A Theory of Simplicity

Gilbert Harman, Michael Ranney, Ken Salem, Frank Döring
Jonathan Epstein, Agnieszka Jaworska

Cognitive Science Lab, Princeton University, 221 Nassau St., Princeton, NJ 08542.
609-987-2824, ghh@princeton.edu

The simplicity of a hypothesis for a person cannot be measured by the simplicity of the person's representation of that hypothesis (for example, the number of symbols used), because any hypothesis can be represented with a single symbol. A better measure of simplicity is the ease with which the hypothesis can be used to account for actual and foreseeable data. But it is also important to allow for different ways in which data might be represented. We suggest that the relevant ways of representing data are those ways in which the person is interested, *i.e.*, those representations that most directly help to answer questions the person wants to answer. In particular, we suggest that the simplicity of a hypothesis for a person is determined by the shortness of the connection between that hypothesis and the data that interest the person, as measured by the number of intermediate steps he or she needs to appreciate in order to appreciate the complete connection.

Keywords: simplicity, explanation, inference, hypothesis

Curve Fitting

Suppose that you have discovered that pressure on a certain sort of ceramic material affects the conductivity of that material and you are interested in being able to determine the relation between these two quantities. You obtain the following replicable data pairs: $P=1$, $C=2$; $P=3$, $C=6$; $P=4$, $C=8$; $P=9$, $C=18$. If we let $C=F(P)$, your data can then be expressed as follows: $F(1)=2$; $F(3)=6$; $F(4)=8$; $F(9)=18$. Given these data, you would normally take the most reasonable hypothesis to be (H1) $F(P)=2P$.

Why? Certainly, (H1) is compatible with your data in a way that various other hypotheses are not, for example:

(H2) $F(P)=3P$.

(H3) $F(P)=3-P$.

But there are infinitely many hypotheses that are like (H1) in being compatible with your data. For example,

(H4) $F(P)=2P+(P-1)(P-3)(P-4)(P-9)$.

What is it that makes you prefer (H1) to a hypothesis like (H4)?

The natural answer to this question is that (H1) is simpler and less ad hoc than (H4). But this answer raises two important and related issues. First, what makes one hypothesis simpler than another? Second, why should the simplicity of a hypothesis be taken as any sort of indication of the truth of the hypothesis?

The preparation of this paper was supported in part by research grants to Princeton University from the James S. McDonnell Foundation and the National Science Foundation. We are indebted to discussions with Paul Thagard and (some years ago) with Hugo Margain.

These are difficult issues. We are going to concentrate on the first and try to say what the simplicity of a hypothesis consists in. But we will also say a little about the second issue, why simplicity, so understood, should be used to select among hypotheses that account equally well for the data.

In the example we began with, the first question becomes this: In what way is (H1) simpler than (H4)? One difference between the hypotheses is that (H1) is expressed in 7 symbols while (H4) uses 28 symbols. So we might consider the following proposal: (H1) is simpler than (H4) in that it is shorter, and, more generally, the complexity (or lack of simplicity) of a hypothesis can be measured by the number of symbols used to express that hypothesis (Sober, 1975).

The trouble with this proposal is that hypotheses can be expressed in various ways. For any given hypothesis we can arbitrarily introduce notation that would allow us to express the hypothesis with very few symbols. For example, we have been using the symbols "(H1)" and "(H4)" to stand for these hypotheses, which shows that each can in fact be represented in 4 symbols. We could have used the numerals "1" and "4" instead, in that way representing each hypothesis with one symbol. Clearly, any hypothesis can be expressed using a single symbol in this way. So, if hypotheses differ in simplicity, this difference cannot be measured simply by counting the symbols used to express the hypotheses.

It might be suggested that "(H1)" and "(H4)" are abbreviations of longer expressions and that we must consider how many symbols it takes to express these hypotheses without using abbreviations. But how are we to tell when abbreviations have been used? How are we to determine what symbols can be used as primitive and undefined? Is it permitted to represent the product of 2 and P as "$2P$" or must we include an operator to represent multiplication as in "$2\times P$"? Can we use ordinary mathematical symbols like "$\times$", "+", and "−", or are these symbols to be defined in even more primitive terms?

One way to answer such questions is to postulate a system of mental representation. In this view, the complexity of a hypothesis for you depends on your actual mental representation of the hypothesis. (H1) is less complex than (H4) for you, even though each can be externally represented by a single symbol, if mental representation of (H1) uses fewer mental primitives than your mental representation of (H4). This is a psychological account of simplicity in the sense that the simplicity of a hypothesis is its simplicity for you, given the way in which you represent the hypothesis in Mentalese.

This leaves the difficult problem of trying to determine what form an inner system of mental representation takes. One test might be to consider how "natural" certain representations are. Consider the following two representations:

(H1) $F(P)=2\times P$.

