Geometric proof checking with diagrams

John Mumma[1]

Abstract. The paper presents an account of the diagrammatic arguments of Euclid's *Elements* as an example of a proof checking process that is guided by specific mathematical concepts rather than universal, abstract logical rules.

1 INTRODUCTION

It's an undeniable sociological fact that a mathematicians result is not accepted until other mathematicians have checked the proof for it. The philosophical question connected to this fact is how the checking process ought to be understood. If one does not take arbitrary convention and caprice to be an option, the problem is to provide a characterization of the standards by which a proof is judged to be correct. What guides the mathematician in checking whether a proposed proof contains gaps?

A natural answer, so natural that it is often simply assumed by philosophers, is a logical ideal of proof. An ideal proof, accordingly, is a sequence of sentences, where the initial sentences express the proof's assumptions, the last sentence its conclusion, and every sentence that is not an assumption is linked to earlier one via a logical inference. What a mathematician does in checking a proof is determine whether it can resolved into such a list of sentences. A gap detected in a proof is a gap in the sequence.

Though this provides a picture that seems generally accurate, it becomes becomes strained once one looks into what exactly the logical ideal demands. Proofs that come to be accepted in practice are not formal in a logical sense, and are a long way from being so. Transforming them into an airtight logical sequence takes a good deal of work. What is taken as obvious in practice expands into long tracts of text in fully explicit logical analyses.[2] And so, though the logical ideal might be part of the story, it cannot (as asserted by the tile of [5]) be the whole story. An account is necessary of the proof steps that for a mathematician are unproblematic and obvious, but require by the logical ideal further justification.

One possible explanation for the disparity is that while logical rules of inference are universal—i.e. their validity is independent of the content of what's inferred—mathematicians allow inferences that depend on the mathematical concepts involved in proofs. As it is, the thesis is too general to be interesting, however. What is needed to substantiate it is an account of how mathematical concepts shape standards of proof for theorems concerning those concepts. My aim with this paper is to provide such an account for the mathematics of elementary geometry. I present some recent work on Euclid's diagrammatic proof method with the goal of showing how, precisely, knowledge of a proof's specific mathematical concepts and techniques can ground judgments on the existence or non-existence of gaps in it.

[1] email:mumma@stanford.edu

[2] This can be seen in work in formal proof verification. The authors of [1] identify the biggest obstacle in their work to be "the gap between those inferences that ordinary mathematicians recognize as obvious, and those that can be verified automatically by conventional proof assistants."

2 EUCLID'S REGIMENTED USE OF DIAGRAMS

Any given geometric diagram manifests a wealth of observable spatial properties. This no doubt is a major reason behind the standard view that such objects cannot used to present rigorous geometrical arguments. In order for a proof to be checked, the information in it must be effectively controlled. It must at no point be unclear what justifies a step. The use of a diagrams as geometric representations might seem to make such a demand impossible to meet. The worry is that once diagrams with their rich array of spatial properties are allowed to represent geometric objects, the ability to isolate the standing of each claim in a geometric argument is compromised. Observing something in a diagram is too complicated, opaque a process to incorporate into mathematically controlled arguments.

This worry loses much of its force in the light of Ken Manders' investigations into Euclid's geometric proofs in [3]. His analysis reveals that diagrams serve a principled, theoretical role in Euclid's mathematics. Only a restricted range of a diagram's spatial properties are permitted to justify inferences for Euclid, and these self-imposed restrictions can be explained as serving the purpose of mathematical control. Manders' central distinction is that between the *exact* and *co-exact* properties of geometric diagrams. Any one of Euclid's diagrams contains a collection of spatially related magnitudes—e.g. lengths, angles, areas. For any two magnitudes of the same type, one will be greater than another, or they will be equal. These relations comprise the *exact* properties of the diagram. How these magnitudes relate topologically to one another—i.e. the regions they define, the containment relations between these regions—comprise the diagram's *co-exact* properties. A diagram of a triangle, for instance, can vary with respect to its exact properties. That is, the lengths of the sides, the size of the angles, the area enclosed, can vary. Yet across all these variations its co-exact properties remain the same. A diagram invariably consists of three bounded linear regions, which together define an area.

