The faithfulness problem

Mario Bacelar Valente

Pablo de Olavide University
Seville, Spain
mar.bacelar@gmail.com
Orcid: 0000-0001-9473-5005

Abstract. When adopting a sound logical system, reasonings made within this system are correct. The situation with reasonings expressed, at least in part, with natural language is much more ambiguous. One way to be certain of the correctness of these reasonings is to provide a logical model of them. To conclude that a reasoning process is correct we need the logical model to be faithful to the reasoning. In this case, the reasoning inherits, so to speak, the correctness of the logical model. There is a weak link in this procedure, which we call the faithfulness problem: how do we decide that the logical model is faithful to the reasoning that it is supposed to model? That is an issue external to logic, and we do not have rigorous formal methods to make the decision. The purpose of this paper is to expose the faithfulness problem (not to solve it). For that purpose, we will consider two examples, one from the geometrical reasoning in Euclid's *Elements* and the other from a study on deductive reasoning in the psychology of reasoning.

Keywords: Euclidean proof; formal proof; deductive reasoning; suppression task

1. Logical model of Euclidean reasoning and the faithfulness problem

How can we be certain that our reasoning is correct? In fact, what could we mean by the correctness of reasoning? In this first part, we will address these issues in relation to a very specific subject: the reasoning in the mathematical proofs in the planar geometry of Euclid's *Elements*. For this work, we will only need to take into account one proof, that of proposition 1 of book 1 (proposition I.1).

Attaining certainty on the correctness of a reasoning process depends on how we define correctness. Here, we adopt a common view in which correctness is achieved by adopting a particular formal language and following its associated rules of inference. In this way, e.g., if we adopt propositional logic and adhere to its rules of inferences, we are certain that the reasonings made with propositional logic will be correct (see, e.g., Hedman 2004, p. 12-19).[1]

Now, in Euclid's proofs one adopts a highly regimented language, but a natural language nonetheless (Netz 1999, p. 89-167). Also, there seems to be a fundamental component of what we might call diagrammatic reasoning related to diagrams (Avigad; Dean; Mumma 2009). How can we determine the correctness of the reasonings?

As it is well-known, through history, doubts on the rigor of Euclidean proofs have been uttered (see, e.g., Venema 2012, p. 7-9). Here, we do not propose to address differences between the notions of rigor and correctness. One might even argue that even if we conclude that Euclidean proofs are not rigorous, they are nevertheless correct. For our purpose, it is enough to consider that doubts regarding the lack of rigor of Euclidean proofs can be further extended to the point of having doubts regarding the correctness of the proofs (or, at least, of lacking a rigorous way of showing the correctness of these unrigorous proofs).

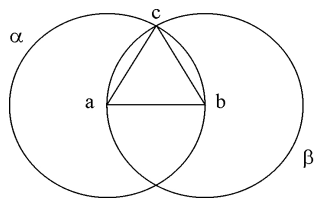
If we could model the reasonings in Euclid's proofs with formal logic, then we could conclude that the reasonings are correct. Avigad, Dean, and Mumma set forward a logical system they called *E* that, they claim, provides a faithful model of the proofs in Euclid's *Elements* regarding planar geometry (Avigad; Dean; Mumma 2009). What do they understand by faithful, and what does it imply? A model in *E* is faithful to the Euclidean proof when it reproduces line-by-line the "argumentative structure" (i.e., the reasoning) of the proof.[2] In particular, *E* mimics the inferences taken to be basic in the *Elements* (i.e., inferences made directly in one step without any further justification). In this

way, when Euclid deploys an inference in just one step, so does *E*; in the same way, when Euclid needs a chain of steps to deploy an inference, so does *E* (Avigad; Dean; Mumma 2009, p. 731).

There are relevant terminological differences between *E* and the regimented language of the *Elements*. That is taken not to impact on the faithfulness of *E*'s models of Euclidean proofs. One example is the meaning of the term line. With Euclid, the term line means line segment. In *E*, lines are, as usually defined in modern mathematics, non-bounded. This is seen as unproblematic since there is a "fairly straightforward translation between Euclid's terminology and [*E*'s]" (Avigad; Dean; Mumma 2009, p. 732). Another example is that the language of *E* does not include the term triangle. This can be addressed by a definitional extension of *E* that enables the definition of triangle from the primitive terms of the language of *E* (Avigad; Dean; Mumma 2009, p. 733). The view of the creators of *E* is that resorting to definitional extensions or other forms of "syntactic sugar" enables us to model more closely the Euclidean proofs (Avigad; Dean; Mumma 2009, p. 734). Accordingly, following the authors, a more precise formulation of the claim that a model in *E* is faithful to the corresponding Euclidean proof is:

> If we use a suitable textual representation of proofs [in *E*], then, modulo syntactic conventions like [the ones above], proofs in [the] formal system [*E*] look very much like the informal proofs found in the *Elements*. (Avigad; Dean; Mumma 2009, p. 714)

We will have to see in practice what "to look very much like" is taken to be. Let us first address the second part of the question above. Granted that the models of *E* are faithful to the Euclidean proof, what does this imply? Avigad, Dean, and Mumma showed that the logical system *E* is sound and complete. That has important consequences regarding the reasoning in Euclidean proofs. Taking into account the faithfulness of models of *E*, we may conclude that the proofs in the *Elements* are closer to formal proofs than one might previously think (Avigad; Dean; Mumma 2009, p. 760). From the perspective of the present work, the faithfulness of the models of *E* to the Euclidean proofs would make these "inherit" the rigor of *E*: the reasonings in the *Elements* (regarding planar geometry) would be sound in the precise sense that there are accurate models of these that are sound. By having a logical model faithful to the reasonings, we can argue that they are correct. We have a procedure to determine the correctness of Euclid's reasonings. We can be sure that they are right. But are we sure that the models are faithful? To put it a bit differently: how do we know with certainty that the models are faithful? To address this question, let us look at *E*'s model of the proof of proposition I.1:

