes or ignorance, the painter referred to by writers such as Vasari, never painted Mona Lisa.\(^{19}\) We took the name from Vasari, and, consequently, we are using it to pick out a certain individual, even though he is not actually the creator of Mona Lisa. Suppose further that by a very bizarre coincidence there was indeed in the Renaissance an Italian solitary painter, who did in fact paint the picture, though he was unknown to people like Vasari and bears no historical connection to us. To which of these two painters would our name “Leonardo da Vinci” refers? Obviously, to the first of the two painters. The way in which we are using the term excludes from the very beginning any mental association between the name and that other person. But, on the other hand, he is in fact the actual creator of Mona Lisa. However, this doesn’t play any role in determining the semantic relation between the name and its bearer.\(^{20}\)

The “da Vinci” example raises the question that, if our names refer directly and rigidly to their bearers, what secures the connection between the name and the named object? The classical view concerning this mechanism of preserving the speaker reference is explained in the framework of The New Theory of Reference by the appeal to a historical chain of users. A name is introduced into the linguistic community by some “initial baptism” in which an object gains a name, and then the name is passed on from link to link to an actual user of the name. Donnellan and Kripke have provided such accounts of securing reference of proper names by means of such causal chains of communication. Hilary Putnam has given a similar account of certain terms designating something by means of a “division of linguistic labor” and a “structured co-operation between experts and nonexperts.” Due to such characterizations the Theory of Direct Reference is often also called “The Causal Theory of Reference.”

\(^{18}\) As Graeme Forbes pointed out in “Proper names”, the very fact that we can understand this “what if” shows that the reference of the name “Gödel” is not fixed as whoever discovered the incompleteness of arithmetic.

\(^{19}\) Giorgio Vasari (1511–1547), who is the chief source book for the history of Italian Renaissance artists.

\(^{20}\) The argument reinforces in fact Donnellan’s criticism to Russelian point of view, stressing the more common referential role of descriptions versus their attributive usage in the most linguistic contexts.
claim that names have senses is destined to fail from the very beginning. But, in
spite of these considerations, it seems to me that Kripke misses the target, in the
sense in which the “Frege” of the intended “Frege-Russell canonical view”
doesn’t really designate Frege’s theory.

**What is the difference, where is the failure?**

It seems that we are now confronted with a kind of dilemma: on one hand, we
have to grant the existence of senses for proper names in order to explain the
informative content of the identities, and, on another hand, due to Kripke, it seems
clear that proper names have no senses, in the sense in which an identity such as
(8) is *contingent* (it could have been otherwise). My claim is that Frege didn’t
fall in this dilemmatic trap because his proper names have senses, so he can
explain why identities like (2) are informative sentences, and these senses could
not be replaced *salva veritate* with definite descriptions. We believe that it seems
to be disingenuous to say that Frege admits that we cannot substitute the name
“Mark Twain” with another name like “Samuel Langhorne Clemens” without
changing the meaning, whereas he agrees that we can replace without problems
“Mark Twain” by a description like “The author of *The Adventures of Tom
Sawyer*.” The misunderstanding is to interpret him (perhaps mostly in the light
of his comments from the mentioned footnote (A) and other similar passages) in
the sense that we can identify senses with definite descriptions. Frege is quite
clear with regard to this point: proper names have senses, and the senses secure
the link between names and the designated objects. But if for Frege senses are not
to be understood as being descriptions, then what are they? As we saw, candidates
for this job are: “modes of presentation,” “modes of determination,” “way of
thinking,” . . .

Whatever they might be, senses must fulfill two main roles: to be
informative and to secure the reference. But, if senses secure the connection
between names and objects, does this mean that always when we grasp a sense of
a proper name we are ensured about the existence of its referent? Not necessarily.
There are empty proper names, names that possess senses but lack references. The
sense is our route to the designated object, and in such cases there is simply no

---

21 For the distinction between ‘modes of presentation’ and ‘modes of determination,’ see
Michael Beaney, *Frege: Making sense*, p. 168:

If we want to make room for there being senses without referents, then we need to talk
not of ‘modes of presentation’ but of ‘modes of determination’. On this more liberal
conception, senses are routes to referents, and it may well turn out that there is nothing -
or at least not what was expected - at the end of the journey. Travelling along the path
set by ‘the greatest Being’, for example, may result in the realization that there can be no
such thing; but (arguably) the term is not senseless. Modes of determination can then be
regarded as the basic type of sense—modes of presentation being modes of determination
plus something else—the object being appropriately present. Frege readily utilizes
both conceptions, but does not distinguish them - just because of his assumption, which
we have emphasized (and criticized) at various points, that names in a logical language
presuppose the existence of objects.
object (waiting for us) at the destination. However, we are using such “mock” proper names as if they were standing for an actual object.

We should distinguish between two types of senses in the case of Frege’s account of proper names. One is the general sense (S1), namely the Fregean sense understood as the vectorial route from the sign to the object; the other one is a particular sense (S2), namely, the content attached by us to the name in various circumstances. This latter sense is determined by the linguistic context and our knowledge of the designated reference. S1 is the way, S2 is the vehicle.

In the practice of language we need first to identify an expression as being a proper name, namely as directing us to the object, and after taking this way we are going to make the trip to the reference with the help of a particular vehicle, namely with a particular cognitive content. The implicit conventions of a natural language ensure the existence of a S1 in the cases of proper names. This S1 entitled Frege to maintain that senses are objective (non-mind-dependent) and communicable (between speakers of the same language). On the other hand, with the help of the constraints of a particular context, we can identify (and express) S2. This is the “meaning” of an object, naming its conventional significance in a determined linguistic context. S2 could be seen as a definite description or as a cluster of such descriptions.

S1 and S2 reveal different semantic roles of proper name in a language. S1 ensures the link between name and object, whereas S2 displays the descriptive content attached to a particular name. S1 belongs primarily to a theory of reference, while S2 is central for a theory of meaning. Thus, we agree with Noonan that: “We can conclude that despite his mistaken assumption that Frege had to hold the sense of proper names to be descriptive, Kripke’s arguments against the “Frege-Russell” theory of names cannot be dismissed from the outset as merely irrelevant to Frege’s position.”

Kripke is relevant in the way that he implicitly pointed out very clearly that we should not mix up S1 with S2, namely, that we should not try to secure the reference of a proper name through a definite description. In the light of the previous distinction it should be clear now that Kripke makes a confusion between S1 and S2. Thus, he is mixing up the way to something with the vehicle itself, which is taking us to the destination. But the road and the car should remain distinct traffic means, if we want to end well the journey. Therefore, the general moral of the story would be that we should keep distinct two interconnected functions of proper names: to refer to something and to mean something.

The theory rejected by Kripke’s may be called “The Descriptivist Theory of Proper Names,” and can be characterized as comprising two main claims: that proper names are concealed definite descriptions and that these descriptions constitute their senses. “Sense” here is taken to be the Fregean Sinn, namely the conceptual content which makes the semantic connection between a sign and its reference (Bedeutung). But, as we have already presented, none of these claims are in fact endorsed by Frege, and therefore Kripke’s enterprise, even highly

\[22\] Noonan, Frege: A critical introduction, p. 216.
valuable and ingenious, does not concern Frege’s view on proper names.\textsuperscript{23}

\textbf{Concluding remarks}

Summing up the discussion, the main tenet of \textit{The New Theory of Reference} (the Theory of Direct Reference) is that proper names are directly referential rigid operators. The strict connection between a name and its bearer is preserved in the virtue of a causal-historical chain mechanism. Even though this rigidity is a questionable issue which deserves further investigation, the problem that concerns us now is the characterization of names as directly referential devices. We have to direct our attention to Mill’s view, and to face again the puzzling problem of identity sentences. This clearly offers us a good basis for the refutation of the directness conception of proper names. In this case we are urged to eliminate this Millian directness from our conception and to preserve exclusively the idea of proper names as rigid designators. In fact there is nothing preventing us from introducing a name by a description, but using it to refer via something else. Thus, the result of the previous critical investigation is that \textit{a name is a “rigid” but not a “direct” linguistic device}. Being Millian inevitably implies being rigid, but the reverse is not necessarily true. Lycan pointed out explicitly

\begin{quote}
Being Millian certainly implies being rigid. But the reverse does not hold. Although Kripke cites Mill and argues that names are rigid, rigidity does not imply being Millian.\textsuperscript{24}
\end{quote}

There are definite descriptions which are rigid, the obvious cases being the descriptions of mathematical truths.\textsuperscript{25} Therefore, the conclusion of the is that a name is a \textit{rigid indirect operator}, and that Frege’s view on proper names is still a gold mine that should be successfully exploited further.

\textbf{Bibliography}

Bell, D.


Beane, M.


Devitt, M.


200? A shocking idea about meaning. Forthcoming in a special issue on Putnam in \textit{Revue Internationale de Philosophie}.

\textsuperscript{23} The attacks do not concern entirely Russell’s view. As we saw, even Kripke acknowledges that, but only his claim that names abbreviate definite descriptions (B). But this may be seen as a separate and independent position from his theory of descriptions, so it may be the case that Russell’s view is valid as well.

\textsuperscript{24} Lycan, \textit{Philosophy of Language}, p. 55.

\textsuperscript{25} Lycan, \textit{Philosophy of Language}, p. 55:

Arithmetical truths (which are necessary truths \ldots) such as “the positive square root of nine” are rigid, because they designate the same number in every possible world, but are certainly not Millian because in order to secure their reference they exploit their conceptual content.
Donnellan, K.

Evans, G.
1982 The Varieties of Reference, Oxford University Press.

Forbes, G.

Frege, G.
1892 Über Sinn und Bedeutung, in Zeitschrift für Philosophie und philosophische Kritik, 100; translated as “On sense and reference”, in Harnish, 1994, from which the quotations are taken.

Harnish, A., ed.
1994 Basic Topics in the Philosophy of Language, Prentice-Hall.

Kaplan, D.

Kripke, S.
1980 Naming and Necessity, Harvard University Press.

Lycan, W.

Marti, G.

Mill, J.S.

Neale, S.

Noonan, Harold

Pears D.F., ed.

Putnam, H.

Russell, B.

Salmon, N.

Searle, J.

Strawson, P.
1950 On referring, in Martinich, 1996.
Generating Necessary *A Posteriori* Truths

Desidério Murcho

One should always say in the first sentence of a paper what is its aim. I ask for your patience; I cannot do that in this case. To state the aim of this paper I’ll have to say a few things first.

Saul Kripke presents a picture of the way we reach necessary *a posteriori* truths:

One knows by a priori philosophical analysis, some conditional of the form “if P, then necessarily P.” . . . On the other hand, then, we know by empirical investigation that P, the antecedent of the conditional, is true. . . . We can conclude by modus ponens:

\[ P \Rightarrow \Box P \]

\[ P \]

\[ \Box P \]

. . . this conclusion is known a posteriori, since one of the premises on which it is based is a posteriori. ²

An example of the form above might be “If water is H₂O, water is necessarily H₂O; water is H₂O; therefore, water is necessarily H₂O.”

Why we should accept a statement like “If water is H₂O, water is necessarily H₂O” is not our concern here. Kripke showed, or tried to show, that we have some clear examples of true conditionals like this. I think he is right, but that is not what we will discuss here. Let’s assume that he is right, that there are true conditionals like this. Does it follow that there are necessary *a posteriori* truths? No, unless he is also right when he claims that the conclusion of the argument above is *a posteriori*.

Our problem is the “since” at the end of Kripke’s quotation above. As we shall see, we may run into trouble trying to spell out exactly what principle, if any, is at work. Why is it that when one of the premises is *a posteriori* the conclusion is also *a posteriori*? What we need is an explanation. We need to explain why the conclusion of a valid argument is *a posteriori* if one of the premises is *a posteriori*. That is the aim of this paper.

This is important because it’s easy to think that all necessary truths are *a priori* and all *a posteriori* truths are contingent. In fact, many philosophers have

---

¹ This paper was presented in Bucharest, Romania, in July 2000, at the 2nd Meeting of the Advanced Reasoning Forum. I am grateful to Mircea Dumitru and everybody at the New Europe College for providing us with such a wonderful meeting and lively discussions. I am also grateful for valuable comments and discussion with Yannis Stephanou at King’s College London, as well as several suggestions and corrections made by Richard L. Epstein and an anonymous reader.