The key observation is that Euclid's diagrams contribute to proofs *only* through their co-exact properties. Euclid never infers an exact property from a diagram unless it follows directly from a co-exact property. Exact relations between magnitudes which are not exhibited as a containment are either assumed from the outset or are proved via a chain of inferences in the text. For an illustration, consider proposition 35 of book I. It asserts that any two parallelograms which are bounded by the same parallel lines and share the same base have the same area. Euclid's proof proceeds as follows.

Let ABCD, EBCF be parallelograms on the same base BC and in the same parallels AF, BC.

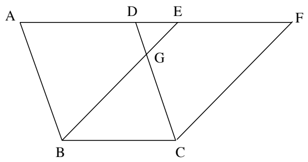

Figure 1. Diagram for proposition 35

Since ABCD is parallelogram, AD equals BC (proposition 34). Similarly, EF equals BC.
Thus, AD equals EF (common notion 1).
Equals added to equals are equal, so AE equals DF (Common notion 2).
Again, since ABCD is a parallelogram, AB equals DC (proposition 34) and angle EAB equals angle FDC (proposition 29).
By side angle side congruence, triangle EAB equals triangle FDC (proposition 4).
Subtracting triangle EDG from both, we have that the trapezium ABGD equals the trapezium EGCF (common notion 3).
Adding triangle GBC to both, we have that ABCD equals EBCF (common notion 2).

QED

The proof is independent of the diagram up until the inference that AE equals DF. This step depends on common notion 2, which states that if equals are added to equals, the wholes are equal. The rule is correctly invoked because four conditions are satisfied: $AD = EF$, $DE = DE$, DE is contained in AE, and DE is contained in DF. The first pair of conditions are exact, the second pair co-exact. Accordingly, the first pair of conditions are seen to be satisfied via the text, and the second pair via the diagram. Similar observations apply to the last two inferences. The applicability of the relevant common notion is secured by both the text and the diagram. With just the textual component of the proof to go on, we would have no reason to believe that the necessary containment relations hold. Indeed, we would be completely in the dark as to the nature of containment relations in general.

An explanation for why Euclid restricted himself to the co-exact properties of diagrams in proofs emerges when we reflect on the general conditions necessary for the transmission of checkable proofs. Any such proof, diagrammatic or otherwise, ought to be easily reproducible. Generating the symbols which comprise it ought to be straightforward and unproblematic. Yet there seems to be room for doubt whether one has succeeded in constructing a diagram according to its exact specifications perfectly. The compass may have slipped slightly, or the ruler may have taken a tiny nudge. In constraining himself to the co-exact properties of diagrams, Euclid is constraining himself to those properties that fellow geometers can be trusted to produce when they construct the diagram themselves.

If Manders's analysis is correct, Euclid's proofs ought to go through with diagrams which are equivalent in a co-exact sense (hereafter *c.e. equivalent*), but differ with respect to their exact properties. This turns out to be the case. The proof of proposition 35, for instance, still works if we substitute either of the diagrams in figure 2 for the given diagram. The diagram need not even satisfy the stipulated exact conditions. The diagram in figure 3 also fulfills the role

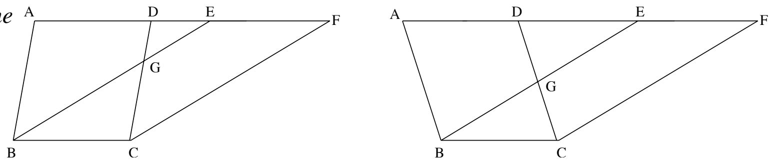

Figure 2. Other possible diagrams for proposition 35

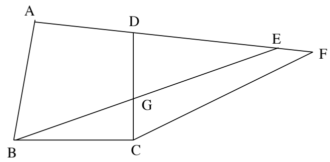

Figure 3. Inexact but acceptable diagram for proposition 35

the proof demands of it. The diagram's burden is to reveal how certain co-exact relationships lead to others. It is not used to show exact relationships. This is the job of the text. The proof must employ a particular diagram, with particular exact relationships. But since the proof only calls on the co-exact relationships of the diagram, its diagrammatic inferences are equally well justified by *any* diagram c.e. equivalent to the particular one used.