> *Assume a and b are distinct points.*
> *Construct point c such that ab = bc and bc = ca.*
>
> 
>
>
> *Proof.*
> Let α be the circle with center *a* passing through *b*.
> Let β be the circle with center *b* passing through *a*.
> Let *c* be a point on the intersection of α and β.
> Have *ab* = *ac* [since they are radii of α].
> Have *ba* = *bc* [since they are radii of β].
> Hence *ab* = *bc* and *bc* = *ca*.
> Q.E.F. (Avigad; Dean; Mumma 2009, p. 734)

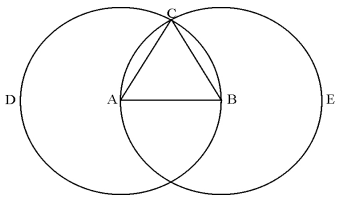
The terms "have" and "hence" are not part of the formal language, they are used to improve readability. In the same way, there are comments in brackets. Also, the drawn diagram is not part of *E*; it is included to improve the readability of the proof.[3]

The first line of the proof is the second construction rule of lines and circles (Avigad; Dean;

Mumma 2009, p. 716). The construction rules establish the accepted constructions in *E*; applying one of them corresponds to constructing an object. Some preconditions must be satisfied for the construction to be possible; also, the construction rules construct objects with some specified properties. In the case of rule 2, it establishes the construction of circles. For that, as a prerequisite, we need two points that do not coincide. That is our case since it is assumed that points *a* and *b* are distinct. The properties established by rule 2 are: *a* is the center of α and *b* is on α. In this way, the first line constructs a circle α with center *a* and with *b* on α. In the second line, another circle is constructed: the circle β with center *b* and with *a* on β. In the third line, there are two rules at play: one inference rule and one construction rule. First, we infer a diagrammatic assertion based on the available diagrammatic information. We do this by applying a rule that enables us to draw a conclusion from the premises. It is the rule 5 of diagrams rules for intersections (Avigad; Dean; Mumma 2009, p. 721). According to it, if *a* is on α, *b* is in α, *a* is in β, and *b* is on β, then α and β intersect. This rule is present implicitly on the third line since this line corresponds to rule 6 of the construction rules of intersections, in which the inferred conclusion of rule 5 – that α and β intersect – is a prerequisite of rule 6 (Avigad; Dean; Mumma 2009, p. 717). The property of the constructed point *c* is that *c* is on α and *c* is on β.

On line four it is asserted that the segments *ab* and *ac* are equal. In *E*, "segment" means the length of a line between two points. The comment in brackets is intended to indicate how the assertion was inferred. One applies the diagram-segment transfer rule 3. According to this rule, if *a* is the center of α and *b* is on α, then *ac* = *ab* if and only if *c* is on α, which is the case. On line five, one applies the same inference to conclude that the segments *ba* and *bc* are equal. Finally, on line six, one applies two metrical inferences – the symmetry of line segments and the transitivity of equality – to conclude that the segments *ab*, *bc*, and *ca* are equal (Avigad; Dean; Mumma 2009, p. 735). This concludes the proof in *E*.

A relevant aspect of Avigad, Dean, and Mumma's approach relates to how faithfulness is determined. It is not. We take for granted that the model is faithful to the Euclidean original. The authors explicitly write the following: "Since the point of this exercise is to demonstrate that proofs in *E* are faithful to the text of the *Elements*, we recommend comparing our versions with Euclid's." (Avigad; Dean; Mumma 2009, p. 734). That is, the faithfulness of the model is supposed to be self-evident by just checking the Euclidean text in relation to the *E*'s model. So, let us do that. The Euclidean text is as follows:

> *On a given finite straight line to construct an equilateral triangle.*
>
> Let *AB* be the given finite straight line.
> Thus it is required to construct an equilateral triangle on the straight line *AB.*
>
> 
>
>
> With center *A* and distance *AB* let the circle *BCD* be described; [Post. 3]
> again, with center *B* and distance *BA* let the circle *ACE* be described; [Post. 3]
> and from the point C, in which the circles cut one another, to the points *A, B*
> let the straight lines *CA, CB* be joined. [Post. I]
> Now, since the point *A* is the center of the circle *CDB, AC* is equal to *AB.* [Def. I5]
> Again, since the point *B* is the center of the circle *CAE, BC* is equal to *BA.* [Def. I5]
> But *CA* was also proved equal to *AB;* therefore each of the straight lines *CA, CB* is equal to *AB.*
> And things which are equal to the same thing are also equal to one another; [*C. N.* I]
> therefore *CA* is also equal to *CB.*
> Therefore the three straight lines *CA, AB, BC* are equal to one another.
> Therefore the triangle *ABC* is equilateral; and it has been constructed on the given finite straight line *AB.*
> (Being) what it was required to do. (Heath 1956, p. 241-242)

One can immediately notice that while in *E* we have the construction of a point *c* such that *ab* = *bc*, and *bc* = *ca* (where *a* and *b* are distinct points), in the *Elements* we have the construction of an equilateral triangle. This situation does not imply a lack of faithfulness on the part of the logical model. As mentioned, we can consider a definitional extension in which we define a triangle from primitive terms of *E*. Accordingly:

> Consider the [Euclidean] phrase "let *abc* be a triangle." Assuming we take this to mean a nondegenerate triangle, we parse this as saying that *a*, *b*, and *c* are points, and there are lines *L*, *M*, and *N*, such that *a* and *b* are on *L* but *c* is not, *b* and *c* are on *M* but *a* is not, and *c* and *a* are on *N* but *b* is not. (Avigad; Dean; Mumma 2009, p. 733)

We could include a new line at the end of E's model of the proof of proposition I.1, something like the following:

> *abc* is an equilateral triangle [by taking into account the definitional extension].