² “Identity and necessity”, p. 88.

© 2001 Desidério Murcho.

thought so, namely Kant and Hume. Kripke rejected this idea, presenting several examples of necessary *a posteriori* truths. They all rely on that “since” in the quoted sentence. To my knowledge, no one has explained that “since.” If there’s no reason to accept that the conclusion of a valid argument is *a posteriori* just because one of its premises is *a posteriori*, then there’s no reason to accept that there are necessary *a posteriori* truths. In that case, we would have no reason to accept one of the most interesting and revolutionary philosophical ideas of the twentieth century.

The point of this paper is not philosophical exegesis. This is not a paper about what Kripke thinks about this matter. I do not know what he thinks. As far as I know, his texts do not address this problem; but that is not the point. The point here is philosophical rather than exegetical: How can we understand the relation between the epistemic status of a conclusion in a valid argument, and the epistemic status of its premises? Is there some special principle that regulates such relation? I will argue that we do not need any special principle to understand that relation; we need only to understand what *a priori* knowledge is.

Let’s rewrite Kripke’s *modus ponens* and call it KMP:

(a) \( P \supset \Box P \)

(b) \( P \)

(c) \( \Box P \)

To explain why (c) is *a posteriori*, we might want to say this:

(1) If one of the premises of a scheme like KMP is *a posteriori*, then the conclusion will also be *a posteriori*.

This principle is not good for two reasons.

First, it speaks only of some obscure “inheritance” scheme: Somehow the epistemic status of the premises of an argument is passed on to its conclusion. However, in the cases at hand, namely the case of water, the first premise is not *a posteriori*; why does the conclusion have to be *a posteriori*? Why is it not that if one of the premises is *a priori* the conclusion will also be *a priori*? At best, (1) tells us what may be going on with KMP, but not why.

Second, it’s not general enough. What’s so special about *modus ponens*? It’s not plausible that whatever is going on has to do with *modus ponens* in particular. We should generalise principle (1) to hold in any valid reasoning.

(2) If one of the premises of a valid reasoning is *a posteriori* then the conclusion will also be *a posteriori*.

This principle is not good, because there are obvious counter-examples:

(3) If water is \( H_2O \), then necessarily \( 2 + 2 = 4 \).

   Water is \( H_2O \).

   *Therefore*, Necessarily \( 2 + 2 = 4 \).

---

3 The epistemic status of a proposition is whether that proposition is *a priori* or *a posteriori*. Below I will spell out what is *a priori* and *a posteriori* knowledge.
There’s a trick here, of course: There’s no content relation between the antecedent and the consequent of the first premise. The premise is true, but it’s something like a vacuously true statement: It is true due to the fact that its consequent is necessarily true. Whatever antecedent you pick up, those premises will always be true.

Principle (2) would have to be reformulated like:

(4) If one of the premises of a valid inference is a posteriori, then the conclusion will also be a posteriori, unless one of the premises relevant to the conclusion has no epistemic relevance.

Of course, we would now have to present a good characterisation of “epistemic relevance.” Maybe we can do that. However, we would have to deal with all kinds of counter-examples, not just “vacuously true” conditionals. Consider this counter-example:

(5) Water is not H₂O or necessarily 2 + 2 = 4.
   Water is H₂O.
   Therefore, Necessarily 2 + 2 = 4.

Here we do not have the “trick” of a vacuously true conditional. Principle (4) is not promising, even if we try to enlarge it to cover all kinds of reasoning; there is so far no systematic way of ruling out counter-examples like (5). Moreover, if we try to rule out counter-examples the way (4) did with (3), then we will end up with a principle that may work just because we ruled out in a completely ad hoc fashion all possible counter-examples. This is not a promising path. We still do not know why the a posteriori character of one premise is passed on to the conclusion, while the a priori character of the other premise is not passed on to the conclusion. At best, principle (4) or some other version of it will state what is going on, but will not tell us why. Anyway, there is another possibility, perhaps more enlightening and surely simpler.

How can we characterise a priori and a posteriori knowledge? One way to do it, suggested by Kripke’s own work, is this:

(6) P is a priori iff P can be known by reason alone.
    P is a posteriori iff P must be known by experience.

This doesn’t mean that all a priori truths have to be known by reason alone. The example of Kripke to illustrate this point is the calculator. When I use a calculator to determine the result of, say, 122 × 98, I get an answer that is true, but I come to know it by experience. It is my confidence in the laws of physics and

---

4. A small detail that is worth noting is this. Knowledge is “factive.” This means we cannot know falsehoods. I cannot know that the earth is flat, because it is not flat. Of course, I can falsely believe that the earth is flat. This shows that belief is not factive, whereas knowledge is. For our purposes, this means that there are neither a priori nor a posteriori falsehoods. We can know a priori that a statement is false—like “2 + 3 = 13”—but that is not to know a priori that 2 + 3 = 13.


6. Ibid., p. 35.
my experience that justifies my belief that the result delivered by the calculator is correct. But that doesn’t mean (7) is *a posteriori*.

(7) $122 \times 98 = 11956$

According to our characterisation of *a priori* and *a posteriori* knowledge, (7) is still *a priori*, since it can be known by reason alone.\(^7\) I may use the calculator to determine the result, but I can also calculate it myself.

Now consider a statement like this:

(8) Water is $H_2O$.

There’s no way I can know this statement is true by reason alone. Therefore, (8) is *a posteriori*.

Now consider the following valid reasoning:

(9) If water is $H_2O$, then water is necessarily $H_2O$.

Water is $H_2O$.

*Therefore*, Water is necessarily $H_2O$.

There is no way of knowing the conclusion is true without knowing that the second premise is true; but the second premise is *a posteriori*. Therefore, the conclusion is also *a posteriori*. Consider again case (3) above. I do not have to know that water is $H_2O$ in order to know that necessarily $2 + 2 = 4$. I can know that the conclusion of (3) is true by reason alone.

Now we can see that the principle at work in order to generate necessary *a posteriori* truths is this:

(10) If $P$ must be known by experience, then $P$ is *a posteriori*.

So, (3) is not a counter-example to principle (10), because the conclusion of (3) can be known by reason alone.

In fact, we do not need (10) as an independent principle to explain what is at work in generating necessary *a posteriori* truths. For (10) is just a corollary of (6). All we need, then, to understand what’s going on is a clear understanding of the notion of *a priori* and *a posteriori* truth. With that in mind, we understand why the conclusion of (9) is *a posteriori*. It’s not due to some metaphorical and

\(^7\) It was pointed out to me that this “can” is obscure. Do we mean “can, in theory” or “can, in practice”? Some highly complex mathematical theorems can only be proved using a computer. Does this mean that those theorems are *a priori* because if our natural reasoning resources were better we could run the proofs ourselves? Or does it mean they are *a posteriori* because we cannot actually run those proofs ourselves? Actually, we should not talk abstractly about “*a priori* knowledge” at all but about particular statements known *a priori* by some specific agent. It’s not hard to see that the concept of *a priori* knowledge has to be agent-relative; God knows perhaps *a priori* many things that we only know *a posteriori*. An advanced mathematician knows many theorems *a priori* that I know of only *a posteriori* because he tells me they are true, and I may never be able to crack them — because I’m no mathematician. Our purpose, however, is to explain why one *a posteriori* premise in a valid argument makes the conclusion also *a posteriori*. What I show is that there’s a general way of understanding *a priori* and *a posteriori* knowledge that explains it. A deep understanding of *a priori* knowledge is not the topic of this paper.
obscure “inheritance.” It’s because that conclusion cannot be known by reason alone; it can only be known by some reasoning in which at least one of the premises is *a posteriori*.

Let’s consider again KMP. Now we understand why premise (b) determines the epistemic character of the conclusion, but not premise (a). We can also understand that what’s wrong with case (3) is not the “vacuously true” first premise. It happens that we can arrive at that conclusion by other means, means that do not involve *a posteriori* premises—and that makes all the difference.

We can now explain what is at work in Kripke’s way of generating necessary *a posteriori* truths. We do not need a principle to explain what is going on; we just need to understand a certain conception of *a priori* and *a posteriori* truth. Whether that conception is right is another matter. If something is wrong with the idea of necessary *a posteriori* truths, it lies not in the “inheritance” scheme.

**Bibliography**

Kripke, Saul


Ex Contradictione Non Sequitur Quodlibet

Walter A. Carnielli and João Marcos

We summarize here the main arguments, basic research lines, and results on the foundations of the logics of formal inconsistency. These involve, in particular, some classes of well-known paraconsistent systems. We also present their semantical interpretations by way of possible-translations semantics and their applications to human reasoning and machine reasoning.¹

1. Do we need to worry about inconsistency?
Classical logic, as we all know, cannot survive contradictions. Among the principles that were gradually incorporated into the "properties of correct reasoning" since Aristotle, the Principle of Pseudo-Scotus (PPS), also known since medieval times as ex contradictione sequitur quodlibet (and also called the Principle of Explosion by some contemporary logicians), states that in any theory exposed to the enzymatic character of a contradiction A and ¬A one can derive any other arbitrary sentence B, so that the theory would turn out to be trivial. Another principle called the Principle of Non-Contradiction (PNC) states that there should be theories from which no such contradictions are derivable. To those principles, one could add the Principle of Non-Triviality (PNT), stating that there should be at least one theory and one sentence B such that B is not derivable from this theory.

In order to fully understand what those principles mean, what the relationship is between them and what their importance is for modeling the concept of inconsistency let us introduce some formalism. This formalism will be apt for the syntactical approach to the logics of formal inconsistency we discuss in the first three sections of this paper. However, the reader should be aware that it is possible to start from a purely syntactical account, as we do in Sections 4 and 5.

Let $\text{For}$ be a collection of formulas of a certain language, and call a theory any subset of $\text{For}$. Let a consequence relation $\vdash$ over $\text{For}$ be a relation between theories and formulas of $\text{For}$, that is, $\vdash \subseteq (\mathcal{P}(\text{For}) \times \text{For})$, where $\mathcal{P}(\text{For})$ denotes the power set of $\text{For}$. If $\Gamma$ is a subset of $\text{For}$, we write $\Gamma \vdash A$ when $\langle \Gamma, A \rangle \in \vdash$. We write $\Gamma \not\vdash A$ when it is not the case that $\Gamma \vdash A$. We define a logic $L$ to be a structure constituted of $\text{For}$ and the relation $\vdash$. The consequence relation of a given logic is often defined by its axioms and rules, or else from some semantical interpretation associated with the logic.

¹ This material was discussed during the II World Congress of Paraconsistency (WCP'2000) in Juquehy, SP, Brazil, and will appear more fully in W. A. Carnielli, and J. Marcos, "A taxonomy of C-systems."

© 2001 Walter A. Carnielli and João Marcos.
Some basic assumptions on what the relation \( \vdash \) would have to obey in order to be considered a consequence relation, known as the Tarskian conditions, are the following:

1. **Reflexivity**  
   If \( A \in \Gamma \), then \( \Gamma \vdash A \).

2. **Monotonicity**  
   If \( \Gamma \vdash A \) and \( \Gamma \subseteq \Delta \), then \( \Delta \vdash A \).

3. **Transitivity**  
   If \( \Gamma \vdash A \) and \( \{ \Delta, A \} \vdash B \), then \( \Gamma \cup \Delta \vdash B \).

These conditions allow for the characterization of an immense number of distinct logics, but they still can be made more permissive (as, for example, weakening the requirement of monotonicity in order to characterize non-monotonic logics). In our present study, we will stick to these three basic assumptions and consider the effect of some additional properties of the relation \( \vdash \).

Fix some logic \( L \) for the following discussion. A theory \( \Gamma \) of \( L \) is said to be:

1. **Contradictory with respect to \( \neg \)** (or simply contradictory)  
   If there exists a formula \( A \) such that \( \Gamma \vdash \neg A \) and \( \Gamma \vdash \neg \neg A \).

