Proofs, Pictures, and Euclid

John Mumma

Abstract. Though pictures are often used to present mathematical arguments, they are not typically thought to be an acceptable means for presenting mathematical arguments *rigorously*. With respect to the proofs in the *Elements* in particular, the received view is that Euclid's reliance on geometric diagrams undermines his efforts to develop a gap-free deductive theory. The central difficulty concerns the generality of the theory. How can inferences made from a particular diagrams license general mathematical results? After surveying the history behind the received view, this essay provides a contrary analysis by introducing a formal account of Euclid's proofs, termed **Eu**. **Eu** solves the puzzle of generality surrounding Euclid's arguments. It specifies what diagrams Euclid's diagrams are, in a precise formal sense, and defines generality-preserving proof rules in terms of them. After the central principles behind the formalization are laid out, its implications with respect to the question of what does and does not constitute a genuine picture proof are explored.

The prevailing conception of mathematical proof, or at least the conception which has been developed most thoroughly, is logical. A proof, accordingly, is a sequence of sentences. Each sentence is either an assumption of the proof, or is derived via sound inference rules from sentences preceding it. The sentence appearing at the end of the sequence is what has been proved.

This conception has been enormously fruitful and illuminating. Yet its great success in giving a precise account of mathematical reasoning does not imply that all mathematical proofs are, in essence, a sequence of sentences. My aim in this paper is to consider data which do not sit comfortably with the standard logical conception: proofs in which pictures seem to be instrumental in establishing a result.

I focus, in particular, on a famous collection of picture proofs—Euclid's diagrammatic arguments in the early books of the *Elements*. The familiar sentential model of proof portrays inferences as transitions between sentences. And so, by the familiar model, Euclid's diagrams would at best serve as a heuristic, illustrative device. They could not be part of the rigorous proof itself. In direct opposition to this, I introduce the proof system **Eu**, which accounts for the role of the diagram *within* Euclid's mathematical arguments. It possesses a diagrammatic symbol type, and specifies rules of proof for these symbols. It thus provides a formal model where Euclid's diagrams are part of the rigorous proof. Though **Eu** has been designed specifically to formalize these arguments, we can subsequently look to it to understand what is distinctive about proving with pictures. **Eu** represents a species of rigorous mathematical

thought falling outside the scope of the familiar model. Not all rigorous reasoning in mathematics proceeds line by line.

After reviewing the history behind the modern, critical stance towards Euclid's diagrammatic arguments, I explain the proof system **Eu**. In a final section, I explore what is novel about **Eu**'s picture proofs, and what questions these novel features raise for the philosopher of mathematics.

1. Historical background

For most of its long history, Euclid's *Elements* was the paradigm for careful and exact mathematical reasoning. In the past century, however, it has been just the opposite. Its proofs are often invoked to illustrate what rigor in mathematics does *not* consist in. Though some steps of Euclid's proofs are respectable as logical inferences, a good many are not. With these, one cannot look only at the logical form of the claims in the proof and understand what underlies the inference. One is forced, rather, to look at the accompanying diagram. The modern opinion is that Euclid's proofs exhibit a deductive gap at such places.

The full historical story behind this opinion is of course a complicated one. Three interrelated factors, however, are consistently tied to its emergence, and it is these I will discuss. They are: the generality problem, the modern mathematical understanding of continuity, and the modern axiomatic method. The first is a puzzle that has surrounded Euclid's proofs from the time they were conceived. The second is a 19th century conceptual development which seemed to expose diagrammatic methods as hopelessly imprecise. And the third is a methodological development, also occurring in the 19th century, which provided a clear and exact way to understand both geometric generality and continuity.

The generality problem arises with Euclid's proofs because the diagram used for a proof is always a *particular* diagram. Euclid clearly did not intend his propositions to concern just the figure on display beside the proposition. They are applied in subsequent proofs to other figures, which are not exact duplicates of the original. And so, for Euclid, consultation of the original diagram, with all its particular features, is somehow supposed to license a generalization. But Euclid leaves the process by which this is done obscure. And so we are left with some doubt as to whether the jump from the particular to general is justified.

Even before the 19th century, when the legitimacy of Euclid's methods was taken for granted, philosophers recognized that there was something to be explained with this jump. The neo-Platonist Proclus asserts that the use of a particular diagram is justified because the

geometer does not "make use of the particular qualities" of the diagram (*A Commentary on the First Book of Euclid's Elements*, (207)). Roughly 13 centuries later, Berkeley reiterates the point. In assailing Locke's theory of abstract ideas, Berkeley argues that we need not invoke such things to account for the generality of Euclid's arguments. He asserted that though Euclid uses a particular triangle, with many particular properties, to establish a general proposition about triangles

> there is not the least mention made of *them* (the particular details) in the proof of the proposition.(Section 16 of the introduction to *Principles of Human Knowledge.*)

The claim of Berkeley and Proclus is that since no explicit connection is made in the proof between the particular qualities or details (such as the obliqueness/acuteness of the triangle's angles) and the conclusion, the conclusion does not depend on them in any way.

Kant saw a deep epistemological fact about mathematics in Euclid's jump from the particular to the general. In proving a proposition with a geometrical concept, the mathematician

> hastens at once to intuition, in which it considers the concept *in concreto*, though non-empirically, but only in an intuition which it presents *a priori*, that is, which it has constructed, and in which whatever follows from the universal conditions of the construction must be universally valid of the object of the concept thus constructed.(*Critique of Pure Reason*, A716/B744.)

A forefather of what seems to be the modern consensus regarding Euclid's proofs is Leibniz. Also commenting critically on Locke, he writes:

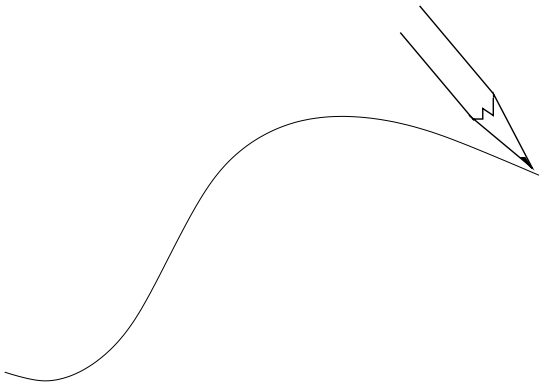
> ...it is not the figures which furnish the proof with geometers, though the style of the exposition may make you think so. The force of the demonstration is independent of the figure drawn, which is drawn only to facilitate the knowledge of our meaning, and to fix the attention; it is the universal propositions, i.e. the definitions, axioms, and theorems already demonstrated, which make the reasoning, and which would sustain it though the figure were not there. (Leibniz, 1949), p. 403

According to Leibniz, the arguments in the *Elements* succeed as proofs insofar as they express universal propositions. The universal propositions are what carry the reasoner through to the conclusion. The diagrams, though perhaps useful for pedagogical purposes, are inessential.

Mathematical developments in the 19th century seem to have confirmed Leibniz's position decisively. In the light of the huge advances made in geometry and analysis, the use of diagrams in geometric argument comes to look at best imprecise or at worst downright misleading.

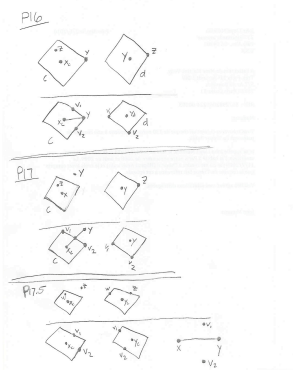
Of these advances, none seems more damaging to the legitimacy of diagrammatic proof than the sharper understanding of continuity.

Our intuitive grasp of the continuous phenomena treated by the calculus is robust. The image of a point tracing out a curve in space comes to most of us easily, and is naturally associated with the basics of differentiation and integration. What the work of 19th century figures like Bolzano, Weierstrass and Dedekind has been taken to show is that this imagery need not, and ought not, play a justificatory role in mathematical arguments. They defined continuity without any reference to geometric intuitions, in terms far more precise than any previous characterization. The connection with pictures arises from the fact that they are often used to convey the intuitive image of continuity. The tip of a pencil moving on a piece of paper becomes a surrogate for the point moving in the plane.



The danger in taking the pictorial representation too seriously is that it can lead to unsound inferences. Understanding curves in terms of such representations, one may conclude that a curve can fail to be differentiable only at a finite number of points (because one can only ever draw a finite number of jagged corners on a curve). That conclusion however is falsified by Weierstrass's construction of a nowhere differentiable curve.

Similarly, we might take the intermediate value theorem to have a quick and easy picture proof.[1] The theorem states that a continuous function f defined on an interval $[a, b]$ assumes all values between $f(a)$ and $f(b)$. Pictorially, it seems clear that if we draw a curve which starts below (or above) a horizontal line and ends above (or below) the line, the curve *must* intersect the line at some point.



But familiarity with the modern theory of continuity leaves room for doubt. Nothing in the picture fixes the underlying field of points. The field could be the reals, but it could also be the rationals or anything in between. All are consistent with what we see in the picture. But only in the case of the reals will the intersection point be guaranteed to exist. Thus the picture on its own fails to have the undeniable force of a mathematical proof.

The general observation here, if accepted, is fatal to the legitimacy of Euclid's diagrammatic proofs. Again and again, Euclid takes intersection points to exist because they appear in his diagrams. Indeed, he does this right away in proposition I,1 when he introduces the intersection point of two constructed circles. The only thing Euclid offers as justification is a diagram where two circles cross. Nowhere do we find the articulation of a continuity principle which by modern standards seems necessary to secure the point's existence.