Apparently, there would be no lack of faithfulness due to the formulation of the model in terms of segments. But a closer look into the Euclidean reasoning shows that E's model is not faithful after all. As it is, in our view, lines 4 and 5 of the model are not faithful to the Euclidean reasoning. We will only address line 4 since they are equivalent. Line four consists of: have *ab* = *ac* [since they are radii of α]. The Euclidean counterpart of this line is: since the point *A* is the center of the circle *CDB, AC* is equal to *AB* [Def. I5]. As we have seen, the reasoning underlying line four consists in applying the diagram-segment transfer rule 3 (if *a* is the center of α and *b* is on α, then *ac* = *ab* if and only if *c* is on α). To be more exact, it consists in applying what Avigad, Dean, and Mumma call a direct consequence of the rule (Avigad; Dean; Mumma 2009, p. 725-7). We can formulate it somewhat as follows: if *a* is the center of α, and *b* is on α, and *c* is on α, then *ac* = *ab*. From line 1, we have that *a* is the center of α, and *b* is on α by construction. From line 3, we have that *c* is on α. The diagram-segment transfer rule 3 licenses us to infer that *ac* = *ab*. This reasoning, however, does not correspond to Euclid's thinking. In this respect, the comment in brackets – [since they are radii of α] – is misleading since it does not agree with the reasoning made in *E*. What is the corresponding reasoning in the *Elements*? Previous to concluding that *AC* is equal to *AB*, we draw the line segments *CA* and *CB*: let the straight lines *CA, CB* be joined. Then we resort to definition 15: a circle is a plane figure contained by one line such that all the straight lines falling upon it from one point among those lying within the figure are equal to one another (Heath 1956, p. 153). Definition 16 makes the previous one clearer: and the point is called the center of the circle (Heath 1956, p. 154). In the reasoning encapsulated in the proof, we have what we might call a component of diagrammatic reasoning. We draw two line segments connecting points *C* to *A* and *C* to *B*. We then see these line segments not merely as such but more specifically as radii of circle α. Taking into account the meaning of radii as given in definition 15, we then infer that they are equal. We could model this reasoning informally as follows:

> Diagrammatic reasoning: seeing *CA* and *CB* as radii of α, and not just as line segments.
> Applying a sort of universal elimination rule: All radii of a circle are equal
> *CA* and *CB* radii of α
> Then, *CA* = *CB*.

To be more faithful to the Euclidean reasoning, in our view, the model should have more parts corresponding to the construction of the line passing through *C* and *A* and the line passing through *C* and *B*. We then would apply an inference rule enabling us to take the segments to be radii of α, and, afterward, we would make another inference to conclude that they are equal because they are radii of the same circle (we would do the same for circle β). It could start by including something like the following:

> Let L_1 be the line through c and a.
> Let L_2 be the line through c and b. (construction rule 1 for lines and circles; Avigad; Dean; Mumma 2009, p. 716)

In *E,* segments are defined as the lengths of line segments from a point to another (Avigad; Dean; Mumma 2009, p. 710). To include the above construction rule enables to approach in *E* the procedure adopted in the Euclidean proof. We would explicitly construct the lines passing by *c* and *a*, and *c* and *b*. Afterward, instead of *ab* = *ac* [by a direct consequence of the segment transfer rule 3], we would have something like the following:

> (definitional extension of radii in *E*).
> inference: *ab* and *ac* are radii of circle α (here *ab* and *ac* are not just segments/lengths as defined in *E* but line segments as defined in the *Elements*).
> variant of transfer rule 3: if *ab* and *ac* are radii of circle α, then *ac* = *ab* (here, in the conclusion *ab* and *ac* return to being "simply" segments as defined in *E* – the length of the line segments connecting points *a* and *b* and *a* and *c*).

This would correspond to the application of a variant of the segment transfer rule 3 (if *a* is the center of α and *b* is on α, then *ac* = *ab* if and only if *c* is on α), in which we would make use of a definitional extension of radii of a circle. The new inference would correspond to the Euclidean practice of seeing an object in different ways (Macbeth 2010). In this case, we would model seeing *ab* and *ac* not merely as the segments we have just constructed connecting the points but as radii of the circle α. This "seeing an object in different ways" occurs throughout the Euclidean proof. After seeing line segments as radii and concluding that, because of this, they are equal, one returns to see them as "just" line segments and concludes by resort to common notion 1 that the line segments CA and CB are equal. From this, one concludes that the line segments "*CA, AB, BC* are equal to one another" (Heath 1956, p. 241). Until this moment, there is no mention of the notion of a triangle. However, immediately after this line of the proof one concludes: "Therefore the triangle *ABC* is equilateral" (Heath 1956, p. 242). For this to be the case, we reason in the diagram as Macbeth puts it (Macbeth 2010, p. 265): we actively go beyond seeing three line segments (proved to be metrically equal) as just line segments to see them as the sides of a triangle. Being metrically equal, we conclude that the triangle is equilateral.

One might argue that even if *E*'s model is not faithful in this part, there is no harm done. But we would need an argumentation that shows that this partial lack of faithfulness does not affect the inheritance of soundness on the part of the Euclidean proof from the model. This argumentation would be made outside logic; it would be informal. This would be another instance of the faithfulness problem: how do we show that the model is faithful enough to guarantee the soundness of the Euclidean reasoning?

Returning to our proposition of a more faithful model of the proof of proposition I.1, we are also stuck on the faithfulness problem: how can we decide with certainty that (a completed version of) this model is faithful? There seems to be no certainty in our claim of the faithfulness of the model. Without this certainty, there is no certainty on the soundness of the Euclidean reasoning also. We might say that our gut instinct is that the model is faithful, and because of this, we see that the Euclidean reasoning is sound. But gut instinct is not enough.