2. **Trivial**  
   If for every \( A \), we have \( \Gamma \vdash A \).

3. **Explosive**  
   If for every \( A \), we have \( \Gamma \cup \{ A, \neg A \} \vdash B \).

These definitions are important for distinguishing theories from their underlying logic: A logic \( L \) is contradictory, trivial, or explosive if, respectively, all of its theories are contradictory, trivial, or explosive.

We can now restate PNC, PNT, and PPS in more formal terms.

- **The Principle of Non-Contradiction (PNC)** for a logic \( L \)  
  \( L \) should have non-contradictory theories, that is, there should be some theory \( \Gamma \) such that for no \( A \), \( \Gamma \vdash A \) and \( \Gamma \vdash \neg A \).

- **The Principle of Non-Triviality (PNT)** for a logic \( L \)  
  \( L \) should have a non-trivial theory, that is, there must exist a theory \( \Gamma \) and some formula \( A \) such that \( \Gamma \vdash A \).

- **The Principle of Pseudo-Scotus (PPS)** for a logic \( L \)  
  \( L \) should have only explosive theories, that is, for every theory \( \Gamma \), \( \Gamma \cup \{ A, \neg A \} \) is trivial.

For classical logic, of course, all these principles hold. It seems intuitively acceptable that PNT should be taken as the most important of these three principles—after all, if PNT does not hold for a certain logic, then every \( \Gamma \) would entail every \( A \) and the relation \( \vdash \) would be total, that is, \( \vdash \) = \( (\emptyset, \text{For}) \times \text{For} \). In that case, \( \vdash \) would not be an interesting deductive relation, for it would stop “making the difference,” failing to give, we argue, any special meaning to the notion of derivability. We shall, accordingly, accept PNT throughout this study, avoiding the consideration of trivial logics.

It is also interesting to remark that, using the property of monotonicity
mentioned above, it is sufficient to consider PPS \( \varnothing \) valid for just the empty theory: Call PPS\( \varnothing \) the statement "\( \{A, \neg A\} \) is trivial." Then, in view of monotonicity, we conclude that PPS\( \varnothing \) implies PPS. The converse (PPS implies PPS\( \varnothing \)) is obvious. Similar reasoning applies to PNC and PNT. Call PNT\( \varnothing \) the statement "the empty theory is not deductively trivial" and call PNC\( \varnothing \) the statement "the empty theory is non-contradictory." By monotonicity, PNT implies PNT\( \varnothing \), and the converse is obvious, and the same for PNC and PNC\( \varnothing \). Thus, for monotonic logics we do not need to care about the distinctions between PPS, PNT, PNC and their respective counterparts, PPS\( \varnothing \), PNT\( \varnothing \), PNC\( \varnothing \). Dealing with non-monotonic logics, the distinctions would have to be taken into consideration, but our primary treatment of the question would still apply.

We intend to discuss here the logics of formal inconsistency, which constitute a large class of paraconsistent logics where the notion of inconsistency can be linguistically expressed. For those logics, we may adopt or not a cautious position to the effect that not only PNT but also PNC should be required—although we want our logics to be able to support contradictory theories, we may not want that our logics derive contradictions.

Paraconsistent logics are often misunderstood as logics that inevitably derive contradictions. This is clearly a mistake. Although there exist some logics (the so-called dialectical logics, or logics of impossible objects) that violate both PPS and PNC and have theses which are not classical theses, this particular case of paraconsistent logics will not be studied here.\(^2\) The logics surveyed in this study just support contradictions and permit reasoning with them, but neither engender contradictions nor validate any bizarre form of reasoning. To the contrary: As we shall see, the logics of inconsistency are, in a sense, "more conservative" than classical logic.

While PNT and PNC can be regarded as ontological principles in that they presuppose the existence of certain theories inside our logics, PPS can be seen as a kind of flexible, operative principle: It describes how the logic works when its theories are exposed to contradictory formulas, and thus it seems likely to be changed or even discarded by some logics and situations in which the trivializing operation that PPS describes is not justified. So, the starting point to our approach to paraconsistency will be the cautious one, and we will ask ourselves: (1) While maintaining PNT and PNC, is there any good reason why we should challenge PPS? (2) If so, should we modify or simply abandon this principle?

The first is a factual question: We should try to challenge PPS if there is sufficient demand, and we will argue that there is. The second is a foundational question of logico-mathematical character, and we will also argue that it is plainly possible to construct a great variety of interesting logics alternative to (propositional and quantified) classical logic maintaining PNT and PNC, while modifying only PPS.

An interesting example of an intellectual activity where holding contradictory or inconsistent hypotheses is more the rule than the exception is abductive reasoning, conceived of as reasoning that looks for explanatory hypotheses and the evaluation of such hypotheses. Scientific activity, and in particular medical diagnosis, generally uses abduction when looking for explanations. According to P. Thagard, and C. Shelley, this is inevitable:

We are not urging inconsistency as a general epistemological strategy, only noting that it is sometimes necessary to form hypotheses inconsistent with what is currently accepted in order to provoke a general belief revision that can restore consistency.\(^3\)

A second example of how contradictions can easily be incorporated into reasoning (independent of any ontological commitments to the actual existence of concrete inconsistent objects in the world) concerns the domain of machine intelligence and the efforts to find rules for automated reasoning. The following example given by A. Rose in “Remarque sur les notions d’indépendance et de non-contradiction” was intended to show that the concepts of independence and triviality in a formal system are themselves independent of each other, but it can also be used as an example of how certain rules and procedures, widely used in the formalization of machine reasoning, must face the problem of contradictions. One such procedure is the closed-world assumption, largely used in databases and logic programming, which proposes that, if from a certain (knowledge-based) system \(S\) one cannot infer information \(A\) and know nothing about its negation \(\neg A\), then one would be entitled to add \(\neg A\) to \(S\) by default. Consider now the following fragment \(S\) of classical propositional logic closed under modus ponens and the substitution rule:

1. \(((A \rightarrow \neg A) \rightarrow A) \rightarrow A\)
2. \((A \rightarrow (B \rightarrow A))\)
3. \((\neg B \rightarrow \neg A) \rightarrow (A \rightarrow B))\)
4. \((A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))\)

Consider also the formula \(D = (\neg A \rightarrow A) \rightarrow A\). It is not hard to show that \(S\) does not entail either \(D\) or \(\neg D\), but if \(\neg D\) is adjoined to \(S\) the resulting system becomes trivial. Since the closed-world assumption has exactly the effect of adjoining \(\neg D\) to \(S\), it is clear that that assumption may produce contradictions, which will result in triviality due to PPS. It is, in fact, not hard to see that PPS holds in \(S\):

a. From \(A\) and \(\neg A\), by axiom 2 derive \(((B \rightarrow \neg B) \rightarrow A)\) and \((\neg B \rightarrow \neg A)\).

b. From \((\neg B \rightarrow \neg A)\) by axiom 3 derive \((A \rightarrow B)\).

c. From \(((B \rightarrow \neg B) \rightarrow A)\) and \((A \rightarrow B)\) by axiom 4 derive \(((B \rightarrow \neg B) \rightarrow B)\).

d. From \(((B \rightarrow \neg B) \rightarrow B)\) by axiom 1 derive \(B\).

Hence for every \(\Gamma\), \(A\), and \(B\), we have \(\Gamma \cup \{A, \neg A\} \vdash_{S} B\).

\(^3\) "Abductive reasoning: logic, visual thinking, and coherence."
As we shall see, such a trivialization does not occur in the logics of formal inconsistency that we will consider, for axiom 3 above will in all cases be valid only under the proviso that A is known to be not inconsistent.

2. The contradictory and the inconsistent

The concepts of contradiction and inconsistency need not be taken to be equivalent. That classical logicians take them as equivalent is explained by two facts: first, for classical logicians, “contradiction,” “inconsistency” and “triviality” are usually seen as synonymous, and, second, PNC and PNT are indeed equivalent for several logics. Yet this equivalence is far from necessary.

It is clear that trivial theories, closed or not, are necessarily contradictory (if there is a symbol for negation in the language) since they derive all sentences, negated and non-negated. From the definitions of PNC and PNT above we obtain the following result, which will hold in all of our logics of formal inconsistency:

**Metatheorem I** Every trivial theory is contradictory, that is, PNC implies PNT.

*Proof:* Suppose PNC holds. Then, there exists a theory Γ such that, for every A, either Γ ⊬ A or Γ ⊬ ¬A. Hence there exist Γ and A such that Γ ⊬ A, and this is what PNT states.

The converse, however, does not hold, because an appeal to PPS is necessary in order for a contradictory theory to be trivial. We may add, as our second founding result:

**Metatheorem II** Assuming PPS, contradictory theories and trivial theories coincide; that is, PPS implies: PNT if and only if PNC.

*Proof:* Suppose that PPS holds. It follows then from the transitivity of ⊬ that if a theory is contradictory it is trivial, that is, that PNT implies PNC.

This last metatheorem can be understood as *ex contradictone non sequitur quodlibet*, in the sense that what follows from a contradiction depends on other underlying principles (in this case, it depends on PPS). So, controlling the explosive power of contradictions is necessary to gain control over the destructive effects of trivialization. But is that enough?

It is clear what contradictory and trivial theories are, as those concepts can be defined in terms of purely set-theoretic conditions. It is also clear what the role of PPS is, since, if PPS holds, every contradictory theory turns out to be trivial.

If we can only modify PPS, the sets of consequences of contradictory theories will not necessarily be trivial. But what is a consistent theory? While contradiction and triviality are definable, respectively, in terms of negation (hence as linguistic) and in set-theoretical terms (hence as metamathematical), consistency is better seen as a semantic concept.

The novelty of our approach is that, although acknowledging the semantic
character of the notion of consistency, we endeavor to internalize it in the
language and treat it then from a purely abstract point of view, independently
from contradiction and triviality. Considering that it is PPS that forces negation
to behave in such absolute terms that contradictory theories collapse into triviality,
it is possible, as we shall see, to define forms of "careful" negation in several
ways so as to avoid such collapse.

In traditional logic, inconsistency and contradiction are taken to be just two
ways of referring to the same concept. For example, neither Aristotle nor
Łukasiewicz, in his celebrated analysis of Aristotelian logic, ever mentioned
consistency or inconsistency, but always contradictions, since in classical logic
the distinction is immaterial—for, in this case, the metamathematical concept of
inconsistency is internalized into the object language by means of conjunctions
of contradictory statements.\footnote{J. Łukasiewicz, "On the principle of contradiction in Aristotle."} This is so, however, due to PPS (Metatheorems I
and II). If PPS is not to be accepted as dogma, for the reasons already presented,
such an internalization will not be guaranteed.

It is possible, nonetheless, to conceive of the concept of inconsistency in
such a way that, while contradictory theories are certainly inconsistent, the reverse
might not be true. In a similar way to the concept of point, which is taken in
geometry as a primitive notion only describable through its relationship to other
concepts, inconsistency can be taken in logic to be a primitive notion as well.
From this point of view, inconsistent and contradictory theories do not coincide,
as our logics of formal inconsistency $\text{bC}$ and $\text{Ci}$ will make clear.

Contradictory theories depend strictly on negation and its properties, while
inconsistent theories do not. So, for example, if we say "It is raining in Ghent"
and "It is not raining in Ghent," those are contradictory statements, but should not
necessarily lead to trivialization. They may lead to trivialization if we add some
extra information, for example, that we are talking about the same instant of time,
and that the concept of instant of time is sharp enough so as to exclude the
possibility of rain and absence of rain at a given instant of time. Or we may say
"It is raining at 11h 34m22s" and say "It is not raining at 11h34m22s," which
again are contradictory statements, but do not necessarily lead to trivialization.
In this case, they may become so if we add extra information, for example that we
talking about the same point in space, and that the concept of point is strict
enough so as to exclude rain and the absence of rain.