It is instructive to consider what the role of such an assumption would be in the context of the *Elements*. In modern theories, order relationships on lines are expressed with a logical relation—e.g. a two place inequality relation or a three place betweenness relation. Such a relation provides the means by which the continuity assumption is formally expressed. But Euclid, famously, fails to develop a logically respectable theory of order. The way points relate on a line is exhibited directly, in diagrams. And so, though Euclid could perhaps formulate a continuity assumption like the least upper bound principle verbally, it would have mathematical relevance insofar as it could be applied in proofs containing diagrams. The situation, by modern standards of mathematical rigor, seems highly unsatisfactory. In logical theories, a continuity assumption is firmly rooted in a clearly defined method of proof. In Euclid's diagrammatic theory, its inferential role seems fuzzy and open-ended.

The inadequacy of Euclid's diagrammatic method in this respect is not limited to the issue of continuity. Whenever Euclid relies on a diagram in a proof, he seems to leave the firm and certain realm of rule-governed proof and enter a foggy, intuitive realm. This comes out dramatically when we compare Euclid's proofs with their counterparts in modern axiomatic theories of elementary geometry. Whereas some inferences in the *Elements* seem to be justified by an unanalyzable experience induced by the diagram, *all* inferences in a modern theory can be seen to be grounded in the application of sound rules to statements assumed as axioms or previously proved.

Fittingly, in the preface to the *Grundgesetze*, the father of modern logic articulates the methodological ideal satisfied by modern axiomatic theories and unsuccessfully pursued by Euclid:

> The ideal of a strictly scientific method in mathematics, which I have attempted to realize, and which might indeed be named after Euclid, I should like to describe as follows. It cannot be required that everything be proved, because that is impossible; but we can require that all propositions used without proof be expressly declared as such, so that we can see distinctly what the whole structure rests upon...

In this general foundational sense Frege puts his work in the same category as Euclid's. But he then distinguishes it from the *Elements* with the remark

> Furthermore, I demand—and in this I go beyond Euclid—that all modes of inference be specified in advance.((Frege, 1964), p. 2)

It is not difficult to understand why Frege took himself to have advanced beyond Euclid. The common notions, postulates, and definitions in book I are Euclid's foundational starting points. If these lay the theoretical groundwork in accordance with Frege's ideal, we ought to trace every move Euclid makes to them. But we cannot. Independent, obscure principles seem to be in play when we consult the diagram to understand the proof. It is as if with each new proposition we are being asked to accept a new proof technique. We do not accept the theory's methods of inference all at once, and then go on to deduce consequences from them. Rather, we approach the proofs on a case by case basis. We decide, *as* we are reading a diagrammatic proof, whether an inference never seen before is acceptable or not.[2]

Frege never produced a formal work on the foundations of geometry, so it is not clear how he thought this flaw in the *Elements* ought to be remedied. Many of his contemporaries, however, tackled the problem with compelling and influential results. Starting with Moritz Pasch's *Lectures in Modern Geometry* in 1882 ((Pasch, 1882)), a body of work emerged in the late 19th century which grounded elementary geometry in abstract axiomatic theories. Euclid's theorems were placed, finally,

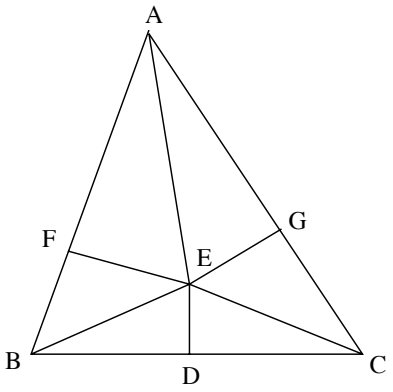


Figure 1. Diagram for fallacy.

in a mathematical context where all modes of inference were laid out explicitly in advance. This development is now universally regarded as a methodological breakthrough. Geometric relations which previously were logically free-floating, because they were understood via diagrams, were given a firm footing with precisely defined primitives and axioms.[3]

Pasch, Hilbert, and Klein are all explicit on the superiority of abstract axioms over diagrams. Pasch declared "that the theorem is only truly demonstrated if the proof is completely independent of the figure." In lectures which would eventually develop into his famous axiomatization, Hilbert reaffirmed the position.

> A theorem is only proved when the proof is completely independent of the diagram. The proof must call step by step on the preceding axioms. The making of figures is [equivalent to] the experimentation of the physicist, and experimental geometry is already over with [laying down of the] axioms. ((Hilbert, 2004), p. 75, as translated by Mancosu in (Mancosu, 2005).)

Klein's argument (1939), showing that it is necessary and not merely pedantic to lay out all axioms in advance, is especially illuminating. At the center of his discussion is the 'all triangles are isosceles' fallacy. It is a diagrammatic argument in the style of Euclid whose putative conclusion is that all triangles are isosceles. As is it happens with many Euclidean proofs, it calls for a construction on an arbitrary triangle. The result of the construction is presented diagrammatically in figure 1. The construction proceeds as follows: 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 the perpendiculars from E to the sides AB and AC. Applying the familiar triangle congruence theorems to the resulting figure, we can deduce that $AF = AG$ and $FB = GC$. Since equals added to equals are equal, side $AB = AC$.

Concluding from this that all triangles are isosceles is obviously unwarranted, but it is not clear what in Euclid's diagrammatic method of proof blocks it. If some diagrams of triangles license general conclusions, why doesn't this one? It is in fact impossible for a physical figure to realize the spatial relationships on display in the diagram *and* to satisfy the metric conditions stipulated in the construction. This is what makes the fallacy possible. When the construction is carried out accurately on a non-isosceles triangle, E lies outside the triangle, and crucially *only one* of the two points F and G lies outside the triangle. The step invoking the equals added to equal rule is thus the place where the proof goes wrong. But the diagram does not do a good job of revealing this to us. Our inability to discern slight divergences from exact metric conditions can lead to fallacies, it seems. Accordingly, Klein stresses the necessity of elucidating *all* the axioms of elementary geometry (even the most obvious ones). Building up the subject along the lines of Pasch and Hilbert, we are safely insulated from the errors diagrammatic reasoning is prone to. Specifically, there is no danger of an unwarranted generalization, as everything is proven with the laws of logic from geometric axioms whose generality is beyond question. (A prime example of such a geometric axiom is the famous axiom from Pasch which states that any line entering a triangle leaves it.) For the same reason, there is no danger of being misled by imprecise visualizations of continuous geometric phenomena. The axioms specify exactly what does and doesn't exist. Given these advantages it is not surprising that Pasch, Hilbert, and Klein championed axiomatic proofs over diagrammatic ones in geometry.

This view has since become standard. Euclid's reliance on pictures in his arguments disqualifies them as rigorous proofs. They do not stand on their own, mathematically. They must be supplemented with the appropriate axioms. Less has been said in support of the stronger view that pictorial presentations of any kind of mathematical argument (geometrical or otherwise) do not count as legitimate proofs. It seems safe to assume however that the default view on the issue simply generalizes the standard assessment of Euclid. Since the pictures in the *Elements* are not taken to prove anything, pictures are not taken to prove anything in other areas of mathematics.

In the past 15 years, a sizable literature consciously opposed to this attitude has emerged. The work ranges from technical presentations of formal diagrammatic systems of proof (e.g. (Shin, 1994)) to philosophical arguments for the mathematical legitimacy of pictures. (e.g. (Brown, 1997), (Dove, 2002)) Despite the fact the Euclid's diagrammatic arguments are often discussed, a satisfactory account of them has yet to be given. The formal system **Eu** was developed to rectify

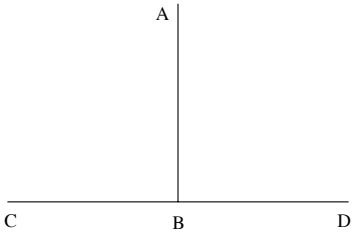


Figure 2. Diagram for proposition 10.

this—i.e. to bring out more clearly the structure of Euclid's diagrammatic reasoning. At the same time, it helps us get a better sense of the characteristic features of picture proofs in general.

2. The Proof System Eu

Discussing Klein's reflections on the all-triangle-are-isosceles fallacy, Ian Mueller is reluctant to endorse Klein's conclusions. Euclid's diagrammatic proofs do not, for Mueller, constitute a serious breach of mathematical rigor. "Perhaps a 'pupil of Euclid' might stumble on such a proof; but probably he, and certainly an interested mathematician, would have no trouble figuring out the fallacy on the basis of intuition and figures alone." (Mueller, 1981), p. 5. Mueller, however, has little to say about the way ancient geometers arrived at sound, general results via intuitions and figures. He simply attributes it to "general mathematical intelligence."

Mueller's desire to soften Klein's assessment is understandable. Though explicit rules for every proof step are absent, a close reading of book I leaves the impression that strict, if implicit, standards are in force. As Oswald Veblen observes, Euclid's purpose was

> to prove every proposition which he could prove, and to prove it with a minimum of assumptions. This required him often to prove statements which are intuitively evident.(Veblen, 1914)

Propositions 14 and 20 furnish good examples of what Veblen is referring to. Proposition 14 states the following: if the sum of two angles which share a side (e.g. angle CBA and the angle ABD below) is equal to two right angles, then the non-adjacent sides of the angles (e.g. BC and BD) lie in a straight line. (See figure 2)

Proposition 20 asserts the triangle inequality—i.e. the sum of two sides of a triangle is always greater than the third. It is hard to see why Euclid troubled himself to prove these statements if he allowed himself to draw conclusions from diagrams in an unconstrained, intuitive way. He seems to have had some conception of what does and does not

require proof. The point does not have to do with the importance of the proposition 20, for instance, to Euclid's theory of geometry. We can consult Proclus's commentary to understand the proposition's role with respect to what comes after it in book I. The question, rather, is why the triangle inequality appears 20 propositions into book I, and not as axiom (as it does in modern characterizations of metric spaces) or as an early proposition which is proved on intuitive grounds (which seems to be Euclid's approach to side-angle-side congruence of triangles with proposition 4). One answer is simply that there is a great deal of arbitrariness in the first twenty propositions—i.e. Euclid relied on intuition when it suited his fancy. An alternative thesis is that definite standards restricted what Euclid could and couldn't do with diagrams in proofs.