2. Logical models of natural language deductive reasoning and the faithfulness problem

We also face the faithfulness problem when addressing human reasoning more generally, not expressed in terms of a regimented language like that of the mathematical practice of the *Elements*. We will address deductive reasoning; that is, the reasoning expressed with natural language in which the form of the argument guaranties that it is valid. One form of a valid argument is the *modus ponens* (Evans 2005). One example of this kind of argument is the following: If it is raining then the ground is wet; it is raining; so, the ground is wet. Using a schematic formulation, the pattern of a *modus*

ponens argument is as follows: if the first, then the second; but the first; so, the second (Novaes 2012, p. 72).

Here, we will consider a particular experimental study of deductive reasoning, the so-called Byrne's suppression task (Byrne 1989); we will focus just on the part relating to *modus ponens*. The participants in the experiment were given a set of premises, and their task was to choose one of three proposed conclusions. They were told that the premises were true. The basic premises were:

> If she has an essay to write then she will study late in the library.
> She has an essay to write.

There were three possible conclusions to choose from:

> (a) She will study late in the library.
> (b) She will not study late in the library.
> (c) She may or may not study late in the library.

A group of participants was faced with the above premises. Of these, 96% of them choose the conclusion (a). This corresponds to a *modus ponens* argument. We can say that they adopted a pattern of a *modus ponens* argument according to the schematic formulation above.

A second group must undertake their reasoning task considering also what Byrne calls an alternative antecedent – an antecedent that could elicit the same conclusion (Byrne 1989, p. 65):

> If she has an essay to write then she will study late in the library.
> If she has some textbooks to read then she will study late in the library.
> She has an essay to write.

In this case, it was obtained the same percentage as with the first group. The presence of another premise did not affect the reasoning; it mainly – for 96% of the participants – corresponds to a *modus ponens* argument.

Finally, A third group was given the initial premises together with what Byrne calls an additional antecedent – an antecedent that "refers to some additional requirement that must also hold" (Byrne 1989, p. 67):

> If she has an essay to write then she will study late in the library.
> If the library stays open then she will study late in the library.
> She has an essay to write.

In this case, only 38% of the participants in this group arrive at the conclusion (a). From a (classic) logical point of view, like in the case of the so-called alternative premise, we should have something like: if p then q or if r then q; p; then q ($p \rightarrow q \vee r \rightarrow q$; p; then q). The additional premise should not affect the reasoning if this is made strictly by considering the logical form of the argument (as prescribed in classical logic).

According to Byrne, the additional premise leads to a suppression of the *modus ponens* argument: due to the context (the presence of an additional premise), the participants are rejecting instances of the valid *modus ponens*.

This result would imply that the mental inferences underlying human reasoning expressed with natural language, even in the case of deductive reasoning, do not comply with logical rules of inference. There would be no *modus ponens* inferences underlying what we might expect to be *modus ponens* arguments. That is so because, in many cases, where we should have a *modus ponens* argument we face a *modus ponens* suppression, and we have a different conclusion. Byrne takes her result as indicating that we do not reason with a mental logic (Byrne 1989). In simple terms, mental logic corresponds to the idea that we reason according to logical rules. One example is *modus ponens*; we would have a logical inference rule system in our minds, literally, and there would exist a *modus ponens* inference (Manktelow 2012, p. 43-46).

Byrne's conclusion was challenged by Stenning and van Lambalgen (2004b). Before addressing their approach, it is important to clarify from the start that Stenning and van Lambalgen do not propose some sort of mental logic. To the best of our knowledge, this aspect of their approach is made clearer in a book by Novaes:

> Stenning and van Lambalgen offer extensive modelling of human reasoning in terms of this framework, but I take it that they do not mean to claim that the very syntactical rules described by the framework are actually and precisely implemented when people reason. Instead, as I read them, the formalism is presented as a model of the phenomena in question, just as a physical theory is a model of physical reality: an approximate description, not the 'real thing'. (Novaes 2012, p. 142)

> In personal communication, Stenning clarifies that they take at least some aspects of the formalism to be accurate representations of psychological phenomena. For example, the formalism does presuppose an asymmetry between positive and negative information, and there are reasons to think that this asymmetry is a real psychological phenomenon (e.g., the discrepancy in reasoning competence with *modus ponens* v. *modus tollens*, which is naturally accounted for in terms of such an asymmetry). (Novaes 2012, p. 142)

I will address Stenning and van Lambalgen`s approach in a way equivalent to the logical system E – as providing a logical model, in this case, of a reasoning task. There is an important difference between E and the logical system proposed by Stenning and van Lambalgen. In the first case, we always have the same logical system E that is applied to all the Euclidean reasonings under consideration. In the case of Stenning and van Lambalgen, we have a more general logical framework that is made more specific for each participant: we model each participant's interpretation of the reasoning task with a particular variant of the logical system. Initially, we model each participant's reasoning to a particular interpretation of the premises (corresponding to reaching a specific setting of the model). Only afterward do we have the modeling of the inferences of the participant using the specific variant of the logical system. We can refer to these two steps as reasoning to an interpretation or model of the premises, and reasoning from this fixed interpretation or model (Varga; Stenning; Martignon 2015; Stenning & van Lambalgen 2004b).

That leads to a completely different view on the results of the suppression task. In Byrne's case, we face a *modus ponens* suppression, conceived as a failure to apply classical logic and leading to a non-sound reasoning. We can now conceive this as the adoption by the participant of a reasoning pattern that is sound according to the variant of the logical system – the specific logical model – that we take to be faithful to the participant's reasoning.