Von Wright suggests that the Kantian conceptions of space and time in the
Critique of Pure Reason\footnote{Compare I. Kant, Critique of Pure Reason, Transcendental Aesthetic II, §4.} (though with different aims) are not alien to this kind
of intuition:

If this representation [he refers to time] were not an a priori (inner) intuition,
no concept, no matter what it might be, could render comprehensible the
possibility of an alteration, that is, of a combination of contradictorily
opposed predicates in one and the same object, for instance, the being and the
not-being of one and the same thing in one and the same place. Only in time
can two contradictorily opposed predicates meet in one and the same object, namely, one after the other.6

The point here is not whether or not one accepts the transcendental ideality of time, but the role of time (and also space) as examples of entities that bind contradictory statements together and make them inconsistent. In other words, contradictory statements A and ¬A by themselves will not be sufficient to entail any other statement, unless we require an extra condition—in our approach, that A is consistent. The consistency here could be understood as introduced by a Kantian rendering of space and time as concepts which exclude the possibility of the concomitant existence of opposed predicates at the same point of space or time.

Based on this intuition, as an addendum to the results in the Metatheorem I and Metatheorem II, we propose the following as our basic metaprinciple:

**Metaprinciple I**
No contradictory theory is consistent, but a contradictory non-consistent theory need not be trivial.

To this metaprinciple we can consider the addition of another (a kind of converse of Metaprinciple I):

**Metaprinciple II**
Every inconsistent theory is contradictory, but not necessarily trivial.

As we will show, distinct classes of paraconsistent calculi arise, depending on whether we take Metaprinciple II in conjunction with Metaprinciple I. It is noteworthy that virtually all known paraconsistent systems in the literature do assume Metaprinciple II.

It is possible to give models for inconsistent theories, even if those might be regarded as epistemologically puzzling. Obtaining models and understanding their role is an extraordinarily important mathematical enterprise: It required enormous efforts of the most brilliant minds, and more than twenty centuries, until mathematicians would allow themselves to consider models where, given a straight line S and a point P outside of it, one could draw not just one line, but infinitely many or no parallel lines to S passing through P, as in the well-known case of non-Euclidean geometries.

3. **A logic for the illogical?**
The challenge is to find mathematically interesting systems that can provide a foundational sense for what contradictions and inconsistency are and suggest an acceptable semantic interpretation with which people would feel comfortable while reasoning with contradictions. The case of (imaginary) complex numbers seems to make a good comparison: Even if we do not know what they are, and may even suspect there is little sense in insisting on which way they can exist in the “real” world, the most important aspect is that it is possible to calculate with

---

6 G. H. von Wright, “Time, change and contradiction”. 
them. Girolamo Cardano, who first had the idea of computing with such numbers, seems to have seen this point clearly—he failed, however, to acknowledge the importance of it. In 1545 he wrote in his *Ars Magna*:

> Dismissing mental tortures, and multiplying $5 + \sqrt{-15}$ by $5 - \sqrt{-15}$, we obtain $25 - (-15)$. Therefore the product is 40. . . . And thus far does arithmetical subtlety go, of which this, the extreme, is, as we have said, so subtle that it is useless.\(^7\)

His discovery that one could operate with a mathematical concept independently of what our intuition says, and that utility (or something else) could be a guiding criterion for accepting or rejecting experimentation with mathematical objects, certainly contributed to the proof of the Fundamental Theorem of Algebra by C. F. Gauss in 1799, before which complex numbers were not fully accepted.

The idea of proposing logics that enable one to operate with what does not appear to be “rational” goes in the same direction: Good underlying mathematical theory plus usefulness would have to constitute the only criteria to evaluate a mathematical formalism that deals with inconsistency or contradictions.

It is time now to give a more precise definition for the logics of formal inconsistency (*LFI*-systems): An *LFI* is any logic where a syntactic notion of formal consistency can be defined in a syntactical way in such a way that this new notion of formal consistency and the already known notion of contradiction can be related in the light of Metaprinciples I and II. In particular, as we discuss below, in many cases this can be done by endowing the language with a new connective $\circ$ and considering new appropriate axioms.

Among *LFI*-systems it is possible to identify a subclass of the so-called *C*-systems that preserve the positive fragment of some other logic and in which consistency or inconsistency are expressible by means of new connectives. As a subclass of the *C*-systems, we define the *dC*-systems to be those in which the notion of formal consistency can be introduced as a defined connective. The *dC*-systems include several classes of paraconsistent systems, including the ones in the hierarchy $C_n$ of N. C. A. da Costa.\(^8\)

The main axioms we will consider here for the study of an interesting class of *C*-systems based on classical logic are the following: Call $C_{min}$ an appropriate axiomatization of the positive fragment of classical propositional logic in the language $\land, \rightarrow, \land, \lor, \circ$, plus the axioms $(\neg A \rightarrow A)$ and $(A \lor \neg A)$ and closed under the rules of *modus ponens* and substitution.\(^9\) Define the basic logic of formal inconsistency, $bC$, as $C_{min}$ plus the following deduction scheme, where $\circ$ is a new unary operator meant to model “$A$ is formally inconsistent”:

**The Gentle Principle of Explosion** $\circ A, A, \neg A \vdash B$

Note that this axiom is in line with **Metaprinciple I**: A contradictory theory (one containing $A$ and $\neg A$) is not consistent and is not necessarily trivial. It would

---

\(^7\) See J. O'Connor, and E. Robertson, "The MacTutor History of Mathematics Archive".


\(^9\) This system was studied in our "Limits for paraconsistent calculi".
become trivial if, besides being contradictory, it were formally consistent. In such a case its very consistency becomes contradictory, and this situation leads to triviality.

Examples of derived consequences of \( \mathbf{bC} \) are:

1. \( \mathcal{A}, \mathcal{\neg A} \vdash \mathcal{\neg \circ A} \)
   If \( \mathcal{A} \) is contradictory, then \( \mathcal{A} \) is not formally consistent.

2. \( \circ A \vdash \mathcal{\neg (A \land \neg A)} \)
   If \( \mathcal{A} \) is formally consistent, then \( \mathcal{A} \) is non-contradictory. (1st form)

3. \( \circ A \vdash \mathcal{\neg (\neg A \land A)} \)
   If \( \mathcal{A} \) is formally consistent, then \( \mathcal{A} \) is non-contradictory. (2nd form)\(^ {10} \)

It is clear that the theorems above are variations on Metaprinciple I. A very important observation is that in \( \mathbf{bC} \) the notions of “not consistent” and “inconsistent” do not coincide. Indeed, even if we introduce the concept of “consistent” as internal to the language through a new symbol \( \bullet \), understanding \( \bullet A \) to model “\( \mathcal{A} \) is formally inconsistent,” \( \circ A \) and \( \mathcal{\neg \circ A} \), and \( \mathcal{\neg A} \) and \( \neg \mathcal{A} \) would not be interdefinable, contrary to what one might assume. This will be further clarified below.

In \( \mathbf{bC} \) a new negation, called strong negation, can be defined as:

\[ \sim A \equiv_{\text{Def}} \mathcal{\neg A} \land \circ A \]

This recovers several features of classical negation, though not all. We have, for instance, \( \mathcal{A}, \sim A \vdash_{\mathbf{bC}} \mathcal{B} \), and thus \( \mathbf{PPS} \) holds relative to this strong negation (that is, this negation is explosive). But \( \vdash_{\mathbf{bC}} (A \lor \mathcal{\neg A}) \) and \( \vdash_{\mathbf{bC}} (A \rightarrow \sim \sim A) \). Consequently, the strong negation \( \sim A \) is not classical, even though it is explosive (intuitively, \( \sim A \) is somehow analogous to intuitionistic negation). But another strong negation, this one having all the properties of classical negation, is definable in \( \mathbf{bC} \) by letting:

\[ \sim A \equiv_{\text{Def}} A \rightarrow (A \land \sim A) \]

Yet another logic of formal inconsistency, \( \mathbf{Cl} \), is defined by first defining \( \bullet A \equiv_{\text{Def}} \mathcal{\neg \circ A} \) and adjoined to \( \mathbf{bC} \) the deduction scheme:

\[ \bullet A \vdash (A \land \mathcal{\neg A}) \]

In the system \( \mathbf{Cl} \), inconsistency and contradictions are equivalent to each other, due to the fact that \( (A \land \mathcal{\neg A}) \vdash_{\mathbf{bC}} \circ A \) (as noted before) and the axiom just introduced, plus the definitions. Omitting the definitions, however, we obtain intermediate logics between \( \mathbf{bC} \) and \( \mathbf{Cl} \) that realize Metaprinciple II.

Moreover, in \( \mathbf{Cl} \) both strong negations mentioned above are equivalent and acquire all the properties of classical negation—but it is still possible to define non-classical strong negations in \( \mathbf{Cl} \), via for example \( \mathcal{\neg \neg A} \) and \( \mathcal{\neg \neg \neg A} \). Some properties of this logic are:

---

\(^ {10} \) In many systems of paraconsistent logic, although \( (A \land \mathcal{\neg A}) \) and \( (\mathcal{\neg A} \land \mathcal{\neg A}) \) are equivalent, \( \mathcal{\neg (A \land \mathcal{\neg A})} \) and \( \mathcal{\neg (\mathcal{\neg A} \land \mathcal{\neg A})} \) are not. See, e.g., João Marcos, Possible-Translations Semantics.
(4) \(\Box A, \Box A \vdash_{\text{Cl}} B\) and \(\Box A, \Box A \vdash_{\text{Cl}} B\).

The Principle of Explosion holds for formally consistent (or formally inconsistent) formulas.

(5) \(\vdash_{\text{Cl}} \Box \Box A\) and \(\vdash_{\text{Cl}} \Box \Diamond A\)

Both consistent and inconsistent formulas are consistent.

(6) \(\Box A \vdash_{\text{Cl}} \Box \Box A\)

If a formula is consistent, its negation is also consistent.

(7) \(\Box \Box A \vdash_{\text{Cl}} \Box A\)

A formula is inconsistent if its negation is inconsistent.

What doesn’t hold in this logic? PPS still does not hold, that is, \(A, \Box A \vdash_{\text{C}l} B\) for some \(A\) and \(B\). De Morgan’s Laws and the rule of contraposition only hold in restricted forms, e.g., \((A \rightarrow B) \vdash_{\text{Cl}} (\Box \rightarrow \Box A)\), though \((A \rightarrow B) \vdash_{\text{Cl}} (\Box \rightarrow \Box B \rightarrow \Box A)\).

Also, Cl does not prove any formulas to be consistent, unless they already refer to consistency or inconsistency. That is, \(\Diamond A\) is provable in Cl if and only if \(A\) is itself of the form \(\Box \Box B, \Box B, \Box B, \text{ or } \Box \Box B\).

The converses of (2) and (3) are still not valid, and thus we may consider the addition of some other deduction schema to Cl, as for example:

- Levo-based scheme for contradictoriness (1): \(11\)
  \(\Diamond (A \land \Box A) \vdash A\).
- Dextro-based scheme for contradictoriness (d):
  \(\Diamond (\Box A \land A) \vdash A\).
- Bi-directional scheme for contradictoriness (b):
  \(\Diamond (A \land \Box A) \lor \Diamond \Box (A \land A) \vdash A\).
- Global scheme for contradictoriness (g):
  \(\vdash (B \leftrightarrow (A \land \Box A)) \rightarrow (\Box B \leftrightarrow \Diamond (A \land \Box A))\).

In formal terms, a dC-system is any C-system where \(\Box\) and \(\Diamond\) can be defined in terms of the usual connectives of the language \(\land, \lor, \rightarrow\). In the case of da Costa’s C_1, the strongest calculus of his hierarchy of paraconsistent logics in his Inconsistent Formal Systems, \(\Box A\) is defined as \(\Box (A \lor \Box A)\) and \(\Diamond A\) as \(\Box \Box A\), so that this logic is an extension of Cil, the logic obtained by the addition of the axiom (l) to Cl.

Several other distinct dC-systems can be defined by choosing adequate axioms of “propagation” for consistency and inconsistency. Many different choices are possible. Depending on the particular ones we choose, we may obtain finite many-valued paraconsistent logics or infinite-valued paraconsistent logics.