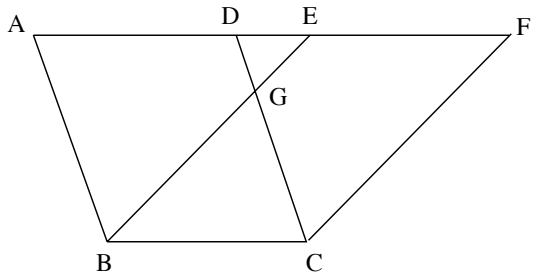
Defending this thesis satisfactorily requires making explicit the standards posited to shape the *Elements* implicitly. Though there have been some recent attempts, none have provided a level of detail necessary to move decisively beyond Mueller's appeal to "general mathematical intelligence."[4] A more compelling account would match the explicitness present in modern axiomatic theories of geometry. That is, it would formulate in advance Euclid's implicit standards in terms of precisely defined, sound rules. The proof system **Eu** is advanced as such an account.

Eu is, in the terminology of Barwise and Hammer in (Barwise and Hammer, 1996), a *heterogeneous* system. It has a conventional sentential syntax *and* a discrete diagram symbol type. The inspiration for the system is Ken Manders' work on ancient geometric proof in (Manders, 2008). His investigations have revealed that both text and diagram have definite roles in establishing a result. The rules of **Eu** have been designed so that its sentences and diagrams fulfill these roles.[5]

To explain the division of labor between text and diagram, Manders distinguishes the *exact* and *co-exact* properties of 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. Diagrams of a single triangle, for instance, vary with respect to their exact properties. That is, the lengths of the sides, the size of the angles, the area enclosed, vary. Yet with respect to their co-exact properties the diagrams are all the same. Each consists of three bounded linear regions, which together define an area.[6]

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. It is not difficult to hypothesize why Euclid would have restricted himself in such a way. Any proof, diagrammatic or otherwise, ought to be 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 stable under such perturbations.

For an illustration of the interplay between text and diagram, 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.
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

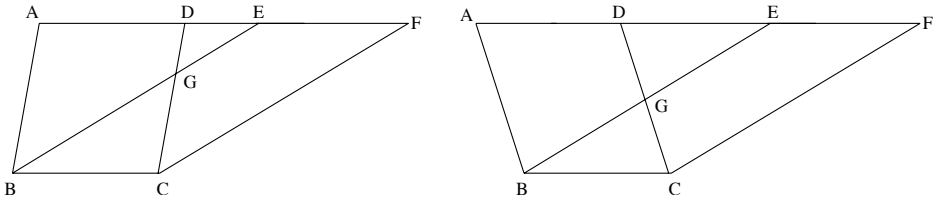


Figure 3. Alternate diagrams for proposition 35

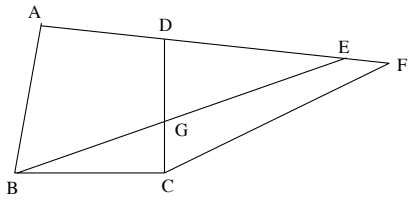


Figure 4. Metrically distorted diagram for proposition 35.

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.

The standard line is that this situation needs to be rectified with something like a betweenness relation. Manders's opposing thesis is that diagrams function in the *Elements* as reliable symbols because Euclid only invokes their co-exact features. Though we may not be able to trust ourselves to produce and read off the exact properties of diagrams, we can trust ourselves to produce and read off co-exact properties. Thus, Euclid seems to be within his rights to use diagrams to record co-exact information. 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 3 for the given diagram. The diagram need not even satisfy the stipulated exact conditions. The diagram in figure 4 also fulfills the

role 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 invariably employ a particular diagram, with particular exact relationships. But since the proof only calls on the co-exact relationships of the diagram, it holds of *all* diagrams which are c.e. equivalent to it. And so, in giving the distinction between exact and co-exact properties its due, we come to see what the generality of Euclid's results consists in.

The formal structure of **Eu** is built around these insights. A well-formed atomic claim of **Eu** has the form

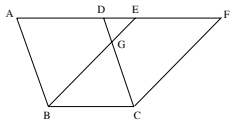
$$\Delta, A$$

where Δ is its diagrammatic component, and A is its sentential component. A is termed the *metric assertion* of the atomic claim. Their syntax is similar to that of first-order predicate logic. The syntax of the diagrams are defined so that the full range of co-exact relations conveyed by Euclid's diagrams are expressible, and the rules governing these symbols in proofs recognize only these relations. For the formal details, see the appendix.

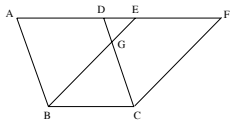
The propositions of the *Elements* are formalized as conditionals

$$\Delta_1, A_1 \longrightarrow \Delta_2, A_2$$

The antecedent Δ_1, A_1 fixes the properties of the figure as stipulated at the beginning of the proof. If one can, via the proof rules of **Eu**, produce the claim Δ_2, A_2 then the conditional is proved. Proposition 35, for instance, is represented as the conditional



$M \quad \longrightarrow$



$ABCD \equiv_{area} EBCF$

where the metric assertion M asserts equality between four pairs of angles: ABC and ADC; BAD and BCD; EBC and EFC; and BEF and BCF. In this case $\Delta_1 = \Delta_2$ and so the conditional represents one of Euclid's *theorems*. When Δ_2 contains additional elements, the conditional represents one of Euclid's *problems*.

In this way **Eu** turns the standard, logical assessment of the *Elements* on its head. In presenting a proposition, Euclid first provides a sentence which in general terms describes what is to be proved.

Following this is a proof which makes reference to a particular diagram. Looked at from a logical point of view, the general statement emerges as an essential part of the proof, and the particular diagram appears decorative. By the lights of **Eu**, in contrast, it is just the reverse. The particular diagram is essential to the proof, and the general statement plays no role. Propositions are represented in terms of diagrams and statements specifying metric relationships between magnitudes of the diagrams. The applicability of a proposition $\Delta_1, A_1 \longrightarrow \Delta_2, A_2$ in a future proof depends on whether the conditions expressed by Δ_1, A_1 obtain in that proof—i.e. whether a configuration c.e. equivalent appears in the proof's diagram, and whether the metric relationships given by A_1 have been shown to hold. There is no need for a general statement to act as an intermediary. Once it is understood how the diagrams function as proof symbols, any particular diagram can comprise a general mathematical claim.

3. The construction stage and generality

The insight that Euclid's proofs rely only on the co-exact properties of diagrams does much to determine what a formalization of the proofs ought to look like. It is not enough, however, to determine a unique formalization which solves the problem of generality. Rules of proof do not fall out immediately once a suitable diagrammatic and sentential symbolism has been specified. A deep difficulty remains. It arises from the fact that the diagram of a Euclidean proof rarely displays just the geometric elements stipulated at the beginning of the proof. They often have a construction stage dictating 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. The building up process is not shown explicitly. All that appears is the end result of the construction on a particular configuration.

As the proof of proposition 35 has no construction stage, it fails to illustrate this common feature of Euclid's proofs. The diagram of the proof contains just those elements which instantiate the proposition's general co-exact conditions. We are thus justified in grounding the result on the co-exact features of the diagram, given that we only apply the result to configurations which are c.e. equivalent to the diagram.

The soundness of Euclid's co-exact inferences is much less obvious when the proof's diagram contains augmented elements. 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 how the co-exact relations within the final diagram work out. 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. Euclid nevertheless draws conclusions from the co-exact features of one such diagram. The vexing question is: how do we know that the co-exact features that Euclid isolates are shared by *all* diagrams which could result from the construction? As there is nothing in the *Elements* addressing the question, it seems that all we have to assure ourselves that inferences from constructed figures are generally sound is not something mathematical, but something empirical: the fact that in the long history of the *Elements* as a canonical text in geometry no counter-examples to one of his proofs was successfully advanced.

Proposition 2 of book I illustrates the problem clearly. The proposition states a construction problem: given a point A and a segment BC, construct from A a segment equal to BC. The proof that a solution always exists is the following:

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.

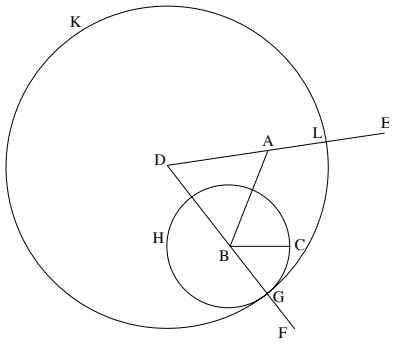


Figure 5.

Since the point B is the center of the circle GCH, BC is equal to BG.
Again, since the point D is the center of the circle GKL, DL equals DG.
And in these DA is equal to DB, therefore the remainder AL is equal to the remainder BG (common notion 3).
But BC was also proved equal to BG, therefore each of the straight lines AL, BC is equal to BG.
And things which equal the same thing also equal one another (common notion 1), therefore AL is also equal to BC.

QEF

The first lines of the proof constitute the proof's construction stage. The diagram shown with them shows how the construction turns out with a *particular* point A and a *particular* segment BC (shown in figure 6). The resulting figure does indeed seem to support the conclusion that AL equals BC. This means at the very least that the problem has been solved for this A and this BC. Yet the force of the proposition, mathematically, is that this construction can be effected on *any* segment and point. The proposition plays an indispensable role in the proof of I,3, and I,3 is applied throughout the *Elements*. Nothing in the proof of I,3, nor in its many applications afterwards, demands that the given segment and point have the particular exact position A and BC have to each other in the given diagram. By Euclid's standards, at least, carrying out the demonstration with this particular diagram is enough to secure the general result.

