TOWARDS A THEORY OF CAUSAL IMPLICATION

Alexander Stepanov
Department of Electrical Engineering and Computer Science

Polytechnic Institute of New York
333 Jay Street
Brooklyn, New York 11201

Abstract

We present a formal theory of causal implication, which allows to reason about cause-effect relationship in a more satisfactory way than the classical calculus of proposition. In this paper only implicational subsystem is presented, but it easily extends to other propositional conjuncts and quantification. The system could serve as an underlying formalism for the development of rule-based expert systems.

l. Introduction

The classical propositional logic is commonly acknowledged to be an underlying formalism for rule-based expert systems [4, 5, 15]. It is also commonly acknowledged that it is not a satisfactory formalism [5]. The material implication is used to represent cause-effect relationship between real world events in spite of the fact that its semantics differs quite essentially from our intuitive notion of causality. The problems connected with the use of material implication were extensively studied by logicians. The book of Anderson and Belnap [l] is a mine of information. But the motivation for most of that work (there are some exceptions such as [2], [8], [14]) came from proof theory and as a rule logical works on relevant implication are not dealing with causality. There are very many valuable ideas scattered around in philosophical literature. [16] is a very good and concise collection of some recent papers on the philosophy of causation. Several very important ideas pertaining to the development of the appropriate theory of causality were developed by AI community. Among them works on non-monotonic logic [9, 10] and logic programming [6, 7] are especially relevant.

Our basic requirement for the theory we are trying to develop is, obviously, that the system corresponds to our intuitive notion of causality. But there are some other secondary requirements, which are worth noting.

It is desirable to have a system that would not be able to derive any sentence, however "true" it may be, if it contains a fact which is not contained in the premises. For example, it rains should not entail Dallas is the capital of New Jersey, or it rains. (This requirement under the name of *principle of containment* was first suggested by William T. Parry in [12].)

It is of great importance that a system be stable when a contradiction is introduced. We do not want to derive that it rains even when we start with premises: it is dark, it is not dark, and if it rains then it is wet outside. This requirement is satisfied by system E of Anderson and Belnap [l] and by some other systems of non-standard logic (see [l] for a bibliography and description of many of these systems; another approach is presented in [13]; readable account of some of this problems addressed to AI audience is given in chapter 4 of [15]).

It is taken for granted that every tautology should be always derivable. Because of that, starting with no premises we can derive that if it rains then it rains or it snows. (It is ambiguous, but both parsings give us a tautology). However true this proposition may be it does not carry any information precisely because it is a tautology. So what we really should want is a system in which no classical tautology is ever derivable. Indeed, more so, instead of deriving that (a and c) implies b from a implies b, we should forget the former the moment we discover the latter.

We definitely do not want to derive from It rains that It rains and it rains, obtaining infinitely many conclusions a finite set of premises using only propositional reasoning.

We do not want to require completeness in the traditional sense because that would require every tautology to be derivable and we do not want any tautology to be derivable. Indeed, any tautology can be added to a theory without making it inconsistent; moreover, a tautology is satisfiable under any interpretation, but we still do not want to derive it. What we want is that if a fact is derivable from a consistent set of premises by a classical inference then we will also derive it.

(In this paper the proofs of all of the propositions are left as exercises for the reader.)

2. Terminology

The term *implication* in classical logic is overloaded. The same is true in natural languages. We use the same word *imply* to say:

(1) eating curare *implies* death,
(2) he is alive *implies* he has not eaten curare, and
(3) eating curare *implies* death and death *implies* burial *implies* eating curare *implies* burial.

We are going to distinguish between these three senses. We will use a word *causes* for a primitive kind of implication, the word *implies* for a secondary relation, and *infers* for the third one. We will denote them by the signs `->`, `=>`, and `)-` correspondingly. When we use)- (the turnstile) we will specify what set of inference rules we mean by prefixing the name of it to the turnstile. For example, we can say `a, a->b (CL)- b`, meaning classical modus ponens. When we use `A (CL)- B` we mean that the set of formulas `A` classically implies `B` when all occurrences of different kinds of implications in `A` and `B` are interpreted as material implication.

Our intuition for *causes* is based on the following premises:

- facts do not cause laws;
- laws do not cause laws;
- laws do not cause facts by themselves;
- facts cause other facts to happen according to the laws.

In other words `->` (*causes*) is a first degree function which accepts only atomic facts. No nesting of `->` is allowed.