The main characteristic of the general logical framework adopted by Stenning and van Lambalgen is how a conditional "if p then q" is represented. It has the form $p \wedge \neg ab \rightarrow q$, which we can read as "if p and nothing is abnormal, then q"; ab stands for an abnormality that would lead to an exception: in the case of an abnormality, we cannot infer q from p; it blocks the inference. One takes the conditional formulas to have conjoined abnormality conditions with the form $r_1 \rightarrow ab_1, \ldots, r_n \rightarrow ab_n$. When there is evidence of some r_i, we take it to be the case that we have the abnormality ab_i. This is an important aspect of this logical framework, which corresponds to the adoption of the closed world assumption: if there is no positive evidence for a proposition, we can conclude that it is false; concerning an abnormality ab, this means that if there is no positive evidence for ab then we conclude that $\neg ab$ is true. In this case the logical form of the conditional reduces to $p \rightarrow q$ (Stenning & Lambalgen 2010, p. 6; Stenning & Lambalgen 2008, p. 184; Besold et al. 2017, p. 45-46).

Let us see Stenning and van Lambalgen's approach at work in the case of Byrne's suppression task. The conditional "If she has an essay to write then she will study late in the library" is represented by the formula $p \wedge \neg ab \rightarrow q$. Both the conditionals "if the library stays open then she will study late in the library" and "if she has some textbooks to read then she will study late in the library" are represented by a formula of the form $r \wedge \neg ab' \rightarrow q$.

In this case, modeling a participant's reasoning to an interpretation of the premises is made by adjusting the meaning of the abnormalities in the previous general formulas. That leads to taking into account, if that is the case, some abnormality conditions. Afterward, it is modeled the reasoning from

the resulting fixed model.

A model consistent with a *modus ponens* argument in the first group has, simplifying, the clauses $\{p; p \wedge \neg ab \rightarrow q\}$. There is no information leading to consider that we have an abnormality. That implies that we have $\{p; p \rightarrow q\}$. In this case, the setting of the model is finalized by replacing $\rightarrow$ by the classical biconditional $\leftrightarrow$ (Besold et al. 2017, p. 47). The end result of this modeling of the participant's reasoning to an interpretation is $\{p; p \leftrightarrow q\}$. The reasoning from this interpretation starts from the logical form $p \leftrightarrow q$ and the premise p, and derives q (Stenning & Lambalgen 2008, p. 197).

Let us now see a model consistent with the suppression of the *modus ponens* argument by a majority of the third group's participants. Besides the conditional clause of the first premise $p \wedge \neg ab \rightarrow q$ (p = "she has an essay to write", q = "she will study late in the library"), we also have a clause representing the additional premise: $r \wedge \neg ab' \rightarrow q$, where r = "the library stays open". Also, the additional premise makes salient the possibility of an abnormality represented in the model by the abnormality condition $\neg r \rightarrow ab$ (Stenning & Lambalgen 2019, p. 7-8; Stenning & Lambalgen 2008, p. 198).

The reasoning to an interpretation starts with a set that contains p, $p \wedge \neg ab \rightarrow q$, $r \wedge \neg ab' \rightarrow q$, and $\neg r \rightarrow ab$. This reduces to $\{p; (p \wedge r) \leftrightarrow q\}$. To be able to infer q from $(p \wedge r) \leftrightarrow q$ ("if she has an essay to write and the library stays open then she will study late in the library") we would need to have as a premise, besides p ("she has an essay to write"), also r ("the library stays open"). According to Stenning and van Lambalgen, "the reasoning from an interpretation is now stuck in the absence of information about r" (Stenning & Lambalgen 2008, p. 198).

This situation does not occur with the second group. In this case, as we have seen, the alternative conditional ("if she has some textbooks to read then she will study late in the library") is also formalized as $r \wedge \neg ab' \rightarrow q$. According to Stenning and van Lambalgen, "by general knowledge, the alternatives do not highlight possible obstacles" (Stenning & Lambalgen 2008, p. 199). As they mention elsewhere, "[the] integration of the third premise does not lead to the addition of information on ab or ab'" (Stenning & Lambalgen 2004a, p. 20-21). In this way, there are no possible abnormalities, and the reasoning to an interpretation fixes the model $\{p; p \vee r \leftrightarrow q\}$. Reasoning from this interpretation/model derives q (Stenning & Lambalgen 2008, p. 199).

From what we have just seen, it is evident that the general framework proposed by Stenning and van Lambalgen is flexible enough to provide models of reasoning compatible with the results in the suppression task with the three groups. But do these models correspond in any way to the actual reasonings of the participants? As it is, this could be an *ad hoc* way of fitting to the experimental results (the choice of the conclusion by each participant). What is at stake is the faithfulness of the models to the actual reasonings.

As Stenning and van Lambalgen mention, regarding another reasoning task, one needs a controlled experiment to provide evidence that the reasoning does take place as modeled (Stenning & Lambalgen 2008, p. 59). For that purpose, after each participant undertakes the reasoning task, they ask him or her for a justification of the chosen conclusion (Stenning & Lambalgen 2004b, p. 40). This unfolds in the form of a dialogue that is supposed to bring some light on the participant's reasoning when making his or her choice. Let us consider two excerpts of dialogues. The first is taken as evidence for the modeling of the suppression of *modus ponens*:

> *Subject* 2.
> *S*: Ok yeah I think it is likely that she stays late in the library tonight, but it depends if the library is open. . . so perhaps I think [pauses]. yeah, in a way I think hmm what does it say to me? I mean the fact that you first say that she has an essay to write then she stays late in the library, but then you add to it if the library stays open she stays late in the library so perhaps she's not actually in the library tonight, because the library's not open. I don't think it's a very good way of putting it.
> *E*: How would you put it?
> *S*: I would say, if Marian has an essay to write, and the library stays open late, then she does stay late in the library. (Stenning & Lambalgen 2008, p. 204)

According to Stenning and van Lambalgen, *Subject* 2's answer is accounted straightforwardly by the

logical model. The conditional has the form $(p \wedge r) \rightarrow q$. In this way, the *modus ponens* is suppressed, unless the premise r ("the library stays open late") is included; in this case, r together with p ("Marian has an essay to write") licenses the inference that q ("she does stay late in the library") (Stenning & Lambalgen 2008, p. 204).