But it is not clear is whether we ought to adopt these standards. There is nothing explicit in his proof, taken by itself, which addresses the worry that the construction will support the same inferences if it is performed on a configuration the exact position of A to BC in figure 7. The result of applying the construction to this figure is given in figure 8 This diagram is distinct from the diagram of figure 5, topologically. In the first diagram, the following relations hold: the point D lies outside the circle H; C and L are both to the right of BA; and E and D are both above BC. The relations between corresponding elements in the second diagram are different. Point D lies inside H, C and L lie on different sides of the segment BA, and E and D lie on different sides of BC.

The one diagram-based inference of the proof occurs with the claim that $AL = BG$. Crucially, the containment relations which justify it *are* shared by both diagrams. The equality $AL = BG$ 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

Figure 6. Initial diagram for particular diagram of figure 5.

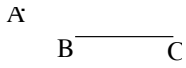


Figure 7. Different initial diagram for proposition 2.

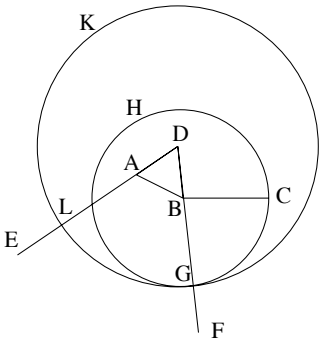


Figure 8. Different final diagram for proposition 2.

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.

Euclid, clearly, possessed the mathematical intelligence to pick out what does and does not hold generally in his diagrams. The main question with respect to formalizing the proofs is whether or not we can characterize this intelligence in terms of a precise and uniform method. The proof system **Eu**, I maintain, answers this question in the affirmative.

The method provided by **Eu** is based on the principle that what is general in a diagram depends on how it was constructed.[7] Consider the diagram of figure 9. Many distinct 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

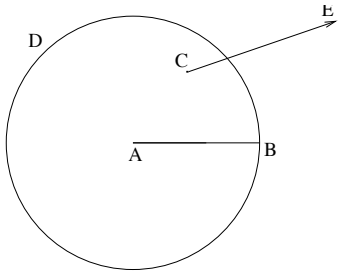


Figure 9.

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 $DL \equiv_{seg} DA$ or that $DL <_{seg} DA$. This contradicts $DA <_{seg} DL$, which follows from the equalities $DA \equiv_{seg} DB$, $DG \equiv_{seg} DL$ and the inequality

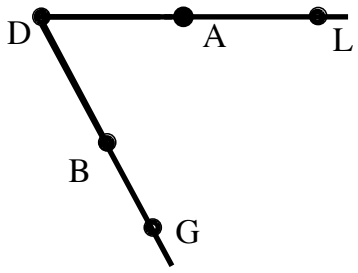


Figure 10.

$DB <_{seg} DG$. (The inequality $DB <_{seg} DG$ 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 as in figure 10. where $DA =_{seg} DB$ and $DG =_{seg} DL$. Accordingly, the equals-subtracted-from-equals rule is applicable, and we can infer that $AL \equiv_{seg} BG$.

Thus runs the proof of proposition 2 in **Eu**. Though the informal version given here is much more compact, each of its moves is matched in the formal version. Generally, proofs of propositions $\Delta_1, A_1 \longrightarrow \Delta_2, A_2$ in **Eu** are two tiered, just as they are in the *Elements*. They open with a construction stage, and end with a demonstration stage. The rules which govern the construction stage are relatively lax. One is free to enrich the initial diagram Δ_1 by adding points, joining segments, extending segments and rays, and constructing a circle on a segment. Presented as a sequence of **Eu** diagrams, the construction stage for proposition 2 is given in figure 11. The last step in the construction yields a diagram Σ, which contains all the objects to be reasoned about in the demonstration. But it is not Σ alone, but the whole construction history of Σ, which determines what can be inferred in the demonstration.

The sequence of steps by which Σ was constructed determine a partial ordering $\triangleright$ of its geometric elements. An element x in the diagram *immediately precedes* y if the construction of y utilized x. For instance, if the points A and B are joined in a construction, the points A and B immediately precede the segment AB. Likewise, if a circle H is constructed with radius BC, the segment BC immediately precedes H. The complete partial ordering $\triangleright$ is simply the transitive closure of the *immediately precedes* relation. It serves to record the dependencies among the elements of Σ. For example, for the representation of I,2 in **Eu**, what the partial ordering works out to be is given in figure 12 As the elements A, B, C and BC are part of the initial diagram Δ_1, they were not constructed from any elements in Σ, and so nothing precedes

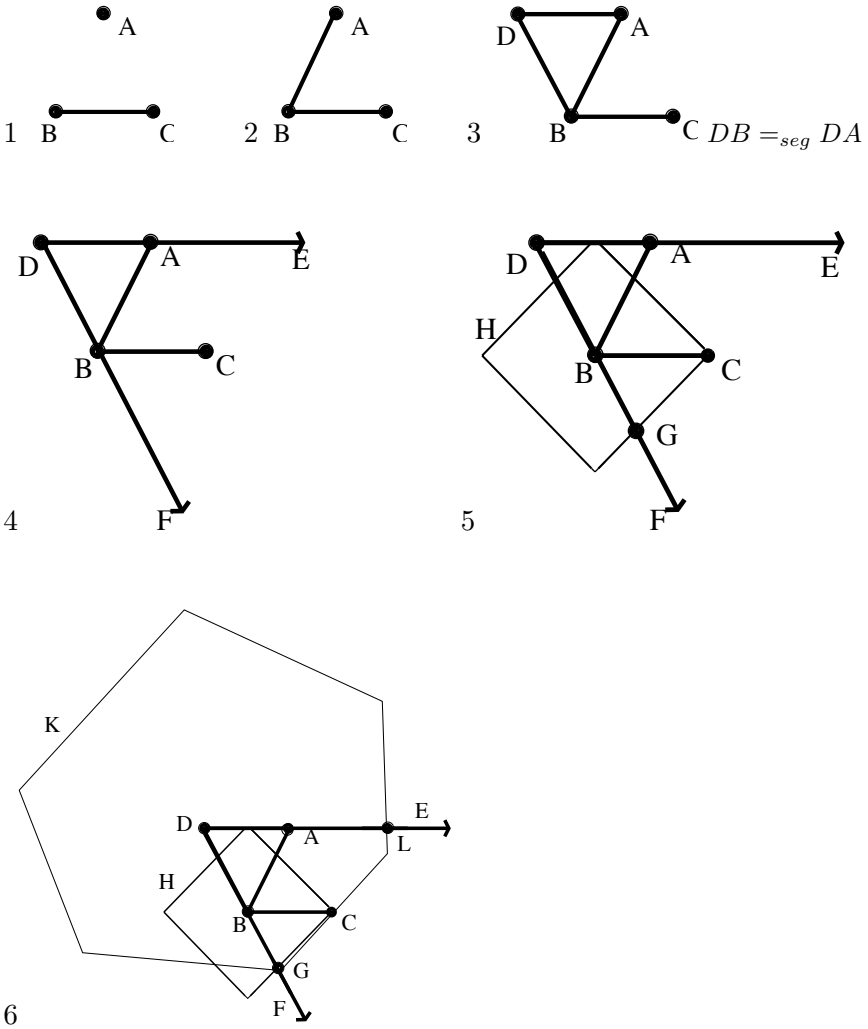


Figure 11. Construction of proposition 2 in **Eu**

them. The rest appear somewhere above these, according to the way they were introduced.

The construction thus produces a tuple

$$\langle \Sigma, M, \triangleright \rangle$$

which is called the *context* of the proof. The term M is the metric assertion which records the exact relationships stipulated from the beginning or introduced during the course of the construction. These three pieces of data serve as input for the demonstration stage. Rules of this stage are of two types: positional and metric.

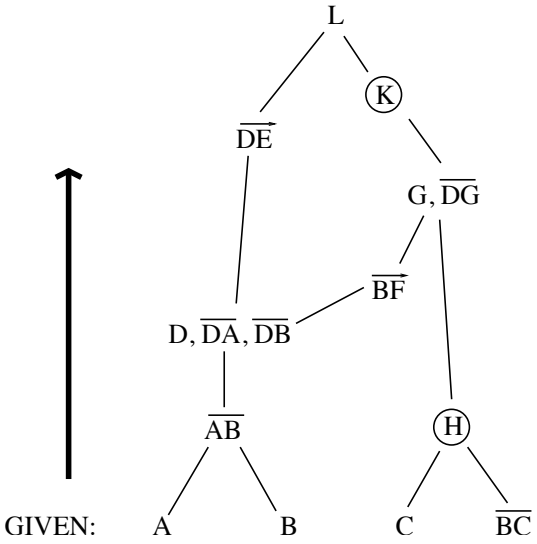


Figure 12. The poset for the construction of proposition 2.

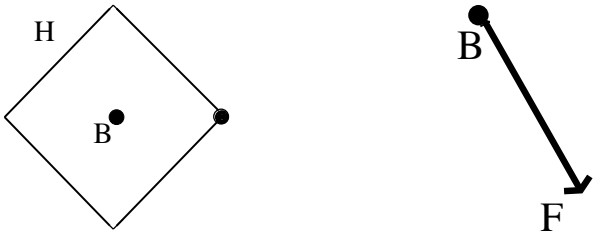


Figure 13.