3 CHECKING GENERALITY

Simply recognizing the restricted role of the diagram for Euclid is not enough, however, to remove all concerns about the mathematical rigor of his diagrammatic method. A feature of the proof of proposition 35 is that it does not contain a construction stage. As a result, a diagram for the proof displays only those geometric relations stipulated at the beginning of the proof. This is not the case for proofs with a construction stage. A construction stage in a Euclidean proof dictates how new geometric elements are to be built on top of the given configuration. The demonstration stage then follows, in which inferences from the augmented figure can be made. In the presentation of the proof, the building up process is not shown explicitly. All that appears is the end result of the construction on a particular configuration.

The soundness of Euclid's co-exact inferences from such diagrams is much less obvious. The construction is always performed on a particular diagram. Though the diagram is representative of a range of configurations—i.e. all configurations c.e. equivalent to it—it cannot avoid having particular exact properties. And these exact properties can influence what co-exact relations come to be exhibited within the final diagram. When the same construction is performed on two diagrams which are c.e. equivalent but distinct with respect to their exact features, there is no reason to think that the two resulting diagrams will be c.e. equivalent. This leads to doubts that the co-exact relations Euclid expects us to read off of an augmented diagram hold for all possible constructions. And there is nothing in the diagram itself to remove these doubts. Euclid's diagrammatic method would thus seem to fail to provide for a way to check for the soundness of certain inferences.

Consider for example the construction of proposition 2 of book I of the *Elements*. The proposition states a construction problem: given a point A and a segment BC, construct from A a segment equal to BC. Euclid advances the following construction as a solution to the problem:

From the point A to the point B let the straight line AB be joined; and on it let the equilateral triangle DAB be constructed.

Let the straight lines AE, BF be produced in a straight line with DA and DB.

With center B and radius BC let the circle GCH be described; and again, with center D and radius DG let the circle GKL be described.

If the construction is performed on the *particular* point A and a *particular* segment BC in the first diagram of figure 4 the result is the second diagram of figure 4. If however the construction is performed on the different particular configuration in figure 5, the result is the diagram to the right of it. This augmented diagram is distinct

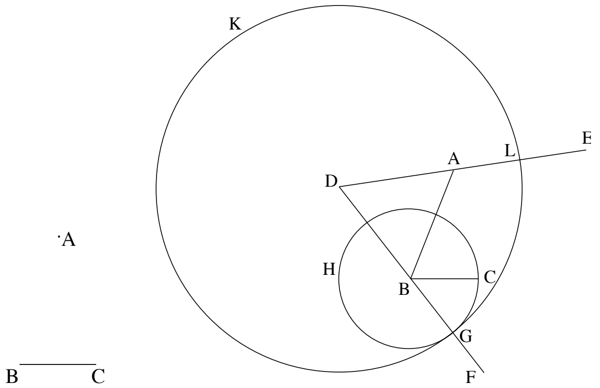

Figure 4. Possible initial and final diagram for proposition 2

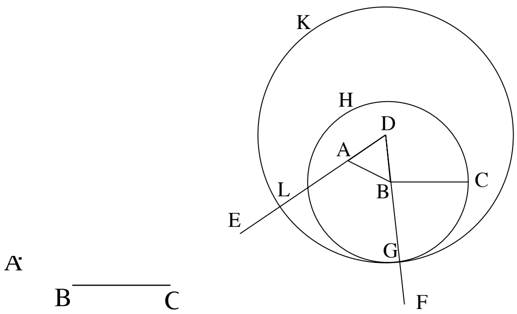

Figure 5. Another possible pair of initial and final diagrams for proposition 2

from the first diagram, topologically. Euclid nevertheless uses the co-exact features of one such diagram to argue that the construction does indeed solve the stated construction problem. The crucial step in the argument is the inference $AL = BG$. This follows from an application of the equals subtracted from equals rule. And for this to be applicable, A must lie on the segment DL and B must lie on the segment DG. So with these two diagrams we see that with two of the possible exact positions A can have to BC the topology needed for the proof obtains. But *prima facie* we have no mathematical reason to believe that it obtains for *all* the other positions A can have to BC.