Stenning and van Lambalgen give the following excerpt as an example of evidence for the adoption of the closed world assumption:

> *Subject* 7.
> *S*: . . . that she has to write an essay. because she stays till late in the library when she has to write an essay, and today she stays till late in the library.
> *E*: Could there be other reasons for her to stay late in the library?
> *S*: That could be possible, for example, maybe she reads a very long book. *But as I understand it she stays late in the library only if she has to write an essay*. (Stenning & Lambalgen 2008, p. 205)

In Stenning and van Lambalgen's interpretation of the dialogue, "the italicized phrase seems to point to closed-world reasoning" (Stenning & Lambalgen 2008, p. 205).

Stenning and van Lambalgen consider that seven out of ten participants behave according to the logical model (Stenning & Lambalgen 2008, p. 212); however, they are aware of the limitations of using dialogues. Accordingly:

> We do not interpret these dialogues as *reports* of reasoning that went on before the dialogue, let alone as transparent and complete reflections of such preceding thought processes. These dialogues *are* the subjects' reasoning with a tutor during a dialogue. Engaging subjects in dialogue undoubtedly changes their thoughts, and may even invoke learning. The relation between the reasoning processes evoked by the standard way of conducting the task, and the processes reflected in subsequent dialogues is a relation that remains to be clarified. (Stenning & Lambalgen 2001, p. 280)

Elsewhere they also remark the following:

> We acknowledge that we cannot be certain that our interpretations of the dialogues are correct representations of mental processes – the reader will often have alternative suggestions. (Stenning & Lambalgen 2008, p. 59)

We face two layers of the faithfulness problem with logical models of reasoning tasks. In the case of the logical model of Euclidean reasoning we had only one: we cannot be certain that the model is faithful to the reasoning as expressed in the proof. The situation here is more complex. Here, we also face the issue of the participant's reconstruction of his or her reasoning. As Stenning and van Lambalgen rightly point to, it is unclear what is the actual relation between the participant's reconstruction expressed in the dialogue and the earlier reasoning. We do not have this problem in the modeling of Euclidean reasoning. What we call Euclidean reasoning is expressed in the proof. We are modeling the proofs while taking them to express an underlying reasoning process. It is here that we face the faithfulness problem: how can we be sure that our model is faithful to the Euclidean proof (as practiced by Euclid)? With logical models of reasoning tasks, we also have this layer of the faithfulness problem. Stenning and van Lambalgen acknowledge that they cannot be certain of their interpretation of the dialogues. That is, even if we took for granted that a dialogue expresses the actual reasoning of a participant, we cannot be sure that we are making the correct interpretation of the dialogue. In this way, in the logical modeling of a reasoning task the faithfulness problem is two-fold: (1) we are not certain that the dialogues express the reasonings of the participants; (2) we are not certain of making the correct interpretations of the dialogues (someone else will often have different interpretations). Without this, we only have logical models that are compatible with the choices of conclusions made by the participants.

By construction, the natural language conditionals, arising from the interpretation of the dialogues, adopted by Stenning and van Lambalgen, correspond to logical conditionals. For example, in the case of the suppression of *modus ponens*, Stenning and van Lambalgen propose the logical model $(p \wedge r)$

$\rightarrow q$ (as the result of the reasoning to an interpretation); this corresponds in *Subject* 2's dialogue to the phrase "if Marian has an essay to write, and the library stays open late, then she does stay late in the library". Stenning and van Lambalgen take this phrase to be accounted straightforwardly by the logical conditional "$\rightarrow$" (and connective "$\wedge$"), so we can consider this phrase as a natural language conditional with which the participant expresses his or her reasoning.

Regarding *Subject* 2's reasoning, Stenning and van Lambalgen take the dialogue as evidence for *Subject* 2's reasoning to an interpretation according to their model, such that *Subject* 2 arrives at the interpretation modeled by $(p \wedge r) \rightarrow q$. Again, like in the case of the *E*'s model, we might say that our gut instinct is that Stenning and van Lambalgen are right in the case of this particular participant (at least regarding the interpretation of the dialogue). However, in general, we are not certain that we are making a rigorous interpretation of the dialogue concerning the participant's reconstruction of his or her reasoning (since we have no formal method to attest this). Neither are we certain that the dialogue corresponds in any clear way to the reasoning of the participant.

3. Further comments

We expect the particular cases of the faithfulness problem we have addressed in detail here not to be the exception but the rule. That is, when developing a logical model of some form of reasoning we expect there to be difficulties in ascertaining the faithfulness of the model.[4]

In our view, in logic literature, there is a clear example of the faithfulness problem that logicians have been addressing without considering that this is a particular case of a much vaster issue. It regards translating a natural language sentence or argument into the formal language of a logic; for example, the sentence "Donald embraced Orman at noon" or the argument "All horses are animals. $\therefore$ All heads of horses are heads of animals" (Michels 2021). In this case, "the correctness of a formalization can never be a completely formal matter, i.e. logic alone can never tell us whether a formula is a correct formalization of a sentence" (Michels 2021, p. 16).

Logicians have tried to put forward criteria for the adequateness of formalizations (see, e.g., Brun 2014 and Peregrin & Svoboda 2017). However, issues have been raised regarding the coherence of formalizations (Dicher 2021) and regarding limitations in the proposed criteria (Reinmuth 2021).

While we are agnostic regarding how logicians are facing the problem of the adequateness of formalizations (which for us is one more example of the faithfulness problem), we do not expect that the very specific approaches they are developing for the case of the formal rendering of sentences or arguments in natural language to be applicable in very different situations, like the two cases we have exposed here.