An application of a positional rule results in a sub-diagram of Σ. Deriving a sub-diagram amounts to confirming the generality of the co-exact relationships exhibited in it. As such, the application of the rules are constrained by $\triangleright$. One can introduce as a premise any sub-diagram of the initial diagram Δ_1—i.e. any sub-diagram consisting of elements which have nothing preceding them by $\triangleright$. Any other sub-diagram must be derived from these by the positional rules, where the derivations proceed along the branches laid out by $\triangleright$. For instance, in the proof of proposition 2, one derives from the segment BC the first sub-diagram in figure 13 and from the points A and B the second sub-diagram in figure 13 From these two sub-diagrams we can then derive the sub-diagram of figure 14 from a rule which encodes the general condition for the intersection of a ray and a circle.[8]

The metric rules are more straightforward. They dictate how metric assertions can be inferred from established metric assertions and derived sub-diagrams. Most codify principles explicitly mentioned by

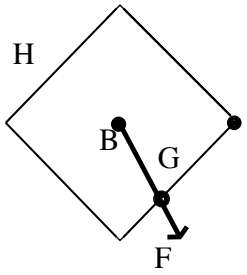


Figure 14.

Euclid—e.g. the transitivity of equality, the equals-added to equals rule, the equidistance of points from a circle's center.

Placing Euclid's proofs next to those of **Eu**, we arrive at a new conception of what the former lack. Euclid's proofs still have gaps, but they appear much smaller than they those which open up when a modern axiomatization serves as the ideal for rigorous proof in elementary geometry. Bringing Euclid up to Hilbert's standard means banishing diagrams from the proofs and replacing them with an abstract theory of order. The evidence for any such theory in the *Elements* is nil. Thus to understand Euclid as an imperfect version of Hilbert does not just reveal flaws in the proofs. A considerable theoretical chunk is posited to be missing. In contrast, a critique which takes **Eu** as the ideal is much less damning. What **Eu** has, and Euclid does not, is an explicit method for judging what is and isn't general in a constructed diagram. The extra steps that these rules require fit in naturally between the steps Euclid actually makes. **Eu** actually fills in gaps in Euclid's proofs. It does not alter their structure completely.

Thus, for a rigorous foundation of Euclid's proofs, one need not dig all the way down to modern logic. The epistemological interest of this lies in the fact that a rigorous foundation for a mathematical subject provides a picture of what justification amounts to in the subject. It sets standards by which a proof is complete, and so marks the point where a defender of the proof is released from the obligation to provide further justifications. When she displays the full proof, she has hit rock bottom. There is nothing more for her to do. If something like Hilbert's axiomatization is understood as the rigorous foundation for Euclid's proofs, there is a great deal more for Euclid to do, and the idea that Euclid succeeded at proving anything becomes strained. If on the other hand **Eu** is taken as providing a rigorous foundation for the proofs, the burden on Euclid is less severe. To furnish a complete proof, Euclid need only verify that the features he reads from the constructed from the diagram are general according to **Eu**.

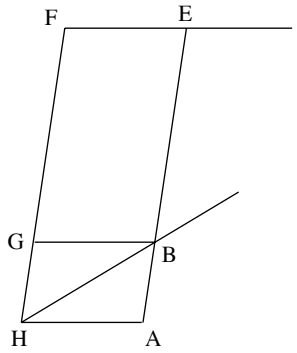


Figure 15. Diagram for proposition 44.

This does not imply, of course, that Euclid and his contemporaries actually carried such verifications out. Clearly, it cannot be plausibly maintained that the rules of **Eu** were followed *exactly*. That is, it is definitely not the case that with every proof in the *Elements* an experienced ancient geometer mentally rehearsed the **Eu** formalization of the proof. The details of the formalization are much too specific. Yet a case can be made, I believe, that constraints roughly similar to those imposed by **Eu** played a part in ancient geometric practice.

The general lesson of **Eu** is that care must be taken in considering positional relationships between diagrammatic elements which by the construction are not directly dependent on one another. In terms of the immediately precedes relation, x is not directly dependent on y if y does not immediately precede x. For want of a better term, call such pairs of elements *unlinked*. The diagram cannot help but display *some* relationship between unlinked elements, but one must refrain from accepting the manifest relationship uncritically. Additional considerations are necessary to confirm that it holds generally. Though Euclid does not exhibit this scruple at every opportunity, there is evidence that he does possess it. In proposition 44, for instance, the construction calls for the segments HB to be extended in the configuration of figure 15. It is clear *in this particular diagram* that HB and FE will meet in an intersection point. But Euclid makes sure to establish this as a general fact in the text with the parallel postulate. The segment HB is not directly dependent on FE, nor is FE directly dependent on HB. Thus their intersection cannot be taken for granted. Something similar occurs in proposition 47 (the Pythagorean theorem) when Euclid takes pains to argue that the perpendiculars constructed on opposite sides of a segment lie on a straight line.

If this is correct, and an awareness of the dependencies between constructed elements guided how ancient geometers read diagrams,

we are then able to flesh out Mueller's assertion concerning the all-triangles-are-isosceles fallacy. The reason that there is little danger that an experienced geometer would be seduced into accepting the putative proof is that crucial elements in the diagram are unlinked. Specifically, the fact that F lies on the segment AB and G lies on the segment AC in figure 1 allows the false result to go through. Yet scrutiny of the construction reveals that neither F and AB are directly dependent one another. The same is true of G and AC. We thus cannot take the position of these elements in the diagram at face value. We need further reasons to accept this aspect of the diagram. We cannot, of course, find any. So the proof does not succeed.

Euclid was not forthcoming with regard to all the reasons which can establish the generality of positional relationships in his diagrams. **Eu** fills in this lacunae with its positional proof rules. The principles fall, roughly, into three groups. One lists the conditions under which intersection points exist. An example is the rule which justifies the existence of the intersection point G above. The way various elements of a diagram can function as a frame of reference form the basis for another group. Finally, a third group grounds the generality of a segment's position within a figure on the convexity of the figure. With these principles one can recover much of the mathematics in the *Elements*. One cannot recover all of it. **Eu** lacks the resources to represent the theory of ratios Euclid develops in book V; nor can it represent the number theory and solid geometry developed in later books. The goal of the proof system is to formalize the elementary plane geometry in books I-IV.[9]

There are places where the **Eu** version of a proof seems needlessly detailed next to Euclid's original. This is usually attributable, however, to the fact that **Eu** is a modern formal system, subject to Frege's ideal. Because within **Eu** one can only use the rules which have been laid out in advance, 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.[10]

The obvious geometrical legitimacy of these rules is attested by the fact that many correspond naturally to axioms which appear in modern synthetic axiomatizations. For instance, the rule already discussed in connection with proposition 2—whereby one reads off the intersection of a ray and circle as general—is in propositional form axiom $A13'$ in Tarski's theory $\mathcal{E}_2''$ (Tarski, 1959). And an **Eu** rule whereby one reads off

the intersection of two lines as general is a more specific version of the Pasch axiom. If one is to engage in synthetic Euclidean geometry, one has to acknowledge certain simple topological invariances of geometric configurations at some point. One can privilege a handful as basic from the outset, and prove the rest, as it is done in a modern axiomatization. Or one can acknowledge them during proofs, by inspecting a diagram according to its construction, as **Eu** portrays Euclid as doing it. On **Eu**'s account, geometric diagrams provide a stable, reliable tool with which certain topological invariances can be checked. It is for this reason, perhaps, that Euclid did not feel the need to provide the axioms which from our modern, Fregean standpoint seem to be missing from his foundation of geometry.

4. Conclusion: illustrating and proving

The standard position against the mathematical significance of pictures in proofs resonates with a broader position on the relationship between picture and text. The natural way to fix the identity conditions of a non-mathematical book is via its sentences—i.e. their order, their arrangement into paragraphs and chapters. The book may come with illustrations, and these will influence how the book is experienced. But when a different edition comes out, with different illustrations or none at all, we would not regard it as a different book. The sentences carry the content of the book, and so define what it is. The illustrations are incidental.

The idea carries over into mathematical books smoothly. Just as an illustration in a work of fiction can depict a scene, a mathematical illustration can depict the relations between various concepts in a proof. Yet just as the story survives without the illustration, so does the proof. A proponent of this view will concede that the illustration may be invaluable in helping one get a mental grip on the proof. But he will go on to maintain that the real activity of the proof is in its inferences, in the transition from premise to conclusion. Pictures may illustrate the transitions, but they are not the means by which the transitions are made.

The analogy has its merits. Pictures which serve only to illustrate proofs pervade mathematical texts. A fitting example, in the present context, is Hilbert's *Foundations of Geometry*. The triangles and circles which fill its pages play no part in Hilbert's derivation of theorems from his axioms. Yet the role of pictures in the *Foundations* does not generalize to all mathematical contexts. Pictures can operate as a means of inference, as **Eu** demonstrates.[11] The proofs of **Eu** go through

the diagrams. The diagrams record positional information and carry positional inferences. And so we cannot dismiss out of hand a proof as incomplete because it relies on a picture.

Acknowledging this forces upon us a difficult question. The view that only lists of sentences prove draws a clear line between proof and non-proof. The line loses its sharpness when we allow proofs to have diagrams or pictures. How should we re-draw it so that Euclid's triangles are separated from Hilbert's? More specifically, are there any general features of **Eu**'s account of Euclid which could serve to distinguish genuine picture proofs from mathematical illustrations?

It is helpful in addressing this question to consider another one first: what about a symbol distinguishes it as pictorial or diagrammatic? Keith Stenning has done some useful work on the question in (Stenning, 2000). He theorizes that the characteristic trait of diagrammatic argument is its *agglomerative* nature. Proofs are commonly conceived of as consisting of lines. The symbols of one line are transformed via inference rules into another line until the desired line is obtained. A paradigm is the solution of an equation by means of elementary high school algebra. The equation is manipulated in a series of steps until the unknown variable is isolated. Each manipulation produces a new configuration of symbols, connected to but strictly separate from the preceding symbols. In contrast, a diagrammatic proof sequence proceeds cumulatively. Each step does not leave the previous configuration behind but builds on it. For example, in the *Elements*, we do not see the state of the diagram at each stage in the construction. We only see the final result.