What seems to be lacking here, and in other proofs of Euclid's that contain a construction stage, is a way to check that his diagrammatic inferences hold in general. Euclid, it seems, has failed to do his part as a proof expositor. It seems, in particular, that he needs to do more to justify the generality of his diagrammatic inferences. This appearance is line with the standard view that diagrammatic justifications are opaque. But in fact a natural technique exists whereby the co-exact relations in augmented diagrams can be verified. This technique has been codified in **Eu**, a formal, diagrammatic system of proof inspired by Manders' work (See [4] for an introduction to the system).

The idea behind the technique is that a diagram with auxiliary elements ought to be understood in terms of the construction that produced the auxiliary elements. Consider the diagram in 6. Many dis-

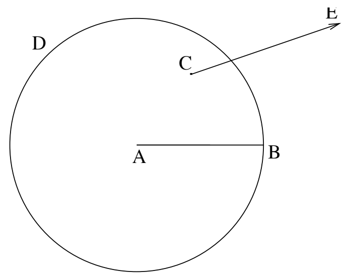

Figure 6.

tinct constructions could have produced it. For instance, the initial configuration could have been the segment AB, and the construction steps leading to the diagram could have been:

- draw the circle D with center A and radius AB.
- pick a point C in the circle D, and a point E outside it.
- produce the ray CE from the point C.

Call this construction **C1**. Alternatively, it is possible that the initial configuration consists of the segment AB and the points C and E, while the construction consists of the following two steps:

- draw the circle D with center A and radius AB.
- produce the ray CE from the point C

Call this construction **C2**.

Now, if **C1** is responsible for the diagram, we are justified in taking the position of C within D as a general property of the diagram. The act of picking C in D fixes the point's position with respect to the circle as general. And since we know the position of C relative to D is general, we can pick out the point of intersection of the ray CE with D with confidence. It always exists in general, since a ray originating inside a circle must intersect the circle. In contrast, none of these inferences are justified if **C2** is responsible for the diagram. Nothing is assumed from the outset about the distance of the point C to A. And so, even though C lies within D in this particular diagram, it could possibly lie on D or outside it. Further, as the position of C relative to D is indeterminate, the intersection point of CE and D cannot be assumed to exist in general, even though one exists in this particular diagram.

Viewing proposition 2 in this way, we can satisfy ourselves that Euclid's diagrammatic inferences are sound. Though the position of segment BC with respect to the triangle ADB is indeterminate, what that segment contributes to the proof is the circle H, whose role in turn is to produce an intersection point G with the ray DF. The intersection point always exists no matter the position of BC to the ray DF. We can rotate BC through the possible alternatives, and we will always have a circle H whose center is B. And this is all we need to be assured that the intersection point G exists. The ray DF contains B, since it is the extension of the segment DB, and a ray which contains a point inside a circle *always* intersects the circle.

A similar argument shows that the intersection point L of the ray DE and the circle K always exists. The argument does not establish, however, that A lies *between* D and L. Here a case analysis is forced upon us. We must consider the case where A coincides with L, or the case where L lies between A and D. These latter two possibilities, however, are quickly ruled out, since they imply that $DL = DA$ or that $DL < DA$. This contradicts $DA < DL$, which follows from the equalities $DA = DB$, $DG = DL$ and the inequality $DB < DG$. (The equalities follow from the properties of equilateral triangles and circles. The inequality $DB < DG$ is entailed by the fact that B lies between D and G, which holds because G was stipulated to lie on the extension of DB.) Thus, Euclid's construction in I,2 can always be trusted to produce a configuration of 7 where

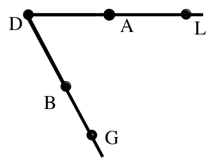

Figure 7.