While our intention in the present work is just to expose the faithfulness problem as a possibly very generalized problem, we will sketch some directions which could be explored to give a "solution" to the faithfulness problem. We must notice that we cannot attain absolute certainty – the kind of certainty we have with logic – when facing the faithfulness problem. This much is implicit in logicians' treatment of the adequateness of formalizations problem. Paraphrasing the above citation, the faithfulness of a logical model can never be a completely formal matter. As such, it is a metalogical issue and needs tools outside logic to be addressed.

What is our second-best option after logic/absolute certainty? Our view is that we should apply the methods of science to face the problem. That is, we might aim to attain what we might call a scientific certainty.

Regarding Stenning and van Lambalgen's logical model, we can fully adopt the view set forward by Novaes mentioned in the previous section: to take the logical model "as a model of the phenomena in question, just as a physical theory is a model of physical reality" (Novaes 2012, p. 142). The faithfulness of the model becomes a scientific question to be addressed by scientific methods, in particular experimentation.[5]

Little is known about the underlying neurophysiological phenomena of the suppression task, and how much we might – as an empirical question – take Stenning and van Lambalgen's formalism to

model the phenomena. However, there is at least one experimental work whose results can be read as evidence that effects of suppression occur as predicted by the logical model (Pijnacker et al. 2010). This one experiment shows that we are already at a point where we can start to address the faithfulness of the model as an empirical issue.

Regarding our first case study, the situation is more cumbersome. Science is still far from providing empirical "criteria" to address the faithfulness of *E*'s models. What could be, for the time being, our third-best option after logic certainty and scientific certainty?

We think our best option is to consider whatever scientific results we might have available (e.g., Hamami; Mumma; Amalric 2021) together with specialized philosophical approaches. We already have many philosophical "results" regarding the reasoning underlying Euclidean proofs, like, e.g., Manders (2008), Macbeth (2010), and Dal Magro & García-Perez (2019). In fact, Manders' work has been taken into account in the development of the formal system *E* (Avigad; Dean; Mumma 2009). Our tentative proposition is to include the issue of the faithfulness of models of *E* (or other formal systems) in our philosophical inquiries into Euclidean proofs.

For the time being we think it is best to have a pluralistic approach to the faithfulness problem, adopting the best available "tools" for each case where we face issues regarding the faithfulness of the logical model.

4. Conclusions

Logic provides a powerful formalism to address the correctness of reasonings. Within logic itself, the soundness of inferences is not subjected to doubt, in the sense that for every logical system we have a collection of sound rules of inference. Outside logic, if we try to address the correctness of reasonings expressed with natural language, we face enormous difficulties due to the lack of a formal approach to address it. One way to deal with this difficulty is to envisage logical models of the reasoning under study. If we can find logical models of the reasoning, then we might say that the reasoning is correct or sound in the sense of having a sound logical model. But for this to be the case, we really must have a logical model of the reasoning. That is, the model must be faithful to the reasoning that it models. In this work we consider two examples of reasonings, the Euclidean reasoning in the proofs on planar geometry in the *Elements,* and the reasoning in Byrne's suppression task. In the case of the Euclidean reasoning, a logical model has been proposed by Avigad, Dean, and Mumma. In the case of the reasoning task, a logical model has been proposed by Stenning and van Lambalgen. In both cases, issues can be raised concerning the faithfulness of these models. The purpose of the present work is to call the attention to what we have called the faithfulness problem, which we suspect to be a generalized issue in logical modeling, by using these two logical models as examples. Like in the case of these two examples, the general case might be that we have no way to decide with (absolute) certainty that a logical model is faithful to the reasoning it is supposed to be modeling.[6]

Acknowledgments

I want to thank the anonymous reviewers for their constructive commentaries and suggestions.

References

Adam, C.; Herzig, A.; Longin, D. 2009. A logical formalization of the OCC theory of emotions. *Synthese* 168: 201-248.

Avigad, J.; Dean, E.; Mumma, J. 2009. A formal system for Euclid's *Elements*. *The Review of Symbolic Logic* 2: 700-768.

Besold, T. R.; Garcez, A. D.; Stenning, K.; van der Torre, L.; van Lambalgen, M. 2017. Reasoning in non-probabilistic uncertainty: Logic programming and neural-symbolic computing as examples. *Minds and Machines* 27: 37-77.

Byrne, R. M. J. 1989. Suppressing valid inferences with conditionals. *Cognition* 31: 61-83.