Given this, one may wonder how diagrams can support useful inferences. A useful proof symbolism ought to allow us to derive relationships which are not immediately transparent. Yet a diagrammatic symbol, being agglomerative, is packed with information which *we have built into it*. The symbol has been constructed according to a stipulated set of conditions. If the situation is analogous to elementary algebra, it can only express those conditions. After we construct an equation expressing a known condition on an unknown quantity x, we go on *to change* the equation to learn what x is. But with an agglomerative diagram, we do not change anything. How then can it yield any useful information?

The answer is simply that the situation with diagrams is not analogous with algebra. A diagram can provide what Shimojima refers to in (Shimojima, 1996) as a *free ride*. A diagrammatic symbol constructed according to a certain set of conditions can express, automatically, conditions *not* in the set. Consider the order relations of points on a line. Suppose we want to express the condition that B lies between A

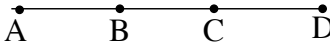


Figure 16.

and C, and C lies between B and D. In a modern theory of geometry possessing a betweenness relation B, we would accomplish this with the sentence

$$B(ABC) \ \ \& \ \ B(BCD)$$

It would then be an axiom or theorem of the theory that

$$\forall\, xyzw \quad ((B(xyz) \ \ \& \ \ B(yzw)) \longrightarrow B(xzw))$$

We could then infer that

$$B(ACD)$$

In **Eu**, we express the betweenness conditions in a single diagram like that of figure 16. We do not have to do anything more with the symbol to see that C lies between A and D. It comes for free once we formulate the first two betweenness conditions diagrammatically.

The capacity of Euclid's diagrams to give free rides is, I maintain, what classifies them as genuine proof symbols. A diagram is constructed to express a certain set of positional conditions. Once it is constructed, one sees in the very same diagram that other positional conditions must hold. In proposition 35, for example, the containment relations which underlie the conclusion are not stipulated at the outset of the proof. These, rather, are pointed out after a diagram is constructed to express positional relationships between two parallelograms. Things are more complicated with diagrams resulting from a construction. Not all the containment relations which come for free can be regarded as holding in general. But some can, as **Eu** confirms, and Euclid restricts himself to these.

Many pictures which appear with mathematical proofs are absent any free-rides. They depict only the relationships which have already been proved, linguistically, and so sit outside the proof. This is what is happening, for example, in the following proof/picture pair from Munkres' *Topology*. The theorem is that if X is a compact Hausdorff space, and every point of X is a limit point of X, then X is uncountable. The first half of the proof is:

First we show that, given a (nonempty) open set U of X, and given $x \in X$, there exists a (nonempty) open set V contained in U such that $\overline{V}$ does not contain x.

The point x may or may not be in U. But in either case, we can choose a point y in U that is different from x. This is possible if x is in U because x is a limit point of X (so that U must contain a point y different from x). And it

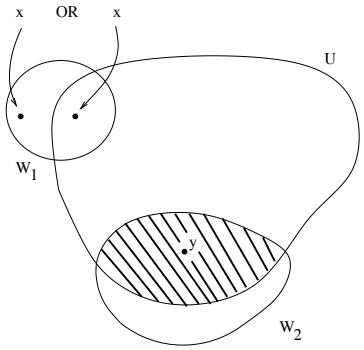


Figure 17.

is possible if x is not in U because U is nonempty. Let W_1 and W_2 be disjoint neighborhoods of x and y, respectively; then $V = U \cap W_2$ is the desired open set, whose closure does not contain x. See Figure 16.

Directly following this appears 'figure 16' (figure 17 here). No topological relationship is read off from the picture. There is none to be read off. All it is meant to do is illustrate the various membership and subset relations between the points and sets of the proof. It thus performs a valuable function. Unifying all the proof's objects into one surveyable image makes the proof easier to grasp. Yet it is important to distinguish this function from that fulfilled by Euclid's diagrams. Nothing is inferred from Munkres' picture. It sits apart from the proof's line of reasoning. Euclid's diagrams, in contrast, support inferences. His proofs travel straight through them.

The diagrams in Euclid and pictures in Munkres exemplifies the contrast between pictorial proof and illustration sharply. Yet the status of other cases with respect to the proof/illustration dichotomy is not so clear. For one thing, the presence of a free-ride in a mathematical picture may be debatable. Consider again the picture of figure 18 and the intermediate value theorem. One may claim that the appearance of an intersection point between the curve and the line is the result of a free-ride. A contrary position is that the intersection point 'appears' only after norms for reading the picture have been fixed. As noted, the picture can be thought to depict a situation where the set of points on the line is not Cauchy complete. Fixing the norms so as to rule out such a possibility seems tantamount to assuming the theorem without proof. There is then no distance for a free-ride to transverse.[12]

Alternatively, a picture may relate two distinct mathematical conditions, but the way it does so may be thought to be merely suggestive rather than mathematically sound. Consider the picture of figure 19 as a candidate proof for the equation in the caption. Grasping the

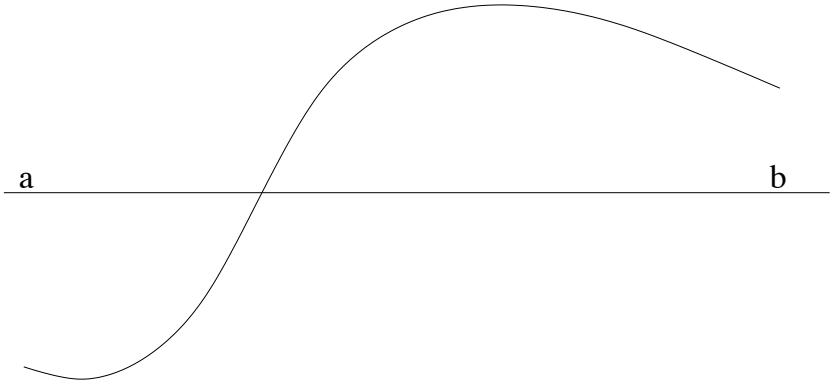


Figure 18.

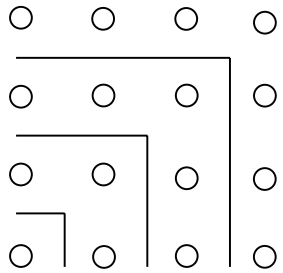


Figure 19. $1 + 3 + 5 + ... + (2n - 1) = n^2$

relevance of the picture to the equation involves, I believe, something like a free-ride. The picture is first seen to express the sum

$$1 + 3 + 5 + 7$$

by the way the dots are partitioned. It is then seen to be a square with sides of length four. (The free-ride could go in the other direction as well. That is, it can first be seen that all the dots form a square, and then it can be seen that the square spits into a sequence of odd numbers.) So the picture at the very least shows that

$$1 + 3 + 5 + 7 = 4^2$$

But does it show the equation

$$1 + 3 + 5 + ... + (2n - 1) = n^2$$

holds for *all* numbers n? It all seems to come down to whether or not the picture shows what a modern proof by induction would require. Namely, the picture needs to establish that the equation is preserved when we move from the nth to the $(n+1)$th case. If we understand the

dot sequence of odd numbers increasing indefinitely up and to the right, does the picture show that each new layer results in a square with sides one unit larger? (Alternatively, if we understand the square as growing indefinitely up and to the right, does the picture show that each new layer of the growing square has two more dots than the previous layer?) By focusing on sub-squares of the four by four square, we can convince ourselves of this for the first three stages of growth. Whether we are justified in extrapolating the pattern to *all* stages of growth, does not seem obvious or non-controversial.

The questions raised by the last two examples deserve further investigation. And there is no reason to stop with them. The enumerability of ordered pairs of natural numbers is often shown with an array, as is the non-enumerability of the powerset of natural numbers. Are the arrays used to present both proofs legitimate vehicles of proof? Why or why not? A whole field of modern mathematics—category theory—uses diagrams extensively. Do any general principles underlie the 'diagram chases' one often encounters in its proofs? And if so, what relation do these principles have to those behind Euclid's diagrammatic proofs? Abandoning the orthodox view of pictures and proof opens up to the philosophy of mathematics a rich range of phenomena. Exploring it promises not only a deeper understanding of the way pictures can prove but, I believe, a deeper understanding of the general nature of mathematical proof.

Appendix: Syntax of **Eu**

As mentioned, the atomic claims of **Eu** have two components: a diagram Δ and a metric assertion A. The purpose of this appendix is to sketch the syntactic structure of these components.

The syntax of a metric assertion is very close that of atomic sentence in predicate logic. There are six relation symbols

$$\equiv_{seg}, <_{seg}, \equiv_{angle}, <_{angle}, \equiv_{area}, <_{area}$$

each representing either the equality or inequality of a certain magnitude. Just as in predicate logic, the relation symbols serve as placeholders for variables $A, B, C, D....$ The arity of the $\equiv_{angle}$, for instance, is 6. A well-formed atomic metric assertion with $\equiv_{angle}$ is

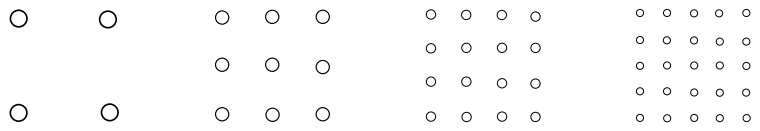
$$ABC \equiv_{angle} DEF$$

There is one connective—&—with which two well-formed metric assertions can be combined into one. An example of such a metric assertion is

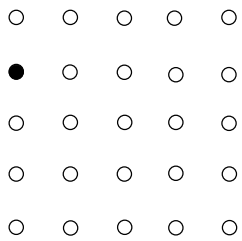
$$ABC \equiv_{angle} DEF \quad \& \quad AB <_{seg} BC$$

Finally, there are two constant metric assertions: $\perp$ and Θ. The first represents contradiction, the second represents the empty metric assertion.

