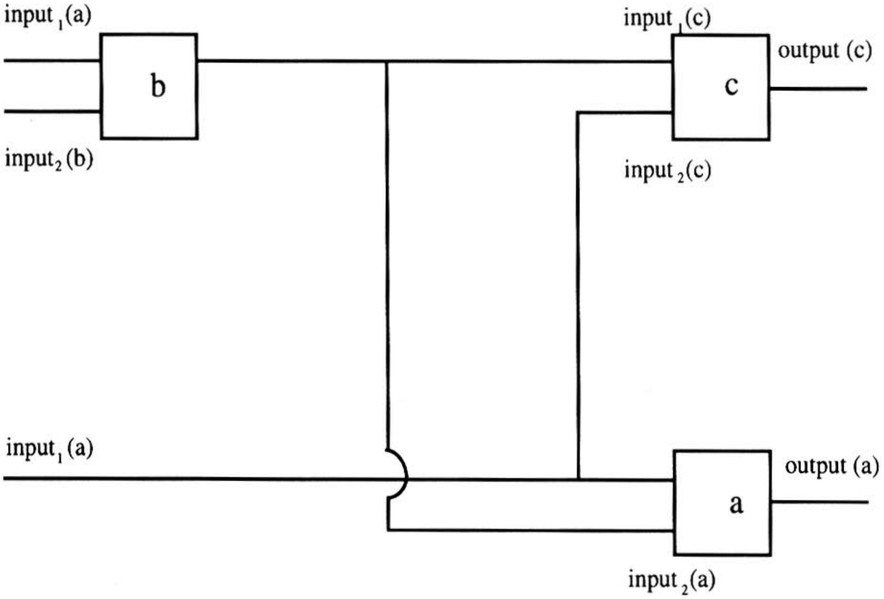
Logique & Analyse **145** (1994), 13-22

DEFAULT REASONING IN THE CORRECTION OF FALSIFIED SYSTEM DESCRIPTIONS

Erik WEBER[*]

1. *Introduction*

Consider the following electric circuit, which contains three gates:



Consider also an inquirer who has the following theory about how this system works:

(1) *a* is an AND-gate.
(2) *b* is an XOR-gate.

[*] Postdoctoral Fellow, Belgian National Fund for Scientific Research. I want to thank Diderik Batens for his comments on earlier versions of this paper.

(3) c is an XOR-gate.
(4) output(b) = input$_2$(a)
(5) output(b) = input$_1$(c)
(6) input$_1$(a) = input$_2$(c)

This theory will be called the original system description (SD_O). AND-gates have output 1 if and only if both inputs are 1. XOR-gates have output is 1 if and only if one of the inputs is 1 and the other 0. I assume that the inquirer's reason to accept statements (4)-(6) is that he can observe that all the wires are properly connected. The statements (1)-(3) are accepted because all gates look the same, apart from a label that has been attached to them by the manufacturer of the circuit. This label tells us which type of gate it should be. So a is labelled AND, while b and c are labelled XOR.

Suppose the inquirer observes that

output(c) = 1 & output(a) = 0

He wants to explain these outputs by means of SD_O and empirical data about the inputs of the circuit. The observed inputs are

input$_1$(b) = 1 & input$_2$(b) = 0 & input$_1$(a) = 1.

From these data and (SD_O) it follows that

output(c) = 0 & output(a) = 1.

So, the attempt to explain the observed outputs leads to a contradiction. As he has a sufficient reason to believe that (SD_O) is false, the inquirer rejects it and starts looking for a new system description.

In section 2 I develop a method for finding a new system description. This method can be applied to all circuits with two or three gates, and makes use of default rules as introduced in Reiter 1980. In section 3 I compare my method with two rivals. I argue that my method is the best tradeoff between speed an reliability. In section 4 I show how the method developed in section 2 can be generalized for circuits with more than three gates.