The negation of a fact is its absence, which can cause other facts. So when we refer to facts, we mean either atomic facts or their negations. We will denote a negation of `a` as `not-a`. The law of double negation is assumed throughout, and `not-not-a` is automatically replaced with `a`. We can for the time being restrict negation only to facts (so it, as well as `->`, is a first degree function) because the negation of a causal law is more or less vacuous. From `a does not cause b` nothing much can be derived about `a` or `b`. It is definitely impossible to derive from it that `a and not-b`.

Another starting point is the fact that `or`, `and`, and other propositional conjuncts are secondary to implication and negation. While it is possible to introduce them, we will not, at present, do it.

We are going to use small letters `a, b, c...x, y, z` for atomic facts and capital letters `A, B, C...X, Y, Z` for sets of formulas.

3. Systems of laws

A system of laws is a set of ordered pairs of facts. For example:

```
a -> b, not-c -> d, c -> not-a
```

is a system of laws.

We can introduce a new relation `=>` (*implie*s) using the following rules:

(Rl) `a -> b )- a => b`

(R2) `a => b , b => c )- a => c`

(R3) `a => b )- not-b => not-a`

This new relation is a product of a secondary causality relationship defined by **Rl** and **R2** and the *contraposition rule* **R3**. While adding **Rl** and **R2** still leaves us within the connotation of the notion of causality, R3 makes this relation not quite causal. We do not

want to *contrapose* smallpox causes fever into absence of fever causes absence of smallpox.

A system of laws is also required to satisfy four axioms.

Axiom I

(Al) No fact implies its own negation

It is an old principle of Aristotle (see *Prior Analytics*, 57a36 - 57bl7). Aristotle explicitly equated implication with causality stating that we infer a fact "when we know the cause on which the fact depends"(*Posterior Analytics*, 71blO); moreover, premises for him are prior to the conclusions "in the order of being" (*ibid.*, 71b20 - 72a5) [3,11]. This principle of his served as a starting point for the development of *connexive logic* of Storrs McCall [8]. (See also [2] and [5].)

From **Al** we immediately derive:

Proposition 1

For no two facts a and b,

```
a => b and a => not-b,
```

nor

```
a => b and not-a => b
```

This proposition is usually known as *Boethius's Thesis*.

And with a little bit more effort we obtain

Proposition 2

For any system of laws `S` and for no fact `a`, S `(CL)- a`

As we noted before we interpret all `->` in `S` as material implications when we use `(CL)-`.

From **P2** immediately follows that any system of laws is consistent in the classical sense when combined with any fact. (Or, as a matter of fact, when it is not combined with any fact.) In other words, a set of implicational clauses S is satisfiable if and only if it satisfies **Al**, and for any variable `x` and any truth-value `V` there is an assignment which will satisfy all clauses of `S` and will assign `V` to `x`. We use classical rules of inference in **P2** because these are the only rules of inference known to the reader at this point, but **P2** is going to remain true for all rules of inference we are going to introduce in this paper. The intuitive meaning of **P2** is that facts do not follow from laws, unless some facts are already known.

Axiom II

(A2) No fact implies itself

This is another of Aristotle's principles. He remarks that those who think that "`a` implies `a`" (remember that for him *implies* and *causes* are largely equivalent) "have an easy way to explain anything" (*Posterior Analytics*, 72b30 - 72b35).

And we immediately derive

Proposition 3

(P3) For no two facts `a` and `b`, `a => b and b => a`

A2 makes *implies* to be a partial ordering, which it should by all means be, since we definitely do not want to have circular causal chains.
(It should be remembered that we are dealing with individual facts and not with classes.)

For what will follow later we need a couple of definitions.

Definition 1

Fact `a` is called *caused* if there is a fact `b`, such that `b -> a`

Definition 2

Fact a is called *initial* iff both `a` and `not-a` are not caused

And we can introduce our third axiom:

Axiom III

(A3) Any finite non-empty subsystem of a system of laws has initial facts

This axiom makes a set of facts of a system of laws into a partially ordered set with the ordering relation *before* which can be defined as a transitive closure of *causes* with negations factored out.

And our forth axiom:

Axiom IV

(A4) It is never the case that a fact and its negation are both caused

The intuitive explanation of this axiom is that for a pair of events `a` and `not-a` only one of them is a “real” event and the other one is just an absence of its opposite. Either cancer has causes and its absence results only from the absence of them, or the other way around. From the practical point of view this axiom makes it impossible for any set of initial conditions to lead to contradiction. In other words, it makes all initial facts independent from each other.