Some choices which have been tried for the axioms of propagation are the following:

**First Choice:** \((\Box A \land \Box B) \rightarrow \Box (A \# B)\) for every binary connective \(\#\).

**Second Choice:** \((\Box A \lor \Box B) \rightarrow \Box (A \# B)\) for every binary connective \(\#\).

---

11 This axiom holds for the system C_1 of da Costa, for example, as discussed in W. A. Carnielli, “Possible-translations semantics for paraconsistent logics” and J. Marcos, Possible-Translations Semantics.
Third Choice \((\circ A \# \circ B)\) for every binary connective \(\#\).

Fourth Choice
\[
\begin{align*}
\circ (A \land B) &\leftrightarrow ((\circ A \land B) \lor (\circ B \land A)) \\
\circ (A \land \circ B) &\leftrightarrow ((\circ A \land \neg B) \lor (\circ B \land \neg A)) \\
\circ (A \rightarrow B) &\leftrightarrow (\circ B \land A)
\end{align*}
\]

Fifth Choice \(\circ (A \# B) \leftrightarrow (\circ A \land \circ B)\) for every binary connective \(\#\).

Sixth Choice \(\circ (\neg A)\) for every formula \(A\).

Seventh Choice \(\circ A\) for every formula \(A\).

The first choice plus (I) above (p. 97) defines the calculus \(C_1\) of the hierarchy of \(C_n\); the second choice plus (I) defines the calculus \(C_1^+\) of N. C. A. da Costa, J.-Y. Béziau, and O. Bueno, "Aspects of paraconsistent logic." Neither of these are many-valued. The third and sixth choices plus (b) define the three-valued maximal logic \(P^1\), introduced in A. M. Sette, “On the propositional calculus \(P^1\).” The third choice plus (b) and \((A \rightarrow \neg A)\) defines another maximal three-valued logic.\(^{12}\) The fourth and fifth choices define, respectively, the three-valued maximal logics \(LFI_1\) and \(LFI_2\) studied in W. A. Carnielli, J. Marcos, and S. de Amo, “Formal inconsistency and evolutionary databases.” And finally the seventh choice defines classical propositional logic. Many other combinations are possible.\(^{13}\)

The fourth choice gives yet another axiomatization for the three-valued paraconsistent calculus \(J_3\), bearing some resemblance to the axiomatization given in R. Epstein’s \textit{Propositional Logics}. There, the axiom \((A \land (\neg A \land \circ A)) \rightarrow B\) is considered among ten other axioms, where \(\circ A\) is defined as \(\neg (\neg (\circ A \land \circ A))\) for a primitive modal operator \(\circ\). The matrix interpretation of \(\circ A\) in that book coincides with our matrix interpretation for \(\circ A\), as discussed in the next section. This logic also coincides with the logic \(CLuNs\), to be found, for instance, in D. Batens’ “A survey of inconsistency-adaptive logics,” and it has appeared quite often in the literature.

So much for the syntactical part of the logics of formal inconsistency. Semantical interpretations are a complicated issue for paraconsistency in general. The first \(C\)-systems were introduced only in proof-theoretical terms, and only some years later were semi-truth-functional bivalued semantics proposed for their interpretation. Those semantics, however, offer a very weak and debatable “meaning” to paraconsistent logics, and we describe in the next section an attractive alternative semantics called possible-translations semantics.

4. The Rosetta stone analogy

Found by Napoleon’s troops in 1799 near the town of Rashid (Rosetta) in Lower Egypt, the Rosetta Stone is a piece of basalt which after several unclear episodes happens to be found today at the British Museum in London, containing

\(^{12}\) See C. Mortensen, “Paraconsistency and \(C_n\),” and J. Marcos, \textit{Possible-Translations Semantics}.

\(^{13}\) See W. A. Carnielli and J. Marcos, “A taxonomy of \(C\)-systems” and J. Marcos, “8K solutions and semi-solutions to a problem of da Costa”.

inscriptions that were the key to deciphering Egyptian hieroglyphic writing. The deciphering was possible only due to the inscriptions appearing in three forms: hieroglyphic, Demotic, and Greek. By comparing the hieroglyphic and Demotic scripts with the Greek version, and assuming that they contained the same text, the British physicist Thomas Young and the French Egyptologist Jean François Champollion were able to decipher the hieroglyphic and Demotic versions in 1822. Moreover, by further comparing the hieroglyphic text to equivalents of the better known Coptic language they could attach a phonetics to the hieroglyphic writings, which were supposed to be only symbolic. Our semantic approach to paraconsistent logics, known as possible-translations semantics, is in many aspects similar to the deciphering of the Rosetta stone.

In very general terms, a translation from a logic system \( L \) into a logic system \( L' \) is just a language homomorphism that preserves derivability, that is, if \( A \) is provable in \( L \) from premises \( \Gamma \), and \( * \) is a translation from \( L \) into \( L' \), then \( A^* \) should be provable in \( L' \) from premises \( \Gamma^* = \{ B^* : B \in \Gamma \} \), that is, if \( \Gamma \vdash A \) then \( \Gamma^* \vdash A^* \). Several specializations and variations of this notion have been studied in the literature, but this general definition is adequate to our present purposes.

In intuitive terms, the idea is to project a given “hieroglyphic” logic by means of translations of it into simpler (usually many-valued) systems, and combine their respective forcing relations in order to obtain a sound and complete semantical interpretation to the initial complicated system. The simpler systems would thus play the role of Greek and Coptic in the Rosetta stone analogy. We may think of this process as working in two distinct directions: When analyzing a complicated logic in terms of simpler components, we call the process splitting logics; but it is also possible to think of this process in the direction of synthesis, by defining a complex logic starting from simpler ones, and in this case we call the process splicing logics. Possible-translations semantics can be seen as a kind of distributed semantics where the meaning of a sentence in the “hieroglyphic” logic is made clear by way of a suitable combination of all the translations of that sentence into the component logics.

We give here an example of how this kind of semantics can help to give meaning to contradictions. For the basic \( C \)-systems \( bC \) and \( C1 \) the logics playing the role of auxiliary languages in the Rosetta stone analogy will be copies of the three-valued logic pictured below, with truth-values \( T \), \( t \) and \( F \), of which \( T \) and \( F \) are absolute “true” and “false,” while \( t \) can be understood as “provisionally true.”

---

14 See E. A. W. Budge, *The Rosetta Stone*.
15 As in W. A. Carnielli, “Possible-translations semantics for paraconsistent logics” and J. Marcos, *Possible-Translations Semantics*. The analogy is due to J. Marcos, who presented it first in his Master Thesis defense at the University of Campinas, Brazil.
16 See W. A. Carnielli and I. M. L. D’Ottaviano, “Translations between logical systems: a manifesto”.
17 As introduced in W. A. Carnielli and M. E. Coniglio, “A categorial approach to the combination of logics”.
For **Ci**, as we show below, the meanings of $\land$, $\lor$, $\rightarrow$ are fixed, but the meanings of $\neg$ and $\circ$ vary: Each of them will be assigned two distinct interpretations, namely a weak and a strong one. For negation $\neg$, the weak interpretation $\neg_w$ regards the value $t$ as careful truth, and assigns to the negation of $t$ also the value $t$. On the other hand, the strong interpretation $\neg_s$ makes no distinction between $t$ and $T$, and assigns $F$ to the negation of $t$. For $\circ$, the weak interpretation $\circ_w$ forgets the distinction between $t$ and $T$, while the strong interpretation $\circ_s$ recognizes the value $t$ as “provisionally true,” and thus potentially inconsistent.

For the system **Ci** the set of all recursive possible translations to be considered is definable by the following clauses, to be obeyed by any translation $*$ in this set:

**Tr1** For atomic $p$, $p^* = p$.
For atomic $p$, $(\neg p)^* = \neg_w p$.

**Tr2** For non-atomic $A$, either $(\neg A)^* = \neg_s A$ or $(\neg A)^* = \neg_w A$.

**Tr3** For $\#$ any of $\land$, $\lor$, $\neg$, $(A \# B)^* = A^* \# B^*$.

**Tr4** If $(\neg A)^* = \neg_w A^*$, then $(\circ A)^* = \circ_s A$.
If $(\neg A)^* = \neg_s A^*$, then $(\circ A)^* = \circ_w A$.

As an example, a formula of the form $\neg \circ \neg A$ will have eight possible distinct translations, according to the above clauses: If $(\neg \circ \neg A)^* = \neg_w (\neg A)^*$, then $(\neg \circ \neg A)^*$ will be either $\neg_w \circ_s \neg_w A^*$, or $\neg_s \circ_s \neg_w A^*$, or $\neg_w \circ_s \neg_s A^*$, or $\neg_s \circ_w \neg_s A^*$, or $\circ_w \circ_s \neg_s A^*$, or $\circ_s \circ_s \neg_w A^*$, or $\circ_s \circ_w \neg_s A^*$ or $\circ_s \circ_w \neg_s A^*$.

In other words, the syntax will be interpreted in different semantic scenarios, and here of course we have infinitely many distinct translations interpreting the formulas of **Ci** into distinct fragments of the above three-valued logics, according to the choices for the interpretation of the connectives $\land$, $\lor$, $\rightarrow$, $\neg$, and $\circ$.

Possible-translations semantics are a powerful tool for combining logics, complementary to other methods such as fibering. In what concerns the two main directions for combining logics cited before (splitting logics and splicing logics\(^{18}\)) possible-translations semantics initially seem to be more apt for splitting while fibering methods are more apt for splicing.

Possible-translations semantics have already been given for the calculi in the

\(^{18}\) See W. A. Carnielli and M. E. Coniglio, "A categorial approach to the combination of logics".
hierarchy $C_n$ and for a slightly stronger version of $C_n$, offering thus a solution to
the difficult problem of finding good semantics for paraconsistent logics.\textsuperscript{19} It is
also possible to give such semantics for some many-valued logics.\textsuperscript{20} Based on
these semantics, connections to other views on logics like dialogical logic could
also be established, as suggested in Shahid Rahman and Carnielli, “The dialogical
approach to paraconsistency.” A categorial treatment of possible-translations
semantics is presented in W. A. Carnielli and M. E. Coniglio, “A categorial
approach to the combination of logics,” where completeness with respect to
possible-translations semantics is characterized by means of limits of categorial
diagrams.

5. Modeling human and computer reasoning, or why we can’t reason
without contradictions

A good example of why the logic of ordinary language reasoning would have to
abor the Principle of Explosion is the following. Suppose that, in the course of
an investigation, you receive information on some given subject (for example, as a
response by two or more people to “Does Dick live in Arizona?”) in the form of
two sentences A and B of type “yes” or “no.” Now there are exactly three
possibilities (in a two-valued logic) concerning the truth-values of those
sentences: Either they are both true, or both false, or they are contradictory, which
occurs only in case they are distinct. Now it happens that the contradictory
answer is very opportune, because it is the only case in which you’re sure you
received wrong information! Human reasoners profit from this possibility, and
we would not exaggerate in saying that this ability is even essential for survival:
In many cases, signs of danger are recognized in this way. So, reasoning with
contradictories, instead of deriving anything else from them, seems to be an
essential trait of human thinking. By itself, this would be a strong case in favor of
developing a logic where PPS is not accepted as universal.

Interesting applications of paraconsistent logics can be found in the domain
of automated reasoning and knowledge-based systems. We mention two
examples of applications. The first refers to automated proof methods and logic
programming. The resolution method and the tableau method are known to be
equivalent for classical logic, and the programming language PROLOG which is
based on such methods permits computer programs to be written almost
axiomatically, directly in a sort of logic programming language, though for non-
classical logics those methods do not necessarily coincide or have comparable
under inconsistent knowledge” we proposed a signed tableau method for the
paraconsistent calculus $C_1$ of da Costa (and the method is immediately
generalizable to other calculi of the $C_n$ hierarchy) which differs from the classical

\textsuperscript{19} See J. Marcos, Possible-Translations Semantics and W. A. Carnielli, “Possible-translations
semantics for paraconsistent logics”.

\textsuperscript{20} A particular case of possible-translations semantics called “society semantics” has been
studied in W. A. Carnielli and M. Lima-Marques, “Society semantics and multiple-valued logics”.

signed tableaux with respect to the rules for negation (all other notions, like closed signed tableaux, etc, being the same as the classical ones).