$DA = DB$ and $DG = DL$. Accordingly, the equals-subtracted-from-equals rule is applicable, and we can infer that $AL = BG$.

Not only does viewing a diagram in terms of its construction in this manner confirm true theorems, it also refutes false ones. Consider the all-triangles-are-isoceles fallacy, an example sometimes used to illustrate the dangers of diagrammatic reasoning. The fallacy proceeds as follows:

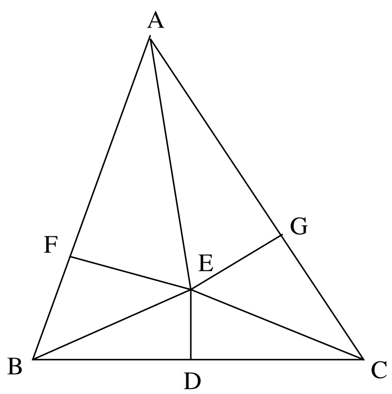

Figure 8. Diagram for fallacy

Let ABC be the given triangle. (Figure 8)
Produce the bisector to the angle BAC and the perpendicular bisector to the segment BC so that they meet in E.
Join E with B and C.
Finally, drop perpendiculars from E to the sides AB and AC.
By AAS congruence, triangles AEF and AEG are equal. So $EG = EF$.
By SAS congruence, triangle EDB and EDC are equal. So $EB = EC$.
Thus, EFB and EGC are equal right triangles, and $FB = GC$.
Since equals added to equals are equal, side $AB = AC$.

QED

On a quick reading the fallacy is puzzling. Yet when one inspects the diagram carefully in terms of the construction in the proof, the key, flawed diagrammatic inference of the argument comes to light. This inference is the position of point F on the segment AB and the position of the point G on the segment AC. On the basis of this, the equals added to equals rule is applied, and the conclusion obtained. Re-tracing the steps of the construction on the augmented diagram reveals that the points are not linked directly by the construction to the segments, and that to link them one would have to make an AD inference with the form of figure 9. That is, one would have to infer

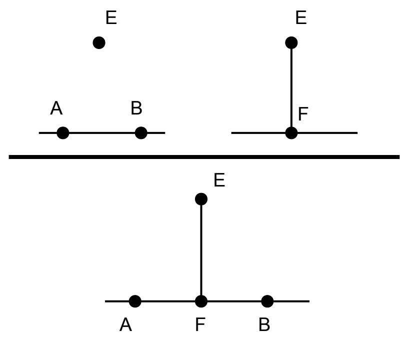

Figure 9. Key inference of fallacy

that given the co-exact position of the point E above the segment AB, and the position of endpoint F on the line (and *not* the segment) determined by AB, then F lies on the segment AB. The inference is clearly suspect. Being co-exact, the relation depicted by the first diagram only specifies that E lies above the line. It does not specify *anything* about its metric relation to the points A and B. It can lie *anywhere* above it. And so the foot of the line F can lie anywhere with respect to A and B on the line.

4 CONCLUSION

The technique illustrated here with proposition 2 and the fallacy is formalized in **Eu**. As a formal system of proof, it contains a finite set of rules for distinguishing the general from the particular in diagrams. An example is a rule which allows one to infer that a ray intersects a circle, given the position of the ray's endpoint inside the circle. The proofs that can be carried out within the system match Euclid's closely, often step for step. There are to be sure places where the **Eu** version of a proof seems needlessly involved next to Euclid's original. This is usually attributable, however, to the fact that **Eu** is a formal system, subject to the constraint that all rules of inference be laid out in advance. Because of this, one is sometimes obligated to prove something which is no less obvious or basic from a geometric point of view than the soundness of the rules licensing the steps. Accordingly, Euclid does not simply demand from his readers that they check that he has applied a pre-accepted list of axioms and rules correctly. They must also check for the soundness of what in **Eu** would be classified as diagrammatic inference rules.