Proposition 4

For no two initial facts `a, b, a => b` or `a => not-b`.

4. Causal systems

Our four axioms made systems of laws so weak that no fact can be derived from them when `->` is interpreted in the classical sense (as material implication). And that was exactly what we wanted. To make things work we need to add to system of laws some facts and some rules of inference.

Definition 3

A causal system `S` is a triple `<L, F, R>`, where `L` is a system of laws, `F` – a system of facts, and `R` – a set of inference rules.

We will say that `S )- a` (a fact `a` can be *inferred* from `S`) iff `a` can be obtained from `L` and `F` by applying inference rules from `R`.

One way to look at this triplet is to view a set of laws as a program, a set of facts as an input data for this program and a set of rules as a computer which evaluates this program. Instead of changing the syntax of our programs we are going to develop a sequence of computers which can compute more and more results applying the same program to the same data. The approach of making a hierarchy of inference rules seems to be more suitable for what we are trying to do than the more traditional approach of strengthening a system by adding more axioms.

5. Weak rules of causal implication

The main reason people use material implication to reason about causality is that it satisfies our most basic intuition about cause-effect relationship, namely, if a cause happened the effect will follow and from the absence of the effect we deduce the absence of a cause. And we would definitely want to make these into our basic rules of inference:

Weak Rules:

(WR)

Modus ponens:

(MP) `a, a -> b (WR)- b`

Modus tollens:

(MT) `not-b, a -> b (WR)- not-a`

In a sense, **MP** and **MT** do exactly as much as the *implies* relation.

Proposition 5

For any causal system `S`, and for any fact a not in the system of facts of `S, S (WR)- a` iff there is a fact b in the system of facts of `S`, such that `b => a`.

Definition 4

A causal system `S` is weakly consistent iff for no fact `a`, `S (WR)- a`, and `S (WR)- not-a`.

Proposition 6

For any weakly consistent causal system `S, S (CL)- a` iff `S (WR)- a`

The meaning of this is that whatever fact can be derived from our causal system using all the rules of propositional calculus we can derive it just with modus ponens and modus tollens.

A system with the weak rules of implication has another nice property. If we denote a set of facts derivable from a system of facts `F` and a system of laws `L` with the weak rules of implication as `SET(L,F)`, then we immediately obtain

Proposition 7

```
SET(L,Union(F,G)) = Union(SET(L,F),SET(L,G))
```

This makes systems with the weak rules of implication as stable when a contradiction is present as anything could be. Namely,

Proposition 8

```
SET(L,Union(F,<a,not-a>)) =
Union(SET(L,F),SET(L,a),SET(L,not-a))
```

So, the system of weak implication satisfies all our requirements, but one. It can be shown that it is impossible to implement "or". Nor is it possible to implement negation. (Axiom II does not allow having a system `not-a -> b, a -> not-b`).

So we need more rules.

6. Closure rule

Here we can use what is commonly known as *the closed universe assumption*. Namely, if we know that none of the causes of some fact happened we can reasonably derive that the fact also did not happen. Or, putting it more formally,

Closure Rule:
(CLR)

If

(l) fact `a` is caused,
(2) fact `a` is not derivable by **WR**
(3) for every fact `b`, such that `b -> a`, it is already derived that `not-b`

then
`not-a`

It should be noted that we do not have a general *negation as failure* rule. If something is not derivable we do not assume that it is false. Only when we know that all of the causes of some event did not happen we derive that the event did not happen. That allows us to control the default reasoning by explicitly including unknown causes in our causal systems.

If we start with a system which is consistent under modus ponens and modus tollens it will remain consistent when the closure rule is added.

Proposition 9

For any weakly consistent system `S` and any fact `a` it is not that `S (WR+ClR)- a` and `S (WR+ClR)- not-a.`

And now we can implement

```
not: a -> not-b

or: a -> c, b -> c

and: not-a -> not-c, not-b -> not-c
```

Proposition 10

Any Boolean function can be implemented by a causal system using WR+ClR.