2. *A method based on default reasoning*

2.1 My method is based on the assumption that, because the inquirer can observe the wiring in the system, he remains convinced that (4)-(6), or the analogous statements in other system descriptions, are true. Because we are dealing with circuits with three gates, and there are sixteen types of gates with one output and two inputs, this assumption entails that 16^3-1 alternatives to the original system description are available to the inquirer. The method I will develop is a fast and reliable instrument for deciding which alternative to accept. For convenience, I give each of the sixteen types of gates a number and most of them a name:

T_1 11 1 10 1 01 1 00 1 TAUT	T_2 11 1 10 1 01 1 00 0 OR	T_3 11 1 10 1 01 0 00 1	T_4 11 1 10 0 01 1 00 1 IMPL
T_5 11 0 10 1 01 1 00 1	T_6 11 1 10 1 01 0 00 0 LEFT	T_7 11 0 10 0 01 1 00 1 NOT-LEFT	T_8 11 1 10 0 01 0 00 1 EQ
T_9 11 0 10 1 01 1 00 0 XOR	T_{10} 11 1 10 0 01 1 00 0 RIGHT	T_{11} 11 0 10 1 01 0 00 1 NOT-RIGHT	T_{12} 11 1 10 0 01 0 00 0 AND
T_{13} 11 0 10 1 01 0 00 0	T_{14} 11 0 10 0 01 1 00 0	T_{15} 11 0 10 0 01 0 00 1 NEITHER	T_{16} 11 0 10 0 01 0 00 0 CONTR

As already mentioned, my method makes use of default rules. We need two *default schemes* to formulate the method:

(I) For all gates *x*:
$$\frac{SD_O \vdash T_\alpha(x) : T_\alpha(x)}{T_\alpha(x)}$$

(II) For all gates *x* and *y*:
$$\frac{SD_O \vdash T_\alpha(x)\&T_\beta(y) : T_\alpha(x)\&T_\beta(y)}{T_\alpha(x)\&T_\beta(y)}$$

T_α and T_β are arbitrary types of gate. T_α and T_β may be identical. SD_O is the original system description. The first default scheme must be read as follows: if the original system description implies that $T_\alpha(x)$, and it is consistent to believe that $T_\alpha(x)$, then assume that $T_\alpha(x)$ is true. The meaning of the second scheme is of course analogous. A default rule of form (I) is said to be *applicable* if and only if there is a gate *x* such that $SD_O \vdash T_\alpha(x)$ and it is consistent to believe that $T_\alpha(x)$. An *application* of a default rule of form (I) is an act consisting of two parts. First, a list is compiled of all gates *x* for which $SD_O \vdash T_\alpha(x)$ and it is consistent to believe that $T_\alpha(x)$. Then for every *x* in this first list, a list is compiled of system descriptions that are compatible with the default conclusion $T_\alpha(x)$. Analogous definitions can be given for defaults of form (II).

The method I propose is the following:

(M_3) (1) Check whether there are applicable default rules of form (II). If there are, apply them all and go to (4). If there are no applicable defaults of form (II), go to instruction (2).
(2) Check whether there are applicable defaults of from (I). If there are, apply them all and go to (4). If there are no applicable defaults of form (II), go to instruction (3).
(3) Compile a list of all possible system descriptions.
(4) Determine which of the system descriptions in the lists obtained by instruction (1), (2) or (3), are falsified by the original observation. Remove them from the lists.
(5) Perform all possible measurements on the system.
(6) Calculate which of the remaining system descriptions are falsified by the results of the measurements. Remove them from the lists. If one candidate survives, accept it. If all remaining system descriptions are falsified by the results of the measurements, go back to instruction (1), but skip instruction (5) from now on.

The original observation is the observation that falsified SD_O. A measurement is the act of bringing the inputs of the system in some state and observing the resulting outputs. It is impossible that more than one candidate survives the measurements. This method can be applied to any circuit with three gates. To clarify it, I will apply it to the example of section 1.

