Deconstructing Material Implication

Óscar Pereira*

Abstract. This is a polished version of a note I once wrote for CS undergrads, who were suffering from the very common confusion of not understanding why the logical connective "implication" is defined the way it is. The reasons for that are not easily found in the literature, and this text purports to fill that gap.[1]

Additionally, a couple of examples are given of how implication is used in mathematics: first a especially contrived one using first order logic (§2), then the different forms that the induction principle can take (§3), and finally proofs by contradiction (§4).

A final remark: the argument laid down in the following pages sits squarely within the bounds of classical logic. In particular, I will freely make use of double negation (and De Morgan's laws), *reductio ad absurdum*, the contrapositive, and *tertium non datur* (excluded middle). Or to put it differently, if the reader knows what constructive and/or intuitionistic mathematics is, then he is well outside of the intended readership for the present text.

Keywords: implication, material implication, logical connective, induction, proof by contradiction.

1 Introduction

Table 1 below displays the logical connective usually named *implication* (or sometimes, *material implication*). When one first learns logical connectives, it is only natural to assume that a connective named "implication" will function according to the meaning that word has in everyday language, viz., that there exists a *causal* relation between φ and ψ.

φ	ψ	$\varphi \to \psi$
0	0	1
0	1	1
1	0	0
1	1	1

Table 1: Truth table for the implication logical connective. As is customary, 0 denotes logical falsehood, and 1 logical truth.

*Contact: {https://, oscar@}randomwalk.eu. Date: 2nd August 2024.
Updated versions of this document and other related information can be found at https://randomwalk.eu/scholarship/material-implication/.

Such an assumption is, however, **wrong**. After all, a truth table describes a well-defined mechanical procedure, which stays the same whether or not there exists causality between φ and ψ. And furthermore, if the role of the connective $\rightarrow$ was really to ascertain that φ somehow causes ψ, then how to explain the first two rows in the table, that apply when the antecedent is false?

So what does the logical connective of implication *actually* represent? To this author, it is a prime example of a situation where being (mathematically) lazy is the right thing to do. Indeed, as causality is just too complicated to establish, mathematicians just *ignore it*—yes, you read that correctly—giving a definition of implication that does not take it into account. And the reason this makes sense, is because despite being a fixed rule, it has the following notable property: whenever there **is** a causal relation between φ and ψ (i.e. if φ happens, then so does ψ), irrespective of how that is to be ascertained, then the logical connective $\rightarrow$ **works as expected**.

What does this mean? That when $\varphi \rightarrow \psi$ has the value 1, and a causal relation exists as described, then the respective truth table lines are coherent with the intuition for causality. Indeed, if the antecedent (φ) is true, the consequent (ψ) must necessarily also be true (last line of the truth table). And when a causal relation exists, if the antecedent is false, then, whether the consequent is true or false, *the causal relation is unaltered* (i.e. it continues to exist). This explains the first two lines of the truth table. As for the third line, note that, while it is usually not trivial to determine when causality does exist, it is dead easy to note when it *cannot* exist: if the antecedent is true and the consequent is false, then most certainly the former does not cause the latter—which means $\varphi \rightarrow \psi$ must have the value 0.

And what about situations where we do not know whether or not there is a causal relation between φ and ψ? Here we cannot have φ be true and ψ be false—for in this case, we would *know* that causality does *not* exist. But if any of the other three possibilities occur, then causality *might* exist, or it might *not*: we simply do not know. For the reasons explained above—coherent behaviour when causality does exist—in all of these three scenarios, the value of $\varphi \rightarrow \psi$ is 1. In other words, $\varphi \rightarrow \psi$ means that having φ be the cause of ψ is not impossible, because φ being true and ψ being false never happens.

Alternative formulations. It is a simple matter to verify that $\varphi \rightarrow \psi$ is logically equivalent to both $\neg\varphi \vee \psi$ and $\neg\psi \rightarrow \neg\varphi$. The latter one is called the *contrapositive*, and it will be mentioned several times in what follows, beginning right below. The former affords us yet another way of thinking about of $\varphi \rightarrow \psi$, namely, as meaning that "either φ does not happen, or ψ happens, or both."