Recalling that in $C_1$ the formula $A^0$ (which plays exactly the same role of the formula $\Box A$ in the axiomatization presented in Section 3 above) abbreviates $\top(A \land \neg A)$, the new rules at the propositional level are the following, where $\#$ stands for any of the binary connectives $\land$, $\lor$, $\rightarrow$. As usual in tableau proofs, rules containing the symbol $\vdash$ represent an “or” tree, with branching nodes.

\[
\begin{align*}
T(\neg A) & \vdash F(A^0) \\
F(A) & \vdash F(A^0) \\
F(\neg A) & \vdash T(A) \\
T(A) & \vdash F(\neg A) \\
F((A \# B)^0) & \vdash F(A^0) \rightarrow F(B^0)
\end{align*}
\]

If $\neg A$ is true, then either $A$ is false or it is false that $A$ is consistent, i.e., $A$ is inconsistent/contradictory.

If $\neg A$ is false, then $A$ is true.

If $A \# B$ is inconsistent, then either $A$ or $B$ is inconsistent.

Notice that the rule for $T(\neg A)$ differs from the classical one just by adding an alternative branch $F(A^0)$. The rule for $F((A \# B)^0)$ does not exist in the classical case, however both could be added as redundant rules to the classical tableau rules.

Completeness of these rules for the propositional calculus $C_1$ (and for its first-order version, $C_1^*$ with additional rules) was proven in W. A. Carnielli and M. Lima-Marques, “Reasoning under inconsistent knowledge,” where other examples for automated reasoning were also given.

A simple yet illustrative example is the well-known “Nixon Diamond,” often repeated in the literature. Suppose we have the following statements:

Nixon is a Quaker: $Q(n)$.

If Nixon is a Quaker, then Nixon is a pacifist: $Q(n) \rightarrow P(n)$.

Nixon is not a pacifist: $\neg P(n)$.

These statements are contradictory, assuming that there is some person having all such properties. At this point, a human reasoner would suspect that one of them should be disqualified, but this maneuver would be blocked if all of the claims had equal confidence status. If all of them are to be taken as true, there is no other rational possibility besides some predicate being inexact, or vague, or subject to contradictions. Let’s suppose that it can always be clarified who is a Quaker and who is not. Then the only candidate for inexactness or possible contradictoriness is “pacifist.” This is exactly the conclusion given by our system\textsuperscript{21}: We can run a tableau for the set of propositions $S = \{ T(Q(n)),$

\textsuperscript{21} In W. A. Carnielli and M. Lima-Marques, “Reasoning under inconsistent knowledge.”
T(Q(n) → P(n)), T(¬P(n))}. The system concludes, instead of becoming blocked, that F(P(n)°), that is, P is contradictory.

The second example is for designing an implementation of databases that are robust enough to work in the presence of contradictions. There are several ways in which a database can be contradictory. Different users having equal access to some given database may introduce new claims, and even new rules or constraints, which, despite being consistent from the point of view of each user, can still be globally contradictory. Traditional databases may detect contradictory information and then start a complicated, and computationally extremely inefficient, procedure for “restoring consistency,” but by no means can they afford modification of constraints, for reasons explained below. It is thus natural to embed databases in logical environments that permit reasoning with contradictory information, while maintaining all other desirable features of traditional logic, such as reasoning with the law of excluded middle, reasoning by cases, and reasoning by means of quantifiers.

In general, information stored in databases must be checked to verify some previous conditions (called integrity constraints) in order to be safely integrated in the database. Integrity constraints are expressed by (fixed) first-order sentences. For example, a database storing information about books may contain the requirement that no book in the collection can have more than one title, a condition that could be expressed by the following first-order formula, where Title(x, y) means that y is a string that is the title of book x:

\[ \forall x \forall y \forall z \ ( \text{Title}(x, y) \land \text{Title}(x, z) \rightarrow y = z ) \]

Updates in traditional databases are only performed if the new database would satisfy the integrity constraints; if not, the database maintains its previous state. So, in a traditional database system, an imperative control never allows contradictory information to be considered.

The situation would be worse for traditional databases if integrity constraints themselves could be changed in time, instead of remaining fixed forever. Such evolving databases, which we call evolutionary databases, seem very interesting for the domain of artificial reasoning. The fact that, traditionally, integrity constraints are defined by the database designer and remain fixed for the users during the database lifetime is a severe limitation on databases, due solely to the logical foundations of classical database theory.

In W. A. Carnielli, J. Marcos, and S. de Amo, “Formal inconsistency and evolutionary databases” we have introduced new logics that axiomatize a formal representation of inconsistency in classical logic, starting from a purely semantical standpoint. Though in that paper we take inconsistency to be equivalent to contradictoriness, that is not necessary according to the observations made in Section 3. That simplification does not affect the treatment of information. Starting from an intuitive semantic account of what contradictory (or inconsistent) data should be, and taking into consideration some basic requirements, we provided two distinct sound and complete axiomatics for such
semantics, LFI1 and LFI2, as well as for their first-order extensions, LFI1* and LFI2*, depending on which additional requirements are considered.

These two formal systems are examples of Logics of Formal Inconsistency (LFI) and form part of a much larger family of maximal paraconsistent three-valued logics. It is important to note that LFI1* and LFI2* are proper subsystems of classical logic extended with a consistency operator, and they entail thus, in intuitive terms, fewer theorems than this extended form of classical logic. We have shown, however, that they can codify any classical or paraconsistent reasoning, because there exist grammatically faithful conservative translations from classical and paraconsistent first-order logics into LFI1* and LFI2*.23

We repeat here an example of an evolutionary database considered in W. A. Carnielli, J. Marcos, and S. de Amo, “Formal inconsistency and evolutionary databases.” Suppose that a claim P is proposed by a certain source. It may enter the database either with the token $\text{\checkmark}$ or the token $\text{\times}$ appended to it. In case not-P is proposed, the claim P enters with the token $\text{\times}$ or does not enter at all. In case that we know nothing about P, nothing is added to the database. As a consequence, in case P and not-P are simultaneously proposed (for instance, by different sources), then P enters the database with the token $\text{\times}$.

As a concrete example, suppose we have a database schema DS containing three relations: Author(Name, Country), Title(Book, String), and Translated(Book, Language). Suppose also that two different sources, I and II, provide information to our database, telling us:

**Source I**

1. Joaquim Maria Machado de Assis was born in Brazil.
2. Gabriel García Marquez was born in Colombia.
3. Machado de Assis is author of *Dom Casmurro*.
4. *Dom Casmurro* has not been translated into Polish.

**Source II**

5. Gabriel García Marquez was not born in South America.
6. Gabriel García Marquez is author of *One Hundred Years of Solitude*.
7. *One Hundred Years of Solitude* has been translated into Polish.

Now claim (4) is negative, and may be stored in DS with a $\times$ or not stored at all. On the other hand, claims (2) and (5) are contradictory, given certain already stored knowledge about geography, and in this case those claims are added with the token $\times$ appended to it. The remaining positive information may be added liberally either having $\text{\checkmark}$ or $\text{\times}$ as a suffix.

So we have already at least one piece of inconsistent information stored in DS. Besides that one, another one may appear if, for instance, Source I adds a new constraint asserting that “No South American author has ever been translated into Polish.” After DS has been updated taking this new constraint into consider-

---

23 See W. A. Carnielli, J. Marcos, and S. de Amo, “Formal inconsistency and evolutionary databases”.
ation, the relation "Translated" will also contain contradictory information. While traditional databases cannot support this situation, our model permits us to reason with these contradictions, even taking advantage of this controversy to get better knowledge about the sources, much in the same way a human reasoner would do.

6. Closure

In this paper we have tried to defend the idea that reasoning with contradictions is not only useful but perfectly well-founded from the logico-mathematical standpoint. We have also discussed some underlying questions of paraconsistency, indicating some applications of it to human and automated reasoning and to database models. We have shown how several axiomatic systems of paraconsistent logics can be formulated, and how possible-translations semantics can be assigned to some of them. We also have tried to show how those interpretations are intuitively clear and philosophically appealing, despite some common objections that could be raised.

Often philosophers make predictions about what is mathematically possible. Some writers, in their efforts to find definitive arguments in favor of their own beliefs, try to blame logic, or more widely, mathematics as a whole, risking their reputation on what could or could not be done in the domain of mathematical possibilities. Some "throw the baby out with the bathwater," prematurely dismissing the possibility of constructing logic systems which are robust enough to carry a good portion of classical reasoning. Karl Popper, while objecting against dialectics, acknowledges the intrinsic interest of contradictions:

Dialecticians say that contradictions are fruitful, or fertile, or productive of progress, and we have admitted that this is, in a sense, true.\(^{24}\)

However, he cautiously warns the reader that from a pair of contradictory premises any conclusion may be deduced in classical logic, and he asserts:

The question may be raised whether this situation holds good in any system of logic, or whether we can construct a system of logic in which contradictory statements do not entail every statement. I have gone into this question, and the answer is that such a system can be constructed. The system turns out, however, to be an extremely weak system. Very few of the ordinary rules of inference are left, not even the modus ponens which says that from a statement of the form 'If \(p\) then \(q\)' together with \(p\) we can infer \(q\). In my opinion, such a system is of no use for drawing inferences, although it may perhaps have some interest for those who are specially interested in the construction of formal systems as such.

Popper's remarks can be seen as a good defense of our position: He concludes not only that contradictions can be seen positively, but also that other systems of logic could be constructed so that PPS, at least as taken for granted in classical logic, would not be valid. Unfortunately, he stops too soon, blocked by

\(^{24}\) Conjectures and Refutations: The Growth of Scientific Knowledge, p. 316.
his own misfortune. But the fact that a great philosopher did not succeed in the mathematical enterprise of constructing a system of this sort, powerful enough to incorporate most of ordinary rules of inference including *modus ponens*, certainly does not mean that this task is impossible. Several mathematicians and philosophers in the course of history have experienced similar difficulties imagining non-Euclidean geometries, complex numbers and the like, but fortunately science does not survive on the imagination (or lack of it) of one individual.

By taking seriously the task of constructing logics for formal inconsistency we may also be helping to understand other philosophical questions: according to S. Knuuttila and A. I. Lehtinen in “Change and contradiction: a fourteenth-century controversy,” the possibility of accepting that contradictory sentences could be true at the same instant of time was already considered (for theological arguments) almost seven centuries ago. Instead of blaming logic or mathematics and risking sibylline prophecies on what cannot be done, it is often better to investigate what can be achieved—even if what is achieved in the end are negative results, as the mathematical impossibilities of solving the classical Greek problems of elementary geometry, the theorems of Gödel, the indecidability of first-order logic and many other results familiar to contemporary logic.

**Bibliography**

Batens, D.


Budge, E. A. W.


Carnielli, W. A.


Carnielli, W. A. and M. E. Coniglio


Carnielli, W. A., and I. M. L. D’Ottaviano


Carnielli, W. A. and M. Lima-Marques


---

25 His system was introduced in K. R. Popper, “On the theory of deduction. I. Derivation and its generalizations” and “On the theory of deduction. II. The definitions of classical and intuitionist negation”. 
Carnielli, W. A. and M. Lima-Marques

Carnielli, W. A. and J. Marcos

Carnielli, W. A. and J. Marcos

Carnielli, W. A., J. Marcos, and S. de Amo

da Costa N. C. A.


da Costa, N. C. A., J.-Y. Béziau, and O. Bueno

Epstein, R. L.

Kant, I.
???? Critique of Pure Reason. Translated by N. Kemp-Smith (second edition, 1787), Macmillan Press.

Knuuttila, S. and A. I. Lehtinen

Łukasiewicz, J.

Marcos, J.
Mortensen, C.

O’Connor, J. and E. Robertson
<http://www-history.mcs.st-andrews.ac.uk/history/index.html>

Popper, K. R.

Rahman, Shahid and Walter A. Carnielli

Rose, A.

Sette, A. M.

Thagard, Paul and Cameron Shelley

von Wright, G. H.
Pedagogy and Examples

This section publishes articles that present ideas or examples for the teaching of critical thinking and logic, or examples that either illustrate or call into question some aspect of current theories of the foundations of reasoning.