2.2 According to (M_3), we first have to determine which defaults of form (II) are applicable. In our example, the original system description entails AND(a)&XOR(b), AND(a)&XOR(c) and XOR(b)&XOR(c). So for the following default rules we have an *x* and *y* such that $SD_O \vdash T_\alpha(x)\&T_\beta(y)$:

$$\text{For all gates x and y: } \frac{SD_0 \vdash \mathrm{AND(x)\&XOR(y) : AND(x)\&XOR(y)}}{\mathrm{AND(x)\&XOR(y)}}$$

$$\text{For all gates x and y: } \frac{SD_0 \vdash \mathrm{XOR(x)\&XOR(y) : XOR(x)\&XOR(y)}}{\mathrm{XOR(x)\&XOR(y)}}$$

All other default rules of form (II) are inapplicable because the circuit contains no gates *x* and *y* for which $SD_0 \vdash T_\alpha(x)\&T_\beta(y)$. The second default rule is inapplicable too: {b,c} is the only pair of gates for which $SD_0 \vdash$ XOR(x)&XOR(y), and it is not consistent to believe that XOR(b)&XOR(c). If *b* and *c* would be XOR-gates, the output of *c*, given the inputs, would be 0; this is contrary to what we have observed. The first default rule is applicable because it is consistent to believe that AND(a)&XOR(c): if *a* is an AND-gate and *c* an XOR-gate, the output of *a*, given the inputs, is 0 if the output of *b* is 0. This is possible, as *b* is not necessarily an XOR-gate. Analogously, the output of *c* is 1 if that of *b* is 0.

According to our method, the inquirer now has to apply the first default rule: it is the only applicable one of form (II). It is inconsistent to believe that AND(a)&XOR(b): if *a* would be an AND-gate and *b* an XOR-gate, the output of *a*, given the inputs, would be 1; this is contrary to what we have observed. So, the only entry in the first list the inquirer compiles is (a,c), and the application of the rule results in a list of 16 system descriptions compatible with the default conclusion AND(a)&XOR(c) (if one assumes that AND(a)&XOR(c) is true, only the type of gate *b* must be determined to obtain a new complete system description).

The inquirer skips instructions (2) and (3). Following instruction (4), he checks which of the 16 candidates are falsified by the observation that falsified SD_0. This is easy: the falsified descriptions are those in which *b* is a gate which has output 1 if the first input is 1 and the second 0. This means that *b* must be a gate of one of the following types:

T_4	(IMPL)	T_7	(NOT-LEFT)
T_8	(EQ)	T_{10}	(RIGHT)
T_{12}	(AND)	T_{14}	
T_{15}	(NEITHER)	T_{16}	(CONTR)

Combined with the assumption that AND(a)&XOR(c), each of these eight possibilities yields a system description that is compatible with the observations the inquirer has made. He has no empirical evidence that allows him to make a choice between them, and no default rule to make a further selection. Instructions (5) and (6) of (M_3) tell the inquirer to make a further selection by means of measurements. He has to bring the inputs of the system in all possible states and register all the resulting outputs. For instance, he has to change the inputs into:

$$\text{input}_1(b) = 1 \;\&\; \text{input}_2(b) = 1 \;\&\; \text{input}_1(a) = 1.$$

Let's assume that for these inputs the same outputs are observed as in the original observation:

$$\text{output}(c) = 1 \;\&\; \text{output}(a) = 0$$

If the inquirer compares the eight possibilities with the results of this measurement, he ascertains that four of them are falsified: if *b* is an IMPL-gate, a RIGHT-gate, an AND-gate or an EQ-gate, then output(b) = 1, and thus output(c) = 0 and output(a) = 1. This contradicts the observations, so the inquirer must conclude that *b* is not an IMPL-gate, a RIGHT-gate, an AND-gate or an EQ-gate. On the other hand, if *b* is a NOT-LEFT-gate, a T_{14}-gate, a NEITHER-gate or a CONTR-gate, then output(b) = 0, and thus output(c) = 1 and output(a) = 0. This is exactly what we have observed, so the inquirer concludes that *b* is a NOT-LEFT-gate, a T_{14}-gate, a NEITHER-gate or a CONTR-gate. By performing more measurements and making the corresponding calculations, he can exclude all possibilities but one.