Implication in prose. We finish this section with a brief detour into language. Statements like $\varphi \to \psi$ can be read in a slew of different ways: besides "φ implies ψ," we can also have "if φ, then ψ," "ψ, if φ," "ψ follows from φ," or "φ only if ψ." To see why this last one makes sense, note that saying that φ happens only if ψ also happens, is tantamount to saying that if ψ does *not* happen, then neither can φ (for otherwise, if φ were to happen, so should ψ, which is against the assumption that ψ does not happen). The latter condition, $\neg\psi \to \neg\varphi$ is, via the contrapositive, equivalent to $\varphi \to \psi$.

Finally, if we have "φ, if ψ" and "φ only if ψ," we can combine them to get "φ if and only if ψ," often shortened to "φ iff ψ," for the *biconditional*: $\varphi \leftrightarrow \psi$.

2 A Practical Example

Here is a good example to test one's understanding of the mathematical meaning of the implication connective, due to Tim Gowers [2]. Given an interval X of real numbers, its *diameter* is defined as the largest absolute difference of any of its elements, i.e. $\mathrm{diam}(X) \overset{\text{def}}{=} \sup|x-y|$, for all x, y in that interval. The following proposition is then clearly true:[2]

$$(\forall x \in X \quad |x| \le 1) \to \mathrm{diam}(X) \le 2$$

However, using the properties of first order logic, we can rewrite the exact same statement in another way. Begin by rewriting the implication as a disjunction:

$$\neg\big(\forall x \in X \quad |x| \le 1\big) \vee \mathrm{diam}(X) \le 2$$

Now, the negation of an universal quantifier, is an existential quantifier: to say that it is not true that for all x, $P(x)$ holds, is to say that there exists at least one x such that $P(x)$ is false, i.e., such that $\neg P(x)$ is true. Hence the above disjunction can be written as:

$$\big(\exists x \in X \quad \neg(|x| \le 1)\big) \vee \mathrm{diam}(X) \le 2$$

However, $\mathrm{diam}(X) \le 2$ does not depend on the quantified variable x, meaning its truth value is not affected by whatever x turns out to be. Hence, we can pull it inside the scope of the quantifier:[3]

$$\big(\exists x \in X \quad \neg(|x| \le 1) \vee \mathrm{diam}(X) \le 2\big)$$

And finally, we turn this disjunction into an implication:

$$\big(\exists x \in X \quad (|x| \le 1) \to \mathrm{diam}(X) \le 2\big)$$

This last statement, though mathematically correct, is clearly nonsense when the implication is taken to have its everyday meaning: there is no way in the diameter of an interval X depends on the absolute value *of only one* of its elements! If the reader can understand the meaning of the statement above, he will have a good grasp of what mathematicians mean when they talk about "implication." The note at the end of the current sentence provides the answer, but the reader will profit from attempting to answer it himself first, before looking at the solution.[4]

3 Another Example: The Induction Principle

The example in the previous section was especially contrived to illustrate the difference between the mathematical (actually, the logical) definition of "implication," and its everyday usage. However, in mathematical practice, we have seen that the most important aspect to keep in mind, is that $\varphi \to \psi$ is true means that when φ is true, ψ must also be true. A good non-contrived example of the use of implication in mathematics is the notion of a *proof by induction*. It has three equivalent formulations:

1. **Weak induction.** If a proposition $\varphi(n)$ holds for 0, and we have $\varphi(n) \to \varphi(n+1)$, then $\varphi(n)$ holds for all $n \in \mathbb{N}$.
2. **Strong induction.** If a proposition $\varphi(n)$ holds for 0, and we have $\big(\varphi(0) \wedge \varphi(1) \wedge \cdots \wedge \varphi(n)\big) \to \varphi(n+1)$, then $\varphi(n)$ holds for all $n \in \mathbb{N}$.
3. **Well-ordering principle (WOP).** Given any subset S of $\mathbb{N}$, if it is nonempty, then it has a smallest element.

Each of the statements relies on implications, and in fact, weak and strong induction use "double" implications, in the sense that in both cases, the antecedent consists of two hypothesis, one of which is itself an implication. Furthermore, saying all three statements are equivalent means that they all imply one another. It should be easy to see that weak induction implies strong induction, but the converse is not so obvious—and neither is how the WOP implies either form of induction, or vice-versa. We will show this by proving that strong induction implies that WOP, and the WOP implies weak induction. To see how this proves their equivalence requires the following lemma—which is also what one would expect from the everyday meaning of the word "implication."