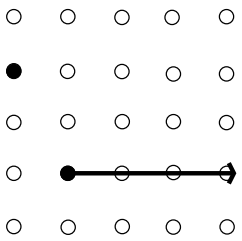
The syntactic structure of diagrams in **Eu** has no natural analogue in standard logic. Their underlying form is a square array of dots of arbitrary finite dimension.



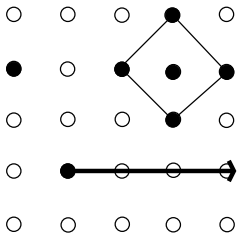
The arrays provide the planar background for an **Eu** diagram. Within them geometric elements—points, linear elements, and circles— are distinguished. A point is simply a single array entry. An example of a diagram with a single point in it is



Linear elements are subsets of array entries defined by linear equations expressed in terms of the array entries. (The equation can be bounded. If it is bounded one one side, the geometric element is a ray. If it is bounded on two sides, the geometric element is a segment.) An example of a diagram with a point and linear element is



Finally, a circle is a convex polygon within the array, along with a point inside it distinguished as its center. An example of a diagram with a point, linear element and a circle is

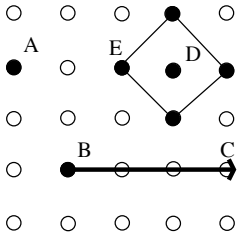


The size of a diagram's underlying array and the geometric elements distinguished within it, comprise a diagram's identity. Accordingly, a diagram in **Eu** is a tuple

$$\langle n, p, l, c \rangle$$

where n, a natural number, is the length of the underlying array's sides, and p, l and c are the sets of points, linear elements and circles of the diagram, respectively.

Like the relation symbols which comprise metric assertions, the diagrams have slots for variables. A diagram in which the slots are filled is termed a labeled diagram. The slots a diagram has depends on the geometric elements constituting it. In particular, there is a place for a variable beside a point, beside the end of a linear element (which can be an endpoint or endarrow), and beside a circle. One possible labeling for the above diagram is thus

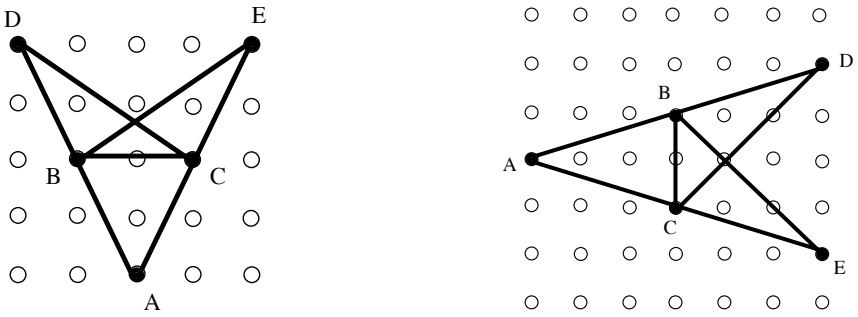


Having labeled diagrams within **Eu** is essential, for otherwise it would be impossible for diagrams and metric assertions to interact in the course of a proof. We can notate any labeled **diagram** as

$$\langle n, p, l, c \rangle [\vec{A}, R]$$

where $\vec{A}$ denotes a sequence of variables and R a rule matching each variable to each slot in the diagram.

Understood as such, labeled diagrams carry too much information. For any one labeled diagram there will be an infinite number of others which convey the same co-exact relationships, and so support the same inferences in a Euclidean proof. For example the differences between the two labeled diagrams



would have no bearing on what one could do with them in a Euclidean proof. It does not matter that the second one is rotated counterclockwise, and the dimension of its underlying array is 7 rather than 5. It expresses the same co-exact conditions with respect to A, B, C, D and E, and so would play the same role in an argument.

To group such diagrams together, **Eu** possesses an equivalence relation $\sim$ which abstracts away all irrelevant information. Roughly, if we understand a diagram as a structure whose objects are its points, linear elements, and circles, the equivalence relation amounts to a characterization of what it is for two labeled diagrams to be isomorphic. Such an isomorphism is a bijection between the geometric elements of two labeled diagrams, whereby the following relations are preserved: point p_1 is on/not on line l_1 or circle c_1; points p_1, p_2 are the same/different side of line l_1; point p_1 is inside/oustide circle c_1; lines l_1 and l_2 intersect/don't intersect; line l_1 does not intersect/is tangent to circle c_1; line l_1 intersects circle c_1 one/two time(s); circles c_1 and c_2 do not intersect; intersecting circles c_1 and c_2 together define k regions. For the details see pages 34 to 40 of (Mumma, 2006).

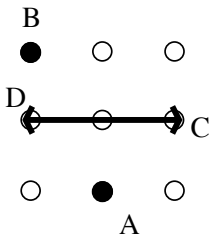
According to **Eu**'s analysis, then, a **diagram**

$$\langle n, p, l, c\rangle[\vec{A}, R]$$

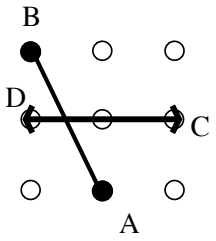
modulo the equivalence relation $\sim$ characterizes what a diagram in the *Elements* is as a proof-symbol. The definition is meant to capture all the

co-exact relationships expressed by Euclid's diagrams. Its suitability is not transparent, however. A few remarks addressing some possible misgivings with the definition are thus in order.

First, one may worry that as discrete objects **Eu**'s diagrams will fail in general to produce the intersection points which appear in Euclid's diagrams. For instance, in the diagram



we can join the points above and below the line to obtain the diagram



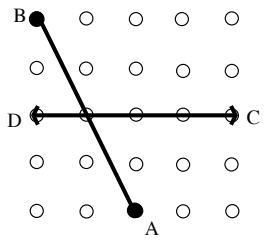
Given what this diagram is intended to represent, we ought to be able to produce an intersection point between the segment AB and the line DC. But the underlying array of the diagram is too coarse. An array entry does not exist where we want a point to be.

Within a proof carried out in **Eu**, this can always be dealt with by *refining* the diagram. The equation which characterizes a line (and the circumference of a circle) is linear, expressed in terms of the coordinates of the array entries. Since the arrays are discrete, the coefficients of the equation are always integers. Thus, the solution for two equations characterizing geometric elements of a diagram will always be rational.

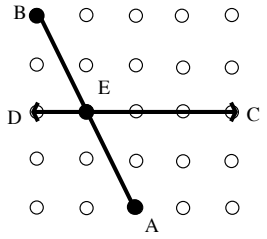
This means that if two geometric elements ought to intersect but don't in a diagram, we can always find an equivalent diagram where they do. The equivalent diagram will just be the original diagram with a more refined underlying array. In particular, if the original diagram has dimension n and the solution between the two equations is a rational

with an m in its denominator, the new diagram will have dimension $mn - 1$.

For the diagram above, then, adding the desired intersection point is a two step process. First, the diagram is refined to an equivalent diagram of dimension 5.



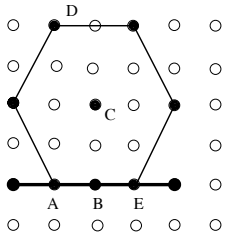
Then the intersection point is added.

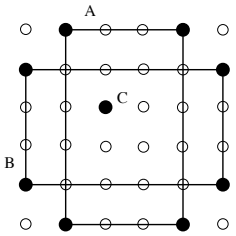


Another natural worry has to do with the circles of diagrams. The circles which appear in Euclid's diagrams actually appear circular. The circles of diagrams, however, are rectilinear. If Euclid exploits the circularity of his circles in his proofs, then the diagrams of **Eu** would fail to capture this aspect of Euclid's mathematics. Euclid, however, never does this. All he seems to assume about circles is that they have an interior. Many features of circles, which would seem to follow from their circularity, are actually proved by Euclid in book III. For instance, he proves that it is impossible for a line and a circle to intersect in more than two points. He also proves that it is impossible for two circles to intersect in more than two points. This is done in propositions III,2 and III,10, respectively.

And so, it is actually a virtue of **Eu**'s diagrams that they are not circular. Because they are convex polygons, they can represent the physically impossible situations of propositions III,2 and III,10. (See

the two diagrams below.) This makes it possible for the reductio proofs of these propositions to be carried out in **Eu**.





Notes

[1] Both Brown in (Brown, 1997) and Dove in (Dove, 2002) defend this position, albeit in different ways.

[2] One may object that holding Euclid to Frege's standards is unfair, as both were pursuing different foundational aims. Working this objection out means specifying the kind of foundational project which does not require that all modes of inference be specified in advance. Euclid at the very least considered it important to specify some modes of inference in advance. He takes pains to tie steps in his proofs to his common notions and postulates, or to theorems previously proved from these. It is thus natural to understand Euclid as driven by a desire to push such a project to the limit. Why stop half-way? In the next section, after presenting **Eu**, I consider the question in the light of its analysis.

[3] Frege, ironically, vocally opposed such an interpretation of the new axiomatic theories. In his famous correspondence with Hilbert and a sequence of articles, he argued that starting with abstract, undefined primitives was illicit—i.e. the meaning of a theory's primitives must be fixed before they are placed into axioms. For him, presumably, this requirement could not be abandoned for the sake of a precise yet abstract elucidation of what one can and cannot do in geometric arguments.