When we perform measurements it may happen that all the system descriptions we are considering are falsified. If our measurements lead to such situation, we have to use defaults of a lower level (instruction (2)) or resort to pure trial and error (instruction (3)). Suppose that when he changed the inputs into

$$\text{input}_1(b) = 0 \;\&\; \text{input}_2(b) = 0 \;\&\; \text{input}_1(a) = 1,$$

the inquirer has observed the following outputs:

$$\text{output}(c) = 1 \;\&\; \text{output}(a) = 1.$$

The eight possibilities are all falsified by this measurement: if output(b) = 0, then output(a) = 0; if output(b) = 1, then output(c) = 0. So whatever the output of *b*, one of the observed outputs must be 0. This contradicts what has been observed. The inquirer must go back to instruction (1). Because of the measurements, it is not consistent any more to believe that AND(a)&XOR(c). Before the measurements, it was already inconsistent to believe that AND(a)&XOR(b) or that XOR(b)&XOR(c). This means that no default of form (II) is applicable. The inquirer has to execute instruction (2). There are two applicable defaults of type (I):

For all gates *x*: $$\frac{\mathrm{SD_O} \vdash \mathrm{AND(x)} : \mathrm{AND(x)}}{\mathrm{AND(x)}}$$

For all gates *x*: $$\frac{\mathrm{SD_O} \vdash \mathrm{XOR(x)} : \mathrm{XOR(x)}}{\mathrm{XOR(x)}}$$

Application of these rules will result in three lists of 256 system descriptions. The first list contains the 256 system descriptions that are compatible with the assumption that AND(a), the second and third respectively the system descriptions that are compatible with XOR(b) and XOR(c). These 768 system descriptions must be compared with the results of the measurements. If all 768 system descriptions in these lists are falsified, instruction (3) must be executed.

3. *Why is the method useful*?

To show that the method is useful, I compare (M_3) with two alternatives. The first alternative is:

(M_A) (1) Take an arbitrary system description and check whether it is falsified by the original observation.
(2) Repeat (1) till a non-falsified system description is found. Accept this system description.

The second alternative is:

(M_B) (1) Compile a list of all possible system descriptions.
(2) Calculate which system descriptions are falsified by the original observation. Remove them from the list.
(3) Perform all possible measurements on the system. After each measurement, calculate which system descriptions are incompatible with the results and remove them. Accept the surviving system description.

If we use the first method the decision to accept a system description is taken without performing measurements. Unless we are very unlucky, only one or a few calculations must be made. So this method is fast and takes almost no effort. But it is very unreliable: the accepted system description may very well be false. The second method is slow and takes a lot of effort, because a large number of calculations are to be made. In our example, there are 16^3 possible system descriptions, so executing instruction (2) of (M_B) amounts to performing 4096 calculations. Further calculations, though diminishing in number, are necessary after each measurement. On the other hand, the second method is very reliable: if our observations are reliable and the gates behave determistically, it guarantees that the system description we accept is true. Besides the obvious differences between (M_A) and (M_B), there is also an important similarity: our grounds for accepting a system description are purely empirical. In the second method we gather more empirical evidence, but in both methods empirical evidence is the only ground for accepting or rejecting a system description. If, on the contrary, we use (M_3) to obtain a new system description, we will have a mixed back-up for the result. Our reasons to accept the new system description consist of observations (original observation and results of measurements) *and* default rules. The default rules are instruments for taking into account the non-empirical evidence supplied by the manufacturer, viz. the labels on the gates. Because the original system description, which was based on information provided by the manufacturer, has been falsified, we know that this information cannot be completely correct. But there is no reason to throw it away completely. The default rules save as much as possible of the information supplied by the manufacturer.