Lemma 3.1 (Transitivity of implication). *Let φ, ψ, τ be propositions, and suppose that $\varphi \to \psi$ and $\psi \to \tau$. Then $\varphi \to \tau$.*

Proof. The lemma is an implication, and it cannot be false because it is not possible to have the antecedent be true and consequent be false.

Indeed the latter means $\varphi \to \tau$ is false, i.e., φ is true and τ is false. But then, if ψ is true, $\psi \to \tau$ is false, and if ψ is false, $\varphi \to \psi$ is false. In either case, the antecedent of the lemma is also false—which means the main implication, i.e. the lemma, cannot be false. This gives the proof. ■

Going back to induction, if we prove that strong induction implies the WOP, and that the WOP implies weak induction, then by lemma 3.1 we will have proven that strong induction implies weak induction—and more generally, that the three formulations are indeed equivalent.

Lemma 3.2. *Strong induction implies the well-ordering principle.*

Proof. Let S be a subset of $\mathbb{N}$, and suppose it has no smallest element. Then $0 \notin S$, for otherwise 0 would be the smallest element. But then, also $1 \notin S$, because then 1 would be the smallest element. Let $P(n)$ be the proposition "$n \notin S$." It is clear that $\big(P(0) \wedge P(1) \wedge \cdots \wedge P(n)\big) \to P(n{+}1)$. By strong induction, $P(n)$ is true for all $n \in \mathbb{N}$—i.e., S is empty. By the contrapositive, if S is nonempty, it has a smallest element. ■

Lemma 3.3. *The well-ordering principle implies weak induction.*

Proof. Let $\varphi(n)$ be a proposition for which the hypothesis of weak induction apply, namely, $\varphi(0)$ is true, and $\varphi(n) \to \varphi(n+1)$ holds. Then $\varphi(n)$ is true for all $n \in \mathbb{N}$. To see this, let S be the subset of $\mathbb{N}$ defined as follows $\{n \in \mathbb{N} \colon \varphi(n) \text{ is false}\}$. S cannot have a smallest element, because if one such element existed, call it m, then as $\varphi(0)$ holds, we would have $m > 0$, and thus $m - 1 \leq 0$ and moreover, $\varphi(m-1)$ would be true—otherwise $m - 1$ would be in S. But $\varphi(m-1)$ would then imply that $\varphi(m)$ holds, which is contradictory. Hence, S is empty, meaning $\varphi(n)$ is indeed true for all $n \in \mathbb{N}$. ■

If the reader could follow this section without difficulty, he is now ready to embark on a more abstract topic.

4 Implication And Mathematical Proofs

Not all mathematical theorems are explicitly of the form $\varphi \to \psi$, but we *can* think of any theorem as an implication, in the following sense: each theorem depends on a set of *hypothesis*—if this were not so, the "theorem" would in fact be an axiom. Hence, for any theorem τ, it will depend on hypothesis $H_1, \ldots, H_n$, and proving the theorem means showing that the theorem being false, whilst the hypothesis are true, cannot happen. This is tantamount to proving that $\big(H_1 \wedge \cdots \wedge H_n\big) \to \tau$ holds.[5]

Furthermore, we can now restate one important proof technique—the so-called proof by contradiction, or *reductio ad absurdum*—in terms of the implication logical connective, and more concretely, in terms of the contrapositive. Recall that in a proof by contradiction, we assume the opposite of what we want to prove, and derive a contradiction. As far as I am aware, the contradiction to be derived can always be seen as (possibly more than) one of the following cases:

1. **Contradict something.** This is the more general form of a proof by contradiction: to prove theorem τ, assume $\neg\tau$ and derive *falsum*, i.e., a condition of the form $\eta \wedge \neg\eta$. This means we show $((\bigwedge H_i) \wedge \neg\tau) \rightarrow (\eta \wedge \neg\eta)$. Via the contrapositive (and because double negation equates no negation) we obtain $(\neg\eta \vee \eta) \rightarrow (\neg(\bigwedge H_i) \vee \tau)$. As the antecedent is always true, so must the consequent, which we can rewrite as $(\bigwedge H_i) \rightarrow \tau$—i.e., we prove theorem τ.
2. **Contradict one of the hypothesis.** To prove τ, we assume $\neg\tau$, and show it implies $\neg H_1 \vee \neg H_2 \vee \cdots \vee \neg H_n$. This is the contrapositive of $\big(H_1 \wedge \cdots \wedge H_n\big) \rightarrow \tau$, and on the first step, we implicitly assume its negation, $\big(H_1 \wedge \cdots \wedge H_n\big) \wedge \neg\tau$.
3. **Contradict the antecedent.** This is a particular form of the previous case: when τ is of the form $\varphi \rightarrow \psi$, we take φ as one of the hypothesis,[6] and then assuming $\neg\psi$, we show that $\neg\varphi$ holds. This establishes its contrapositive $\varphi \rightarrow \psi$, which is what we wanted to show. Note that we again begin by assuming its negation, i.e., $\varphi \wedge \neg\psi$.

