Why *Reductio ad Absurdum* arguments work

PHI 154 (Eliot) Fall 2022

Negation Introduction (¬I), Negation Elimination (¬E), Indirect Proof (IP), and Explosion (X) are powerful new rules for natural deduction. They are powerful because they offer an entirely new strategy for proofs. Even proofs we have been doing without them can be done with these new rules, and sometimes faster. Using them models an argument strategy which is at least as old as Socrates, who used it often. It is called *Reductio ad Absurdum*, or "reduction to absurdity."

The reasoning at the core of every *Reductio ad Absurdum* argument is this: if you can derive a contradiction from some claim, then you have shown that that claim is False.

For instance, when Socrates uses *Reductio ad Absurdum*, he typically derives a contradiction between his interlocutor's claim and one of the interlocutor's other beliefs. Socrates asks Hippias what makes things beautiful. Apparently without thinking carefully, Hippias replies "covering them in gold." Socrates asks if Hippias thinks a beautiful woman is beautiful, and of course Hippias agrees. Socrates asks if covering a beautiful woman in gold makes her more beautiful, and Hippias acknowledges that it does not. But since Hippias's first, hastily-expressed claim entails that for *anything*, covering it in gold makes it more beautiful, he must accept that covering a beautiful woman in gold makes her more beautiful. He is therefore committed to a contradiction: that it *does and does not* make her beautiful. Since contradictions cannot be True, he is committed to a False belief. He can avoid holding this False belief by allowing that his original claim was False.

Learning how to find and expose contradictions yields a useful intellectual skill, and using the negation rules on symbolized arguments should help you internalize that skill.

But moreoever, what makes *Reductio ad Absurdum* arguments so powerful is that they can be used to prove the conclusion of *any* valid argument, even when the given premises do not entail a contradiction.

If you are trying to prove a negated sentence (e.g., $\neg P$)—whether it is the conclusion of the argument or just a sentence you need to prove along the way—you can *always* assume the unnegated version of that sentence (e.g., P), derive *any* contradiction, flag the contradiction with a contradiction symbol ($\bot$) using ¬E, discharge the assumption, and conclude the negated sentence (e.g., $\neg P$). Your justification is the entire subproof and the rule ¬I (as shown top right). How you reach the contradiction is not important, other than that you follow the rules of TFL! Similarly, if you're trying to prove an un-

1	P	
2	⋮	
3	Q	
4	$\neg Q$	
5	$\bot$	¬E 3,4
6	$\neg P$	¬I 1–5

In conversation, philosophers sometimes abbreviate the Latin name and call an argument of this form a "reductio" for short, as in "Doesn't that observation offer an easy reductio of his claim?"

This occurs in Plato's dialogue *Hippias Major* which is about what's *kalon*, or "fine." Here I am translating *kalon* as "beautiful."

This unsettling image was explored in detail by Guy Hamilton (1964), `http://www.imdb.com/title/tt0058150/`

This strategy is called "indirect proof," as opposed to direct proof, which is just the usual strategy not involving contradictions.

1	$\neg P$	
2	⋮	
3	Q	
4	$\neg Q$	
5	$\bot$	¬E 3,4
6	P	IP 1–5

negated sentence (e.g., P), assume its negation (e.g., $\neg P$), and follow the same strategy, seeking a contradiction, then using the rule IP.

Happily, this strategy often makes proofs shorter. But not always. Sometimes they're longer. Yet, it *always* works. So if you don't see a direct route to some conclusion or subgoal, it may be worth a try. But *why* are ¬I and IP acceptable? And why do they always work?

How the negation rules work

You may have noticed the resemblance between ¬I/IP and →I, in that both rules use subproofs. That similarity reveals why ¬I is justified. When you use ¬I, you are essentially proving a conditional sentence of the form $P \rightarrow \bot$. We don't actually write that conditional, but look at the subderivation and you'll see that you *could* if you wanted to.

Now think about that conditional sentence with the form $P \rightarrow \bot$. Its consequent is a contradiction. A contradiction is simply, by definition, a sentence which is never True, for logical reasons. So, we know that it is a True conditional sentence with a False consequent.

Now what does that tell us? Look at the truth table for the conditional. Remember that the rows of the truth table describe cases or scenarios. Which case are we in? We're in a case where the consequent of the conditional is False. That focuses our attention on rows 2 and 4. But we also know that the conditional is True, and that rules out row 2 where the conditional is False. So, we must be on row 4. In the case row 4 describes, the antecedent is False. So, it must be False.

And that makes the case: the antecedent we've shown False is the sentence assumed at the top of the subproof, P. Since we've shown it's False, we can negate it! $\neg P$.

Alternatively, instead of thinking about the truth table of the conditional in general, let's examine the truth table of $P \rightarrow \bot$. In using ¬I, you show that a sentence with the structure $P \rightarrow \bot$ is True. So, look at its truth-table. Under what circumstances is it True? Just row 2. So, it's True exactly whenever P is False. So, that is: $\neg P$!

A	B	A → B	
T	T	T	
T	F	F	
F	T	T	
F	F	T	←

P	Q	P→(Q∧¬Q)	
T	T	F	
T	F	F	
F	T	T	←
F	F	T	←

P	⊥	P→⊥	
T	F	F	
F	F	T	←

Why the negation rules always work

In a deductively valid argument, it's impossible for the premises to all be True and the conclusion False. Saying that the conclusion is False is the same as negating it. So, the premises must not be jointly possible with the negated conclusion. If they jointly impossible, they must involve a contradiction. So, for a valid argument, a contradiction can always be derived from its premises and the negation of its conclusion.