My method (M_3) is better than (M_B) because it is considerably faster, without loss of reliability. That the reliability of (M_3) is equal to that of (M_B) follows from the fact that no system description is accepted unless it has been checked against a complete set of measurements. If our observations are correct and the gates behave deterministically, (M_3) will result in a correct system description. In most cases, (M_3) is faster than than (M_B). If there is only one faulty gate, a default of type (II) will be applicable. The number calculations we have to make will be small (in our example: 16 in the first round, 8 in the second round, 4 in the third round, ...). If there are two faulty gates, at least one default of type (I) is applicable. The number of calculations will be larger (in our example, 768 in the first round after the application of the default rules of type (I)), but much smaller than for method (M_B). Only if all three gates are faulty, the number of calculations is of the same order as for (M_B).

Why is (M_3) better than (M_A)? It is more likely that only one component does not function in agreement with its label, than that the manufacturer has made two or three mistakes. So, the probability that there is one faulty gate (i.e. the probability that method (M_3) requires only a few calculations and is quite fast) is very high. The probability that there are three faulty gates (i.e. the probability of a case in which (M_3) is a very slow method) is low. This means that, in most cases, (M_3) is only a bit slower than (M_A).

Because (M_3) is fully reliable and (M_A) has an extremely low reliability, (M_3) is to be preferred to (M_A).

4. *Extension to more complex circuits*

If we consider circuits with three gates, default rules of the following form are useless:

(III) For all gates *x*, *y* and *z*:

$$\frac{SD_O \vdash T_\alpha(x)\&T_\beta(y)\&T_\gamma(z) : T_\alpha(x)\&T_\beta(y)\&T_\gamma(z)}{T_\alpha(x)\&T_\beta(y)\&T_\gamma(z)}$$

The reason is that $SD_O \vdash T_\alpha(x)\&T_\beta(y)\&T_\gamma(z)$ will be false if and only it is consistent to believe that $T_\alpha(x)\&T_\beta(y)\&T_\gamma(z)$. E.g., the original system description in our example entails that AND(a)&XOR(b)&XOR(c), but it is inconsistent to believe that AND(a)&XOR(b)&XOR(c). However, default rules of the form (III) become useful if the circuit contains more than three gates. A method for dealing with problems involving four-gate-circuits can be easily obtained by putting the following instruction in front of the instructions of (M_3):

> (1') Check whether there are applicable default rules of form (III). If there are, apply them all and go to instruction (4). If there are no applicable default rules of form (III), go to instruction (1).

Higher-level methods can be obtained in the same way: to obtain a method for level *n*+1, we put an instruction in front of the first instruction of the method for level *n*. The additional instruction refers to a default scheme that involves *n* gates.

As for (M_3), the speed of the higher-level methods depends on the degree of approximation of the original system description to the truth. The more we have reasons to believe that the original system description is approximately true, the more rational it is to apply (M_3) or a higher-level variant of it.

5. *General conclusions*

In section 2 I have developed a method for solving the problem outlined in the introductory section of this article. In section 3 I have defended this method, while in section 4 I have clarified how it can be generalized. The problem which I have used throughout this article, is a typical instance of a class of similar problems. These problems may be characterized as follows. An inquirer has a theory, SD_0, about how a system works. He tries to explain some observation, viz. an output of the system, by means of SD_0. However, he does not manage to show that the explanandum follows from SD_0 and the observed inputs. Instead, he ascertains that SD_0 and the observed inputs entail that the observed outputs cannot have occurred. The inquirer's original system description is falsified, and he starts looking for a new one. I think a general conclusion, pertaining to all problems of the type just outlined, may be drawn from the discussion in section 3. The conclusion is that, if we have a reason to believe that our original system description is approximately correct, it is rational to use a method which combines default reasoning and empirical tests, in the way illustrated by (M_3) and its generalizations. The defaults have the following form:

> If your original theory implies the hypotheses H_1,. .., H_n, and it is consistent to believe that H_1, ..., H_n are all true, then assume that H_1, ..., H_n are true.

If the original system description is approximately correct, methods which use such defaults will result in a fast and reliable problem solving process.

Universiteit Gent

REFERENCES

Reiter R. (1980), 'A Logic for Default Reasoning', *Artificial Intelligence* 13, pp. 81-132.