The proof of lemma 3.3 can be seen as an example of case 3: we show that if there exists an m such that $\varphi(m)$ is false, then assuming the hypothesis that $\varphi(0)$ holds, the hypothesis $\varphi(n) \rightarrow \varphi(n+1)$ must be false, because we have $\varphi(m-1) \wedge \neg\varphi(m)$. It can also be seen as an example of case 1, because assuming that the lemma is *false*—i.e., assuming that the well-ordering principle does *not* imply weak induction—we derive $\varphi(m) \wedge \neg\varphi(m)$.

Remark 4.1. Case 1 above can also be understood without recourse to the contrapositive: if the implication $((\bigwedge H_i) \wedge \neg\tau) \rightarrow (\eta \wedge \neg\eta)$ is always true, and the consequent is always false, then the antecedent must also always be false—which of course means that its negation, $\neg(\bigwedge H_i) \vee \tau$, must always be true, or equivalently, that $(\bigwedge H_i) \rightarrow \tau$ must be always true. △

Notes

[1] There might be *other* reasons—besides the one put forth in the present text—to define material implication the way it is defined, but in my experience this one usually suffices to satisfy students' curiosity. Moreover, as van Dalen writes, '[t]here is no compelling reason … to stick to the notion of implication that we just introduced [i.e., the one in table 1]. Various other notions have been studied [but] for mathematical purposes our notion … is, however, perfectly suitable.' [1, p. 16f].

[2] $\forall x \in X$ is read "for all x in (belonging to) X (some property holds)." $\exists x \in X$ is read "there exists (at least) one x in X (such that some property holds)."

[3] In case this step is not entirely clear, here is a more detailed explanation. Say P is a proposition that does not depend on x. If it is true, than so is $\exists x\, P$: one can just pick any x one wants! And if P is false, then so is again $\exists x\, P$: regardless of what x one chooses, P continues to be false—which of course means that $\exists x\, P$ is also false.

[4] To say that there exists an $x \in X$ such that the implication $(|x| \leq 1) \rightarrow \text{diam}(X) \leq 2$ is true, simply means that either there is an x such that the antecedent is false—i.e., an x such that $|x| > 1$—or there is an x such that the consequent is true, i.e., such that $\text{diam}(X) \leq 2$. Thus if the antecedent is true—i.e., if there exists no x such that $|x| > 1$—then there must exist an x such that $\text{diam}(X) \leq 2$ holds, or equivalently, $\text{diam}(X) \leq 2$ must also be true. It should be clear to see that this is a correct statement.

[5] As an ancillary remark, note that when τ *is* of the form $\varphi \rightarrow \psi$, then we can rewrite $\big(H_1 \wedge \cdots \wedge H_n\big) \rightarrow (\varphi \rightarrow \psi)$ as $\big(H_1 \wedge \cdots \wedge H_n \wedge \varphi\big) \rightarrow \psi$—this is an immediate consequence of being able to turn implications into disjunctions: $a \rightarrow (b \rightarrow c) = \neg a \vee (\neg b \vee c) = (\neg a \vee \neg b) \vee c = \neg(a \wedge b) \vee c = (a \wedge b) \rightarrow c$.

[6] Cf. note 5.

References

1. van **Dalen**, Dirk, 2013 [1980]. *Logic and Structure*. Boca Raton, FL: Springer, 5th edition. Cited once on page 7.
2. **Gowers**, Timothy, 2013. *A little paradox*. See `https://gowers.wordpress.com/2013/12/09/a-little-paradox/`, last access on June 19, 2024. Cited once on page 3.