Brun, G. 2014. Reconstructing arguments: Formalization and reflective equilibrium. *Logical Analysis and History of Philosophy* 17: 94-129.
Dal Magro, T. & García-Perez, M. J. 2019. On Euclidean diagrams and geometrical knowledge. *Theoria. An International Journal for Theory, History and Foundations of Science* 34(2): 255-276.
Dicher, B. 2021. Reflective equilibrium on the fringe. *Dialectica* 74(2): 71-94.
Evans, J. 2005. Deductive reasoning. In: K. J. Holyoak & R. G. Morrison (eds.), *The Cambridge handbook of thinking and reasoning*, p. 169-184. Cambridge: Cambridge University Press.
Hamami, Y.; Mumma, J.; Amalric, M. 2021. Counterexample search in diagram-based geometric reasoning. *Cognitive Science* 45(4): e12959
Heath, T. H. 1956. *The thirteen books of Euclid's Elements*, second edition unabridged. New York: Dover Publications.
Hedman, S. 2004. *A first course in logic*. Oxford: Oxford University Press.
Macbeth, D. 2010. Diagrammatic reasoning in Euclid's *Elements*. In: B. v. Kerkhove; J. de Vuyst; J. P. v. Bendegem (eds.), *Philosophical perspectives on mathematical practice*, p. 235-267. London: College Publications.
Manders, K. 2008. The Euclidean diagram. In: P. Mancosu (ed.), *The philosophy of mathematical practice*, p. 80-133. Oxford: Oxford University Press.
Manktelow, K. 2012. *Thinking and reasoning: an introduction to the psychology of reason, judgment and decision making*. London: Psychology Press.
Martin, B. & Hjortland, O. 2021.Logical predictivism. *Journal of Philosophical Logic* 50: 285-318.
Michels, R. 2021. The formalization of arguments. *Dialectica* 74(2): 1-33.
Netz, R. 1999. *The shaping of deduction in Greek mathematics*. Cambridge: Cambridge University Press.
Novaes, C. D. 2012. *Formal languages in logic: A philosophical and cognitive analysis*. Cambridge: Cambridge University Press.
Payette, G. & Wyatt, N. 2018. How do logics explain? *Australasian Journal of Philosophy* 96(1): 157-167.
Peregrin, J. & Svoboda, V. 2017. *Reflective equilibrium and the principles of logical analysis*. New York: Routledge.
Pijnacker, J.; Geurts, B.; van Lambalgen, M.; Buitelaar, J.; Hagoort, P. 2010. Reasoning with exceptions: an event-related brain potentials study. *Journal of Cognitive Neuroscience* 23: 471-480.
Reinmuth, F. 2021. Holistic inferential criteria of adequate formalization. *Dialectica* 74(2): 115-149.
Sattler, U.; Calvanese, D.; Molitor, R. 2007. Relationships with other formalisms., In: F. Baader; D. L. McGuinness; D. Nardi; P. F. Patel-Schneider (eds.), *The description logic handbook*, p. 142-183. Cambridge: Cambridge University Press.
Stenning, K. & van Lambalgen, M. 2001. Semantics as a foundation for psychology: A case study of Wason's selection task. *Journal of Logic, Language and Information* 10: 273–317.
Stenning, K. & van Lambalgen, M. 2004a. *Evolutionary considerations on logical reasoning*. Retrieved from https://www.researchgate.net/publication/2896327_Evolutionary_Considerations_on_Logical_Reasoning. Access: 29/09/2021.
Stenning, K. & van Lambalgen, M. 2004b. *A working memory model of relations between interpretation and reasoning*. Retrieved from https://www.researchgate.net/publication/239015102_A_working_memory_model_of_relations_between_interpretation_and_reasoning. Access: 29/09/2021.
Stenning, K. & van Lambalgen, M. 2008. *Human reasoning and cognitive science*. Cambridge: MIT Press.
Stenning, K. & van Lambalgen, M. 2010. The logical response to a noisy world. In: M. Oaksford & N. Chater (eds.), *Cognition and conditionals: Probability and logic in human thinking*, p. 85-102. Oxford: Oxford University Press. (Version adopted: retrieved from https://www.researchgate.net/publication/286784975_The_logical_response_to_a_noisy_world. Access: 29/09/2021.)
Stenning, K. & van Lambalgen, M. 2019. Reasoning and discourse coherence in autism spectrum disorder. In: K. Morsanyi & R. Byrne (eds.), *Thinking, reasoning, and decision making in autism*, p. 135-155. London: Routledge. (Version adopted retrieved from https://www.researchgate.net/publication/330324480_REASONING_AND_DISCOURSE_COHERENCE_IN_AUTISM_SPECTRUM_DISORDER. Access: 29/09/2021.)
Varga, A.; Stenning, K.; Martignon, L. 2015. *There is no one logic to model human reasoning. The case for interpretation*. Conference Paper. Retrieved from https://www.researchgate.net/publication/280114264_There_is_no_one_logic_to_model_human_reasoning_The_case_for_interpretation. Access: 29/09/2021.
Venema, G. A. 2012. *Foundations of geometry*. Boston: Pearson.

[1] Unless stated, we will take certainty to mean the "absolute" certainty provided by logic.

[2] Here, we adopt Novaes' view that "a formal language […] can characterize directly the target phenomena without the mediation of ordinary languages" (Novaes 2012, p. 99). In this way, the idea of logic as translating natural language statements (being a model of these) might best be understood as a metaphor (unless we are considering the particular case where we are directly modeling natural language sentences). In this part of the paper, we take logic to (try to) model the reasonings expressed in the argumentative structure. In this way, we will be liberal in our terminology and sometimes speak of models of Euclidean proofs others of models of Euclidean reasonings.

[3] According to the authors, "in *E* the diagram *is* nothing more than the collection of generally valid diagrammatic features that are guaranteed by the construction. In other words [...] we identify the diagram with the information provided by [a] construction […] and all the direct diagrammatic consequences of these data" (Avigad; Dean; Mumma 2009, p. 706).

[4] And this can be the case also when the modeling is not of human reasoning. One example is the modeling with

Description Logics of data models used in databases; more specifically, the modeling of the Entity-Relationship (ER) model. Accordingly, "several features of the ER model and desired reasoning tasks could not fully be captured by the proposed translation" (Sattler; Calvanese; Molitor 2007, p. 168). Another example might be the logical modeling of the OCC theory of emotions. It is claimed concerning the proposed logical modeling its "faithfulness to the OCC theory" (Adam; Highers; Longin 2009, p. 513). Considering the difficulties, we have found in our two case studies, we feel that the putative faithfulness of the modeling of the OCC theory needs more thorough scrutiny.

[5] In our view, this corresponds to adopting an anti-exceptionalist stance on logic. On anti-exceptionalism, see, e.g., Payette & Wyatt (2018) and Martin & Hjortland (2021).

[6] According to the view sketched in the previous section, this implies that we cannot, without somehow addressing faithfulness issues, use the existence of a logical model to decide with (some sort of) certainty that the reasoning being modeled is correct. For example, if we were to attain scientific certainty on the faithfulness of the logical model of a reasoning task, we could use the existence of this model to decide with scientific certainty that a participant's reasoning is correct.