Understood in terms of **Eu**, then, Euclid's diagrammatic proof method is a quasi-formal one that does indeed provide enough information for his results to be verified. Though a diagram considered in isolation is opaque as to what is and isn't representative in it, there is no need to consider the diagram in isolation. We can take the way it was constructed into account. This information can be used to analyze the diagram into a sequence of sub-diagrams ordered by the dependencies between the geometric objects represented. The geometric situation represented by these sub-diagrams (e.g. a ray whose endpoint is positioned within a circle) is simpler, and as a result the issue of geometric generality becomes manageable. Judgements as to the generality of co-exact relations manifest in the augmented diagram become possible. The apparent gaps induced by proofs with a construction stage can thus be filled.

Moreover, **Eu** shows that filling in the gaps does not require all the intellectual work involved in filling them in according to the logical ideal. On the latter approach, the co-exact information of the proof that is represented diagrammatically must be reformulated in sentential form. This amounts in the end to developing an abstract theory of order, as is done famously in [2]. Primitives must stipulated, axioms identified, and definitions given. Then it must be shown, in purely logical terms, that the theory of order so specified agrees with the spatial properties of the concepts of elementary geometry. Though it is a straightforward exercise, it takes time. Further, it involves many choices that are arbitrary from a geometrical point of view. In Hilbert's theory of order, for instance, there is just one primitive: the one dimensional relation of betweenness. The two dimensional relation of two points lying on the same sides of the line is defined in terms of it. But from a geometrical point of view, neither is more fundamental than the other. One could just as easily define the one dimensional relation in terms of the two dimensional one, as it is done in [7].

In verifying Euclid's proofs according to the technique laid out by **Eu**, one does not have to makes such arbitrary choices, and engage in such theoretical development. What is required is possession of the concepts of elementary geometry, and how they are represented in diagrams. To verify the generality of the intersection of a ray and circle in a diagram, for instance, one must understand that the physical diagram of a circle and a point within it represents a continuous range of geometric possiblities, and that the intersection of the ray constructed from the point and the circle is independent of the point's variation over the range. On the basis of the diagrammatic inference rules of **Eu** one can in fact specify a general schema for the diagrammatic inferences needed to verify Euclid's diagrammatic proofs:

$$\frac{R_1(\vec{f},\vec{x}) \quad C(\vec{x},y)}{R_2(\vec{f},y)}$$

In the schema, $R_2(\vec{f},y)$denotes the positional relation in the diagram inferred as general (e.g. the intersection of the point and the ray). $C(\vec{x},y)$ denotes a construction step (e.g. the construction of a ray), which produces y from the tuple $\vec{x}$, and $R_1(\vec{f},\vec{x})$ denotes a collection of positional relations already established as general between objects $\vec{f}$ and $\vec{x}$ (e.g. the position of the point in the circle). The tuple $\vec{f}$ is best understood as a frame of reference. $R_1(\vec{f},\vec{x})$ is the condition that the objects of $\vec{x}$ have a certain position with respect to $\vec{f}$. This position along with the nature of the construction $C(\vec{x},y)$ licenses the inference $R_2(\vec{f},y)$—i.e. y has a certain position with respect to the frame of reference $\vec{f}$. What's behind the inference, specifically, is the recognition of a positional invariant. $R_1(\vec{f},\vec{x})$ does not fix the *exact* position of $\vec{x}$ with respect to $\vec{f}$ but specifies a range of positions over which $\vec{x}$ can vary. This limitation in $\vec{x}$'s positional relationship to $\vec{f}$ forces a positional relationship of y to $\vec{f}$.