[4] Both Netz in chapter 6 of (Netz, 1999) and Norman in chapter 10 of (Norman, 2005) seem to be moving in the right direction. Yet their comments fall short of a complete account. On the other hand, Miller develops a formally impeccable account of Euclid's diagrammatic reasoning, similar to that of **Eu**, in (Miller, 2007). The work fails however to be satisfactory in that it demands consideration of a staggering number of cases, all but a few of which are ever considered by Euclid (not least because most of the cases are not physically realizable). For a discussion of the drawbacks of Miller's approach, see my review of Miller's book (Mumma, 2008).

[5] For all the formal details of **Eu**, see (Mumma, 2006).

[6] This is admittedly only a suggestive characterization of the distinction between exact and co-exact. Manders' definition categorizes the co-exact exact "as those conditions unaffected by some range of every continuous variation of the diagram" and the exact as "those which, for at least some continuous variation of the diagram, obtain only in isolated cases." (Manders, 2008). The formal characterization of co-exactness within **Eu** takes the form of an equivalence relation between its diagrammatic symbols. See the appendix for a brief description of this relation $\sim$. Its complete description occurs on pages 34-40 of (Mumma, 2006).

[7] The principle is perhaps close to what Kant is talking about when he speaks of the "the universal conditions of the construction" in the passage quoted above. For a discussion which relates Kant's philosophy of mathematics to Euclid's geometric constructions, see (Shabel, 2006).

[8] Here then is a specific instance of **Eu**'s treatment of geometric continuity. How it compares to the modern treatment of continuity is too involved a question to discuss fully here. Some brief comments are possible, however. In **Eu**, the intersection points which in a modern theory are secured by a continuity axiom appear directly in its diagrammatic symbols. One must employ rules (such as the one just described) to establish that such a point as exists in general. But the *particular* existence of an intersection point is read directly from a diagrammatic symbol. If there is a crossing in a **Eu** diagram—between two circles, two lines, or a line and a circle—a point of intersection exists in the diagram or an equivalent one. Building intersection points into diagrams in this way may be thought to be illicit. As brought out in the discussion on picture proofs of the intermediate value theorem, we need not understand the crossing of curves in a diagram to indicate a point. And so in a sense a principle of continuity is present in **Eu**. But it is present in a way which is different from its presence in modern theories. It does not serve to rule out other mathematical possibilities, in the way that a modern continuity axiom rules out certain models. Rather it is embedded into the way the proof system's symbols are used.

[9] It is not known at present how exactly **Eu** relates to standard axiomatizations of Euclidean plane geometry. It is straightforward to check the relative consistency of **Eu** to any such axiomatization. The conditionals $\Delta_1, A_1 \longrightarrow \Delta_2, A_2$ have a natural interpretation in terms of first order formulas made up of the theory's primitives. The soundness of **Eu**'s proof rules is then easily checked in terms of this interpretation. An open question, however, is the strength of **Eu** with respect to a modern theory. That is, can **Eu** prove everything an axiomatization can prove? One can give a partial, negative answer immediately. Since it is only possible to solve quadratic equations with the intersection of lines and circles, it follows that one cannot prove anything which, when interpreted in a axiomatization, requires a continuity assumption stronger than the condition that the plane's underlying field is closed under square roots. Appropriately, fields which satisfy this property are termed *Euclidean*. Thus, an axiomatization whose continuity assumption guarantees Euclidean fields but nothing more sets an upper bound for what is provable in **Eu**.

To address the question of the completeness of Euclid's diagrammatic method, a proof system inspired by **Eu** has recently been developed. The new system, termed E, is complete. That is, it can be shown that modern axiomatizations of elementary geometry are conservative extensions of E. For a description of E and the completeness proof, see (Avigad et al., 2009). Though inspired by **Eu**, the system E is not so similar that its completeness settles the question of **Eu**'s completeness.

[10] The picture of Euclid's method which thus emerges seems akin to the account of proof Azzouni provides in (Azzouni, 2004), whereby a mathematician introduces axioms into a proof *as* she is carrying out the proof. In Azzouni's words, she *augments* the proof system she is working within. How exactly this process of augmentation relates to **Eu**'s account of Euclid is a question worth exploring, for it seems a case could be made that the account also supports the view Azzouni opposes in the piece. The view, defended by Yehudi Rav in (Rav, 1999) and (Rav, 2007), is that mathematical proof is grounded on a formalism-independent knowledge of what the terms in the proof mean mathematically. Accordingly, the Euclidean inferences

codified as new rules within **Eu**'s framework could be said to be based on the way diagrams bring out the meaning of the geometric concepts involved in the proof.

[11] **Eu** is not the first to demonstrate it. As mentioned, formal diagrammatic systems of deduction have already been developed. **Eu** contributes to the case for picture proofs by formalizing picture proofs of enormous historical importance. Not only is rigorous reasoning with pictures possible. If **Eu**'s account of the *Elements* is accurate, rigorous reasoning with pictures was once commonplace and played a huge part in shaping mathematical thought and practice.

[12] See note 8 for a brief discussion of how fixing the norms for diagrammatic symbols differs from assuming a continuity axiom.

Acknowledgements I am grateful to Jeremy Avigad, Clark Glymour, Ken Manders, Dirk Schlimm, Dana Scott and two anonymous referees for helpful comments on earlier drafts.

References

Avigad, J., Dean, E., and Mumma, J. (2009). A Formal System for Euclid's Elements. To appear in *The Review of Symbolic Logic*.

Azzouni, J. (2004), Derivation Indicator View of Mathematical Practice. *Philosophia Mathematica, 12(2)*, 81-106.

Barwise, J. and Hammer, E. (1996). Diagrams and the Concept of a Logical System. (In G. Allwen & J. Barwise (Eds.), *Logical Reasoning with Diagrams*. New York: Oxford University Press.)

Brown, J. R. (1997), Proofs and Pictures. *British Journal of the Philosophy of Science, 48(2)*, 161-180.

Dove, I. (2002), Can Pictures Prove? *Logique & Analyse, 45(179-180)*, 309-340.

Frege, G.(1964). *The Basic Laws of Arithmetic, translated by Montgomery Furth.* (Berkeley: University of California Press.)

Hilbert, D. (2004). *David Hilbert's Lectures on the Foundations of Geometry: 1891-1902*. (Edited by M. Hallet & U. Majer. Berlin: Springer.)

Klein, F. (1939). *Elementary Mathematics from an Advanced Standpoint*. (Dover Publications.)

Leibniz, G. (1949). *New Essays Concerning Human Understanding*. (LaSalle, Illinois: Open Court Publishing.)

Mancosu, P. (2005). Visualization in Logic and Mathematics. (In P. Mancosu, K. F. Jorgensen, & S. A. Pedersen (Eds.), *Visualization, Explanation, and Reasoning Styles in Mathematics*. Springer.)

Manders, K. (2008). The Euclidean Diagram. (In P. Mancosu (Ed.), *Philosophy of Mathematical Practice*. Oxford: Clarendon Press.)

Miller, N. (2007). *Euclid and His Twentieth Century Rivals: Diagrams in the Logic of Euclidean Geometry*. (Stanford, California: Center for the Study of Language and Information.)

Morrow, G, editor. (1970). *Proclus: a Commentary on the First Book of Euclid's Elements*. (Princeton: Princeton University Press.)

Mueller, I. (1981). *Philosophy of Mathematics and Deductive Structure in Euclid's Elements*. (Cambridge, Massachusetts: MIT Press.)

Mumma, J. (2006). Intuition Formalized: Ancient and Modern Methods of Proof in Elementary Euclidean Geometry. PhD Dissertation, Carnegie Mellon University. Online at www.andrew.cmu.edu/jmumma

Mumma, J. (2008) Review of *Euclid and His Twentieth Century Rivals: Diagrams in the Logic of Euclidean Geometry. Philosophia Mathematica, 16(2)*, 256-264.

Netz, R. (1999). *The Shaping of Deduction in Greek Mathematics: A Study of Cognitive History.* (Cambridge: Cambridge University Press.)

Norman, J. (2005). *After Euclid: Visual Reasoning and the Epistemology of Diagrams.* (Stanford, California: Center for the Study of Language and Information.)

Pasch, M. (1882). *Vorlesungen über Neuere Geometrie.* (Leipzig: B.G. Teubner.)

Rav, Y. (1999), Why Do We Prove Theorems? *Philosophia Mathematica, 7(1)*, 5-41.

Rav, Y. (2007), A Critique of a Formalist-Mechanist Version of the Justification of Arguments in Mathematicians' Proof Practices. *Philosophia Mathematica, 15(3)*, 291-320.

Shabel, L. (2006). Kant's Philosophy of Mathematics. (In P. Guyer (Ed.) *The Cambridge Companion of Kant*, 2nd edition. Cambridge University Press.)

Shimojima, A. (1996). Operational Constraints in Diagrammatic Reasoning. In G. Allwen and J. Barwise (Eds.) *Logical Reasoning with Diagrams.* New York: Oxford University Press.

Shin, S. (1994). *The Logical Status of Diagrams.* (New York: Cambridge University Press.)

Stenning, K. (2000). Distinctions with Differences: Comparing Criteria for Distinguishing Diagrammatic from Sentential Systems. (In *Proceedings of the First International Conference on Theory and Application of Diagrams.* London: Springer Verlag.)

Tarski, A. (1959). What is Elementary Geometry? (In L. Henkin, P. Suppes, & A. Tarski (Eds.) *The Axiomatic Method, with Special Reference to Geometry and Physics.* Amsterdam: North Holland Publishing Company.)

Veblen, O. (1914). The Foundations of Geometry. (In *Monographs on Topics of Modern Mathematics, relevant to the elementary field.* Longsmand, Green, and Company.)