Knowing the Exact Proportions is Not Enough

Stanislaw Krajewski

To describe a population we give the percentage of its members that have a certain property. This is a convenient method for dealing with large groups. This is also the way to concentrate on the properties we’re interested in. For instance, one may be interested to know how many men and how many women visit a museum in order to know whether the exhibition is equally interesting to both genders. The exact numbers are not that important; we want to know the percentages.

We are all familiar with the need to be very careful when dealing with samples. To infer a property of the whole population, from which the sample is drawn, we must be reasonably sure that the sample is representative and not too small. Elaborate theories of statistics tell us how to choose samples. To know how many visitors to a museum are male we study an appropriate sample and infer the percentage with a probability that can be estimated. It may seem that exact knowledge of the percentage guarantees safe inferences. It turns out that we can err even then. Thus, we may have exact information concerning the ratio of men and women among the visitors, but we can misunderstand the significance of the ratio.

A well known joke explains the danger of statistics: When my neighbor beats his wife every day and I never do, statistically we beat our wives every second day. It could seem that this is so trivial that a little care is sufficient to overcome the danger. Let us see how most of us could fall into the same trap.*

Example 1

In a prestigious college, The School of Marketing and Merchandising, only those candidates are admitted who have passed an entrance examination. A journalist of the local newspaper Valuable News became interested in the results of the last exams, and she found out that 36% of women candidates and 64% of men candidates were admitted. “Therefore,” she said to herself, “there has been discrimination here!” Things looked obvious, especially in view of the fact that an equal number of men and women took part in the exams, namely, 500 women

* I first learned about an example of this kind from Dr. Lech Kubik, Warsaw.

© 2001 Stanislaw Krajewski.
(only 180 of whom were admitted) and 500 men (320 admitted). The journalist was ready to write a story about the lack of equal opportunities; she felt that she had enough evidence to make a strong case against the policy of the college. Before writing the final draft she went to interview the president. The president felt embarrassed, and a little threatened, so she decided immediately to check the reason behind the inequality in admissions. There were two separate departments in the college, the Department of Marketing (let us call it Department I), and Department of Merchandising (let’s call it Department II), with separate examinations. The president hoped that only one of the departments had been guilty, but the disparity, 36% vs. 64%, was so big that she also felt she had to prepare herself for the possibility that neither of the departments had provided fair examinations. This seemed highly probable because an equal number of candidates, 500, tried their luck in each of the departments. The president called the deans of the departments to report their respective percentages. To her amazement each of them maintained that in his department the chance to be admitted were actually better for women. Both mentioned a 5% margin. Actually, in Department I, from among their candidates, 25% of the women and 20% of the men were admitted. In Department II, 80% of the women and 75% of the men were admitted.

The president didn’t believe her deans, and the journalist was sure that they tried to cheat her. “How come,” she said, “it isn’t obvious that whenever relatively more men than women are admitted, then if not in both departments, at least in one of them it must have been easier for men to be admitted than for women?” The president had no answer to that. She asked the deans to deliver the complete data at once. Only the secretary, who was witnessing the conversation when she brought coffee to the president’s room, remarked humbly that if more men were admitted, then more men must have been admitted in at least one of the departments, but perhaps the chance of admission was something else. This looked like casuistry to both her boss and the journalist. The latter said, “If the chance to be admitted is higher for women at both Department I and Department II, and the college has just these two departments, the chance must be higher for all women candidates to the college.” “This looks almost like a syllogism to me,” admitted the president. Soon, however, they got the Table with all details.

<table>
<thead>
<tr>
<th>Candidates’ gender</th>
<th>Department I</th>
<th>Department II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitted</td>
<td>400 f</td>
<td>100 f</td>
</tr>
<tr>
<td></td>
<td>100 m</td>
<td>400 m</td>
</tr>
<tr>
<td>Percentages</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>College globally</td>
<td>500 Candidates</td>
<td></td>
</tr>
<tr>
<td>Admitted</td>
<td>500 f</td>
<td>500 m</td>
</tr>
<tr>
<td></td>
<td>36%</td>
<td>64%</td>
</tr>
</tbody>
</table>
Knowing the Exact Proportions is Not Enough

The president and the journalist studied the table briefly and did feel humbled, so rather than asking the secretary they brought Dr. E, the school’s Critical Thinking instructor, to analyze the situation.

Example 2
Everything is exactly the same. Only when they are waiting for the exact data, another journalist comes. He checked that in both departments women had higher chances to be admitted. He was puzzled because he had always thought that men were much better in marketing and in merchandising. He suspected too much of a political correctness at work, and contemplated an article for the local weekly, Invaluable News. When confronted with the table with all the details he felt as bad as did the others.

It was easy for Dr. E to describe the table. Everyone can see that the uneven distribution of men and women among the candidates to the departments is the key fact. Had the distribution been even, the peculiar situation could not have arisen. More women were admitted to Department I, but many more men are admitted to Department II, so in total there are more men. Still, the percentage is always on the side of women because many more women applied to Department I, where only 120 were admitted (while in Department II as many as 380 were admitted). Clearly, women concentrated in the more difficult of the two departments. These remarks of Dr. E seem to have explained the situation well enough.

The journalist (as well as her colleague of Example 2) was, however, unsatisfied. She did not feel she knew what was happening. Was there discrimination or not? If so, in what direction? She did not know how to ask. Fortunately, the secretary helped to ask the right question again: “Why were more women concentrated in the difficult department?”

The question seems hard to answer, and it is surely hopeless if all we know is the Table. Now, one important issue is as follows: What is the reason for the disparity in the number of students admitted (120 to 380)? One possibility (Case 1) is that the number of admissions in each department had been determined in advance. (They were not equal, because, for example, Department I needed special lab classes and could admit no more than 120.) Another possibility (Case 2) is that there was no prearranged numbers, but that they just took all those (top 500) who got good enough grades; it so happened that the candidates to Department II got so much better grades. How does our principal problem, that concerning the gender discrimination, look in both cases?

Case 1. If it had been prearranged that only 120 would be admitted to Department I, why were so many women going there? It was clear that the exams would be more difficult than in the other department. Was it a coincidence? Were women more ambitious? Misinformed? If they chose Department I consciously, there was no discrimination. Were women more attracted because of a specific character of Department I? Did they know that no more than 120 would succeed? If so then, again, there was no discrimination.
It could conceivably have happened that the decision to admit only 120 to Department I was made only after it had been learnt that women were dominating among the candidates to that department. If *that* were the reason for the decision, that is, the fact of women being the majority, then we have a blatant case of discrimination. If, however, the decision was made on other grounds, without taking into account the relative number of women, then there was no discrimination.

Case 2. If the departments admitted just those who got better grades, the problem is why the women came out inferior. We can’t exclude the possibility that in that specific place and time men were simply better, on the average. However, before admitting this, we must ask one important question: Were the exams in the two departments of the same degree of difficulty?

If not, was this connected to the fact that there were so many women trying to get to Department I, or not? If yes, then there may have been discrimination; yet, before saying this our journalist should find out the basis of the decision to make the exams to Department I harder. If not, that is, if the exams for Department I were made more difficult, but this had no connection to the relative number of women, then perhaps women just had bad luck (or rather, relatively more women than men turned out to have bad luck). This is rather clear if among those who prepared exams there had been no awareness that women were over-represented among candidates to Department I.

If, on the other hand, the exams were the same all over the college there seems to have been no discrimination. Unless, that is, some subtler reasons intervened.

There may be other factors that could affect the outcome of the exams. For example, the questions could have been posed by a computer, whose software contained an unintended mistake and gave harder questions to those with longer first names. It so happened that women had, on average, longer first names than did men. Then there would be an unfair discrimination, although not intended by anyone, so it would be hard to assess guilt.

We just need to imagine the possibilities. It seems plausible that there can be, theoretically, a large, practically unlimited, number of them.

It should be clear by now that we can easily add additional complexity to the picture. Let us distinguish another category among the candidates, not just according to gender. The new group, call it A, would be, typically, a minority. For example: Athletes, or Asians, or Arabs, or Aristocrats. Or, say, Jews, or Cyclists, or just the opposite. (By the way, we can easily imagine that Asians had, on average, shorter names and then they would be put in a privileged position by that unintended error in software.) Now we are ready for the third example, which is a slight extension of the initial one.

**Example 3** Everything is as before, only when they were contemplating the Table and waiting for Dr. E, another journalist came, a columnist for *Valuable News*. She was worried because it has turned out that A-candidates were
discriminated against. Namely, only 30% of them were admitted while 50% of all candidates were. And, of course it turned out again that in no department had there been a problem. In Department II there were simply no A-candidates at all. In Department I, 30% of A-candidates were admitted and only 20% of non-A-candidates. The table for Department I was the following (‘a’ stands for A-ness, ‘n’ for non-A-ness):

<table>
<thead>
<tr>
<th>Candidates’ A-ness</th>
<th>200 a</th>
<th>300 n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitted</td>
<td>60 a</td>
<td>60 n</td>
</tr>
<tr>
<td>Percentages</td>
<td>30%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Combined with gender, the table for Department I was:

<table>
<thead>
<tr>
<th>Candidates’ A-ness</th>
<th>400 f = 100 a + 300 n</th>
<th>100 m = 100 a + 0 n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitted</td>
<td>100 f = 40 a + 60 n</td>
<td>20 m = 20 a + 0 n</td>
</tr>
<tr>
<td>Percentages</td>
<td>25% 40% 20%</td>
<td>20% 20%</td>
</tr>
</tbody>
</table>

It is also apparent that A-women candidates fared better than the non-A-women. The A-candidates had bad luck to try to get to the department with harder examinations. Or, was it just bad luck? What was the mechanism of the examinations? Were the A-women privileged? If the exam was more difficult for Department I, and within it they were not treated better than others, they were actually the best scoring group of all the specific subgroups of candidates considered so far.

So: Was the A-group at a disadvantage? Or was it the non-A-group? Was any group discriminated against? If so, by whom? Not enough was known to offer the readers of Valuable News a reasonable conclusion: The percentages are so-and-so, therefore . . .

Our overall conclusion is that there may be many circumstances, and we must not assume that we know all of them before we analyze the situation. Knowing the exact percentages is not sufficient for drawing conclusions about the reasons for differences or disparities among them. Sounds discouraging? I would certainly grant that in most everyday situations we don’t have to be so penetrating. We assume that the factors other than those obviously affecting the situation are irrelevant. But if we want to accuse we should do better. The journalist’s task is not just to learn the percentages and infer the existence of discrimination or the lack of it. Having come across the percentages, the journalist still has to do the real work.

Our journalist from Valuable News has learnt a valuable lesson. She had a strange dream at night. Various percent figures were dancing and singing to a haunting tune with bizarre words: “It is so hard, so hard for us all, to know the right way of saying ‘therefore’.” (Once she heard “sixth” rather than “right.”)
Discussion: On the Ethics of Killing Dogs

The following concerns whether an acceptable solution to the problem of stray dogs in a city is to kill them. That solution had been proposed in Bucharest, Romania, where the problem of stray dogs was acute, as we learned during the ARF meeting. But it had not been implemented because of an outcry by animal rights activists. A few weeks after the meeting, the following article appeared in newspapers:

At least three dozen dogs are being killed in Tehran every day in an effort to rid Iran’s capital city of canines. The Qods daily reported that at least 1,000 dogs, which are regarded as impure in the Islam belief system, have been killed in the last month alone. The figure is a 50 percent increase over the previous month. There are no animal shelters in the country and roving canines are frequently seen going through sidewalk garbage can looking for scraps. While many affluent city residents secretly keep dogs as pets, the animals are regularly denounced by the country’s ruling clergy.

Richard L. Epstein suggested that it was a sin to kill dogs. This is the discussion that ensued.