For more examples of such inferences, represented diagrammatically, see figure 10. The ray/circle intersection inference is shown in the upper left corner of the figure 10. In each of the inferences, the

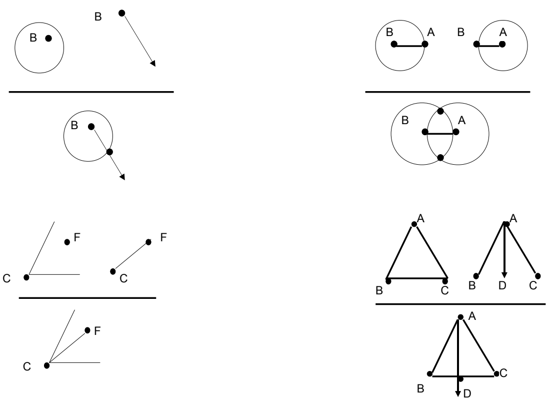

Figure 10. Some basic diagrammatic inferences

diagram on the top left depicts $R_1(\vec{f},\vec{x})$, the diagram on the top right $C(\vec{x},y)$ and the diagram below the line depicts the invariant relation $R_2(\vec{f},y)$. Every example is an inference that occurs in the *Elements* by the analysis of **Eu**. The inference of the top left inference occurs in the **Eu** analysis of I,2, that of the top right in the analysis of I,1, that of the bottom left in the analysis of I,6 and 1,16, and that of bottom right in the analysis of I,10. It is worth noting that the legitimacy and fundamental nature of inferences of this form is attested by the fact that they appear as axioms in modern axiomatizations of elementary geometry. Pasch's axiom, found in [2], and axiom $A13'$ in Tarski's theory $\mathcal{E}_2''$ [6] are examples.[3]

In isolating a basic form of geometric inference, **Eu** provides an account of what is fundamental to geometric reasoning. More broadly, the analysis may fruitfully serve as a model for investigations into reasoning in other areas of mathematics. The relation of Euclid's *Elements* to modern geometric theories such as Hilbert's suggests the relation of informal proofs of mathematical practice to

[3] In words, the Pasch axiom is 'any line constructed from a point on one side of the triangle intersects one of the triangles two other sides.' It is thus a proposition of the form

$$\forall \vec{f}, x, y \;\; (R_1(\vec{f}, x) \;\&\; C(x, y)) \longrightarrow R_2(\vec{f}, y)$$

where $\vec{f}$ is a triangle, R_1 expresses that the point x lies on a side of the triangle, $C(x,y)$ expresses that the line y is constructed from x, and R_2 expresses that the line y intersects one of the triangles other two sides. Tarski's $A13'$ is a proposition of the same form corresponding to the upper left inference in figure 10.

the logical ideal. In fact, Hilbert's treatment of Euclid's more intuitive method is often used to illustrate what the logical ideal demands. It thus seems worth looking into whether informal proofs of mathematical practice relate to artificial formal ones in the way Euclid's diagrammatic method relates to Hilbert's logical one. Can we in particular locate in other ares of mathematics symbols analogous to Euclid's diagrams, symbols subject to formal constraints but at the same used in tandem with a primitive understanding of mathematical concepts? If we can, we will be closer, I believe, to a satisfactory account of how proof checking actually works in mathematical practice.

REFERENCES

[1] J. Avigad and P. Raff K. Donelly, D. Gray, 'A formally verified proof of the prime number theorem', *ACM Transactions on Computation Logic*, **9(1:2)**, 1–23, (2007).

[2] David Hilbert, *Foundations of Geometry*, Open Court Publishing Company, La Salle, Illinois, 1971.

[3] K. Manders, 'The euclidean diagram', in *Philosophy of Mathematical Practice (Mancosu, 2008)*, 112–183, Clarendon Press, Oxford.

[4] J. Mumma, 'Proofs, pictures, and euclid'. to appear in *Synthese*.

[5] J.A. Robinson, 'Logic is not the whole story', in *First Order Logic Revisted (Hendricks, Neuhaus, Pedersen, Scheffer, Wansing, 2004)*, 287–302, Logos Verlag, Berlin.

[6] A. Tarski, 'What is elementary geometry?', in *The Axiomatic Method, with Special Reference to Geometry and Physics (Henkin, Suppes, Tarksi, 1959)*, 16–29, North Holland Publishing Company, Amsterdam.

[7] J. von Plato, 'A constructive theory of ordered affine geometry', *Indagationes Mathematicae*, **9**, 549–562, (1998).