More than that, the closure rule guarantees that if we know the initial state of the world we can derive everything about its future:

Proposition 11

For any causal system `S = <L, F, WR+CLR>` if `F` contains for every initial fact either it or its negation then for any fact `a` in `L, S )- a` or `S )- not-a.`

And that if we start with the consistent set of initial facts we shall never derive a contradiction:

Proposition 12

For any causal system `S = <L, F, WR+CLR>` if `F` contains only initial facts or their negations (but not both for the same fact) then for no fact `a` in `L, S )- a` and `S )- not-a.`

7. Backward deduction

Now we can do all possible inferences from causes to their effects. But in many cases we cannot make a reasonable deduction from effects to their causes. For example, let us look at the causal system `S` with set of laws

```
L = < a -> d, b -> d, c -> d>
```

and a set of facts:

```
F = < not-a, not-c, d >.
```

We know that something has caused d to happen. And we know that it was not `a` or `c`. So it was `b`. But we cannot derive it with the help of WR+CLR. In many cases the following rule will do what is needed:

Weak Rule of Backward Inference:

(WRBI)

If

(l) `a` happened,
(2) for no `b`, such that `b -> a`,
`(WR+CLR)- b`,
(3) there is unique `c`, such that `c -> a`,

and it is not that (WR+ClR)-not-c
then
c

In other words if there is only one explanation for something, this explanation better be true.

Unfortunately, this rule is not sufficient. But the rule which is sufficient is not that easy to use. In the worst case it will take exponential number of steps (assuming that P is not equal NP). And here it is:

Rule of Sufficient Reason:

(RSR)

Something which is caused happens if and only if some of its causes have happened

This rule is so strong that all the other rules we introduced so far can be inferred from it.

Proposition 13

For any system of laws L and a system of facts F if L, F (WR+ClR+WRBI)- x then L, F (RSR)- x.
And if a value of an argument of a Boolean function is derivable with the help of propositional calculus we will derive it with the help of the rule of sufficient reason. (No wonder, since it brings a decision procedure for propositional calculus into our system.)

Proposition 14

For any Boolean function y = B(xl...xn) there exists a causal system S such that for any system of facts F containing some of <y,xl...xn> or their negations B,F (CL)- z iff S,F (RSR)- z, where z is one of <y,xl...xn> or a negation of one of them.

8. Bibliography

1. Anderson, Alan Ross and Nuel D. Belnap, Jr., *Entailment. The Logic of Relevance and Necessity*, Princeton University Press, Princeton, 1975

2. Angell, R.B., A Propositional Logic with Subjunctive Conditionals, *The Journal of Symbolic Logic,* Volume 27, Number 3, Sept. 1962, pp. 327-343.

3. Barnes, Jonathan, *Aristotle's Posterior Analytics*, Clarendon Press, Oxford, 1975

4. Buchanan, Bruce G. and Edward H. Shortliffe (eds.), *Rule-Based Expert Systems*, Addison-Wesley, Reading, Mass., 1984

5. Charniak, Eugene and Drew McDermott, *Artificial Intelligence*, Addison-Wesley, Reading, Mass., 1985

6. Clark, K.L., Negation as Failure, in *Logic and Data Bases*, (Gallaire, H. and J. Minker, Eds.), Plenum Press, New York, 1978

7. Kowalski, Robert, *Logic for Problem Solving*, North-Holland, New York-Amsterdam-Oxford, 1979

8. McCall, Storrs, Connexive Implication, *The Journal of Symbolic Logic*, Volume 31, Number 3, Sept. 1966, pp. 415-433.

9. McCarthy, John, Circumscription - A Form of Non-Monotonic Reasoning, *Artificial Intelligence* 13 (1980), pp. 27-39

10. McDermott, Drew and Jon Doyle, Non-monotonic Logic I, *Artificial Intelligence* 13 (1980), pp. 41-72

11. McKeon, Richard (ed.), *The Basic Works of Aristotle*, Random House, New York, 1941

12. Parry, William T., Ein Axiomensystem fur eine nue Art von Implikation (Analytische Implikation), *Ergebnisse eines mathematischen Kolloquiums*, vol. 4, 1933, pp. 5-6

13. Rescher, Nicholas and Robert Brandon, *The Logic of Inconsistency*, Rowan and Littlefield, Totowa, N.J., 1979

14. Routley, R. and Montgomery, H., On Systems Containing Aristotle's Thesis, *The Journal of Symbolic Logic*, Volume 33, Number 1, March 1968, pp. 82-96.

15. Sowa, John F., *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley, Reading, Mass., 1984

16. Sosa, Ernest (ed.), *Causation and Conditionals*, Oxford University Press, Oxford, 1975