I don’t mean to step on anybody’s paws, but . . . if the dogs are digging scraps in garbage cans, they probably aren’t enjoying life very much. While dogs are keen olfactory, and can remember signs of pleasant and painful stimuli, and can thus anticipate the very near future, I am unaware of evidence that suggests they form hopes for the far future. If that is so, they don’t dread or grieve loss of life, and they don’t have plans beyond the next meal or mounting that will be frustrated by early demise. On the other hand, as we know first hand, when not cared for as pets, they can be dangerous to people, who can suffer anxiety about this possible source of danger to themselves and their children over long periods of time. On balance, therefore, it seems morally acceptable to end the lives of dogs that are not being cared for as pets, especially in cities where humans are numerous and food is scarce. (There is, of course, no argument for doing this in any fashion other than the most painless available. Nor for clerics to indulge in insults or condemnations of Doghood as such, or of individuals that have formed particular human attachments.)

William S. Robinson, August 7, 2000

Very interesting argument indeed, except at least one premise is dubious. Ever see a dog waiting for his master at the door, or at a gate, hours after the master left through it? Mine waits for me at the gate, (apparently) hoping I’ll come home and play with him when I’m away for a whole day in Albuquerque. Now I’m not

claiming that this means he’s planning ahead, but it gives us good reason to doubt your premise that he doesn’t plan.

Richard L. Epstein, Aug 8, 2000

I have enjoyed the ethical exchange between Dick and Bill on dogs and Doghood. But I have a question about Bill’s argument, and the response. Is that argument using a premise like “Animals, human or non-human, are only to be protected if they cause no substantial harm to the well-being of humans, and they plan for the future.” Or is it, as I suspect, assuming a more complex premise about the nature and scope of the plans (and, to avoid the obvious reply that young infants are implicated, something about at least having the potential for such planning)? I suspect that the line:

[Dogs don’t] form hopes for the far future. If that is so, they don’t dread or grieve loss of life . . .

is more important to the argument than the line:

They don’t have plans beyond the next meal or mounting that will be frustrated by early demise.

Anyway, this kind of uncertainty about the structure and content of the argument is something we are all familiar with, except we tend to forget it when we do analysis of arguments with students (at least my tutors do). Analysing arguments is an art.

Fred Kroon, Aug 14, 2000

I agree that I didn’t give canine cognitive abilities their full due in my argument. They may pine for the return of the Great Arf. (I think this is actually included under remembering signs of pleasant and painful stimuli, but I should have made that explicit, and the time frame is longer than I implied.) I’m also convinced that they respond to signals in play in a way that must be considered symbolic. (Where did I learn this? Allen & Bekoff, *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. I recommend the book, for those who are interested in abilities of nonhuman animals. (If you aren’t already, you might become so by reading this book.) No time to read a book on this? You can look at my review of it in *Psyche*, at <http://psyche.cs.monash.au/v5/psyche-5-26-robinson.html>.)

However, I don’t think these omissions undercut the essential point of the argument I was making. “And what was that argument?” asks Fred.—Two preliminaries: (a) As Fred can be taken to suggest, I don’t think I was appealing to the first of the possible premises he mentions. (b) I agree that the absence of dread and grief is more important than the absence of plans. Plans become relevant in at least two (possible) ways: (i) One might dread failure of one’s plans by premature interruption (which, of course, doesn’t arise if one doesn’t have plans). And (ii) There’s a value that might be called “aesthetic,” but which might be a value even for plans that aren’t yours and that you don’t like. Suppose
someone is writing a book that you're sure defends a false view. You might still regard it as a bad thing if the work never comes to completion because the author dies in a crash—that is, you'd think it was awful in a way that is something additional to the regret at losing the author. Calling this kind of value "aesthetic" may risk making it sound like something not too serious, but I don't take it too lightly.—But yes, the dread and the grief weigh more.

So, what's the argument? As I was aiming for brevity, I spoke as a utilitarian. But I don't really premise that utilitarianism is true. I know about justice and rights issues against utilitarianism, and I'll return a bit to them. But the main drift is that those problems for utilitarianism don't overturn a utilitarian view in this case. The main drift, that is, is that (a) Canine abilities (to plan) and susceptibilities (to dread, grieve) are quite limited, (b) They're not having much pleasure, while (c) Human fears about attack are substantial and (d) There are actual pains due to actual attacks. The disvalues on the human side seem to me to far outweigh the values on the canine side—not because canines are a different species, but just because of the contingent facts about the capabilities and conditions of each species.

If I thought that word would get around among the dogs, and they'd be living in fear of their lives, my view would be different. As it is, I think conspiracy theories are beyond their imagination. I think there would be some "Where's Fido got to?" thoughts (well, actually, images of Fido accompanied by vague unease), but when it comes to "imagin(ing) the possibilities," I suppose the dogs are going to be quite limited.

But don't they have a right to life? This question opens a large topic; where do rights come from anyway? No, you're not going to get a treatise! At most a plausibility consideration. Namely, if they had such a right, then it would be unjust for us to violate it. On many views that would imply that (at least) it would be unfair to violate it. But putting the matter this way returns us to the kinds of factors already mentioned. For now we can ask: Wouldn't it be unfair to us to demand that we go out of our way to preserve the lives of dogs, given that there are the differences in abilities and susceptibilities already noted? It can't be fair to demand equal treatment for dogs and humans when they are so unequal in relevant respects. (I haven't argued that the respects are relevant, but I expect that view to recommend itself to you without argument.)

Now, this shows at most a reduced right to life (compared with humans), and so there'll be a question about how much the reduction is. But I think the argument will proceed along the lines above, i.e., the same factors I've identified will be the ones agreed to be relevant, and the large disparity will have the effect of reducing the right to life claim to a very small weight.

No one is going to let me off the hook without my saying something about infants. Imagine a species that's just like infants, except they never progress. Such a species would have less ability than dogs, and less claim on our consideration. So, yes, infants don't make it into a class we ought to protect by virtue of their actual abilities. But I don't see any attraction in the idea that we
have to bring about potentialities, so I don’t think it is mere potentiality that can justify protection for infants. I think the connection goes through parental love and correlative parental fear. Most parents want strong laws of protection for their children (to put it mildly). Of course, they expect them to grow up. Attitudes toward children would be quite different if people thought of them as never growing up. But it’s the attitudes and fears, and not the mere potentiality, that justify the protection.

This view leads to the question of what to say about cases where a parent doesn’t care about his or her infant. Well, it’s not possible to maintain respect for law while making an exception of an infant’s own parent. But, morally speaking, I don’t see infanticide as so awful, IF (it’s a big if) it’s not going to result in a terrible sense of loss to one of the parents. (This would usually mean that it’s the parents, or at least one parent, who did the infanticide.)

This may sound a little stark at first sight, but I think it’s actually close to most people’s sensibilities. Not many years ago there were two cases of infanticide within about a year of each other, both in East Coast states of the US. These were cases in which no one (except the father, in one of the cases) knew that the mothers were pregnant. (Amazing, but true.) One of them delivered alone in a campground, then abandoned the baby. The other went to a motel with her boyfriend (the father) and the two cooperated in disposing of the body in a dumpster. Of course, they were identified, prosecuted, and convicted. (I can’t say on exactly what charge—it could have been manslaughter instead of murder, and murder in most US states comes in several degrees.) The point here is that the sentences were on the order of two or three years. I take this to reflect a kind of official judgment that infanticide is a considerably less serious offense than other murders. (There were also many expressions of pity for the perpetrators, who must have been wildly estranged from most social goods.)

The point of these cases is, of course, not to argue that the courts’ judgments were right. I only mean to deflect an objection to the effect that my ruminations lead to a stance that severely diverges from everybody’s actual moral sensibilities.

Well, this is more than I said before, but obviously not enough! What would be enough? At least a book! Probably two or three. But (as you may be glad to hear) this is all I’m going to write about it today.

Arf!

William S. Robinson, August 16, 2000
Essay Prize on the History and Philosophy of Computability
An annual prize of the Advanced Reasoning Forum for US $250.

The inaugural prize in 2001 was awarded to:

Mr. Giorgos Barbalias
University of Thessaloniki, Greece and University of Leeds, England
“Self-reference and paradoxes”
Faculty Sponsor H. Dugald Macpherson

Terms of the Prize:
1. CLOSING DATE:
   January 15 of each year.

2. ELIGIBILITY
   Any student, graduate or undergraduate, currently registered or who has graduated no
   earlier than May of the preceding year.

3. TOPIC
   Any subject in the history and/or philosophy of computability and/or the foundations of
   mathematics. The topic should fall broadly within the range of topics discussed in
   *Computability* by Richard L. Epstein and Walter A. Carnielli (2nd edition, Wadsworth,
   2000) and the accompanying timeline. Strictly mathematical papers are not acceptable,
   nor are papers whose main emphasis is the relation of mind and machine. New theories
   or models of computation may be discussed.

4. LENGTH and FORMAT of ESSAY
   The essay must not exceed 10 double-spaced pages in 12 point font. The essay must be
   in English. Footnotes should appear on the page of the text to which they apply.

5. SPONSOR of ESSAY
   The essay must be sponsored by a faculty member of the school of the student. No
   faculty member may sponsor more than two students in any one year.

6. METHOD of SUBMITTING ESSAY
   The essay must be submitted with a cover letter with the author’s name, address, and
   name, and university address of the faculty sponsor, signed by both the student and
   sponsor. No identification of the author should appear in the body of the essay. Three
   copies of the essay should be submitted to:

   Richard L. Epstein
   P. O. Box 217
   Socorro, NM  87801  USA

   If a reply is requested that the essay has been received, a enclose a self-addressed
   stamped postcard or an e-mail address for response must be included.

7. AWARD of PRIZE
   The essays will be judged by the members of the Advanced Reasoning Forum, and the
   prize will be announced by April 15. A certificate and the prize money will be sent to the
   student’s faculty sponsor for presentation to the student at his or her school.
Members

Peter Adams
Philosophy Editor
Wadsworth Publishing Co.
10 Davis Drive
Belmont, CA 94002 –3098 USA
peter.adams@wadsworth.com

Walter A. Carnielli
Centre for Logic, Epistemology, and the History of Science
University of Campinas
C. P. 6133
13083-970 Campinas, SP, Brazil
carniell@cle.unicamp.br

Mircea Dumitru
Dean of the Faculty of Philosophy
University of Bucharest
Spl. Independentei 204, sect. 6
Bucharest, Romania
dumitru@fil.unibuc.ro

Richard L. Epstein
Advanced Reasoning Forum
P. O. Box 217
Socorro, NM 87801 USA
rle@AdvancedReasoningForum.org

Stanisław Krajewski
Department of Philosophy
University of Warsaw
Krakowskie Przedmiescie 3
Warsaw, Poland
stankrajewski@poczta.gazeta.pl

Fred Kroon
Department of Philosophy
University of Auckland
Private Bag 92019
Auckland, New Zealand
f.kroon@auckland.ac.nz

Desidério Murcho
Department of Philosophy
King’s College London
Strand, London WC2R 2LS
United Kingdom
desiderio.murcho@kcl.ac.uk

Alex Raffi
1142 Sierra Laurel Ct.
Henderson, NV 89014 USA
landofnott@aol.com

William S. Robinson
Department of Philosophy
Iowa State University
Ames, Iowa 50011 USA
wsrob@iastate.edu

Student Member

Célia Teixeira
21 A Camden Road
London E17 7NF, United Kingdom
celia.teixeira@btinternet.com

Corresponding Members

Benson Mates
Philosophy Department
University of California, Berkeley
Berkeley, California 94720-2390 USA

Peter Eggenberger
365 Perkins St. #101
Oakland, California 94610 USA
PeterEggenberger@msn.com

Piergiorgio Odifreddi
via Superga 6
10099 San Mauro, Torinese Italy
piergior@dm.unito.it
Additional Contributors
for this volume

Sorin Costreie
Department of Philosophy
Talbot College
The University of Western Ontario
London, Ontario N6A 3K7 Canada

João Marcos
RUG, Ghent, Belgium and
IFCH, University of Campinas
13083-970 Campinas, SP Brazil
vegetal@cle.unicamp.br

Copyeditor of the Bulletin of Advanced
Reasoning and Knowledge

Carolyn Kernberger
P. O. Box 760
Socorro, NM 87801 USA
cakernb@aol.com