(H4) $F(P)=2\$P$.

Here the operator "$\times$" stands for multiplication and the operator "$\$$" satisfies the following rule:

$$A\$B=A\times B+(B-1)\times(B-3)\times(B-4)\times(B-9).$$

The operator "$\times$" seems quite natural as an operator that is suitably represented by a single symbol, whereas the operator "$\$$" so defined may seem "unnatural"—one that is *not* suitably represented by a single symbol.

But this does not meet the objection already considered, since it does not preclude there being various ways to mentally represent any given hypothesis. Just as you can use a single symbol to stand for any given hypothesis while you are talking to someone else, you ought to be able to do the same thing in the systems of mental representation you use for thought. If you can, the complexity of a hypothesis for you is not a function of the number of mental symbols you use to

express the hypothesis in your inner language of thought.

In other words, it is unclear why you can't introduce a mental symbol for the operation "*$*" so that you can *mentally* represent (H4) as "*F(P)=2$P*"? That may not be a natural way of representing (H4) in some sense, since the symbol "*$*" is unfamiliar, but that fact cannot prevent "*F(P)=2$P*" from being a possible mental representation of (H4).

One thing that may make this particular representation seem unnatural is that, in order to use it, you need to unpack it. If you are to use this representation of (H4) to calculate the value of "2$7," you need first to translate the representation into the equivalent form "2×7+(7–1)(7–3)(7–4)(7–9)." You can then see that this is equivalent to "14+6×4×3×(–2)," which is equal to "14–6×4×3×2," which is equal to "14–24×6," which is equal to "14–144," which is equal to "–130."

This calculation is more complicated than simply multiplying "2×7." That is what seems to make (H4) a more complicated hypothesis. Notice that the complexity of the calculation has little to do with how (H4) is represented. No matter how the hypothesis is represented, a relatively complicated calculation is needed in order to determine the value $F(P)$ for particular values of P. The complexity of the hypothesis depends not just on how it is represented, but rather on the calculations needed to use the hypothesis.

Let us generalize this idea. Instead of measuring the complexity of a hypothesis by the number of symbols used to represent the hypothesis in either English or Mentalese, we can measure complexity by the amount of processing required to use the hypothesis in order to connect it with the data -- or more generally, the amount of processing that is needed to use the hypothesis in whatever way the hypothesis is to be used. As we have seen, this would count (H1) as simpler than (H4) on the grounds that determining $F(P)$ for various P requires fewer calculations under (H1) than under (H4)

This is still a "psychological" account of simplicity and complexity, but the psychological simplicity or complexity of a hypothesis is taken to depend on the complexity of certain mental processes rather than depending directly on the complexity of particular mental representations. Changing notation to make a hypothesis shorter will not make that hypothesis simpler, unless the new notation allows a psychologically shorter derivation of the data.

Immediately Obvious Steps

The complexity of the connection between a hypothesis and an implication of that hypothesis might be measured by the number of steps needed for a person to recognize that connection. For example, the connection between "2×7" and "14" is relatively immediate once you have mastered the multiplication table for "2". No further calculation is needed to see the connection between "2×7" and "14." There is not an equally immediate connection between "(2×7)+(7–1)×(7–3)×(7–4)×(7–9)" and "–130". You can appreciate this connection only by virtue of noting intermediate steps such as those mentioned above, namely, "14+6×4×3×(–2)," "14–6×4×3×2," "14–24×6," and "14–144."

The relevant steps are "immediately obvious steps." That is, you can recognize that each step follows from a previous step without having to recognize some intermediate steps. We claim that this notion of an "immediately obvious step" is quite important for the theory of inference. It is important not only for understanding simplicity, but for other reasons as well.

Harman (1987) argues that the concepts of logic might be at least partially explicated in terms of immediately obvious steps of implication between propositions involving these concepts. The operation of logical conjunction, "*&*," has the property that a conjunctive proposition "*A&B*" implies and is immediately implied by its conjuncts "*A*" and "*B*". This is not true of the

more complex but logically equivalent concept "%" explicated as follows: "*A %B*" = "not either not *A* or not *B*." If the meaning of a concept depends in part on the immediately obvious implications of propositions involving the concept, logically equivalent concepts can differ in meaning.

It is worth noting that steps that are not immediately obvious at one time can become immediately obvious at some later time. For example, at one time it may not have been obvious to you that 4×7 is 28. You may have had to note such intermediate steps as these: 4×7; 7+7+7+7; 14+7+7; 21+7; 28. At a later time, you can immediately see the equivalence. Your recognition of it does not depend on the conscious recognition of such intermediate steps. In our view, when you came to acquire this sort of mathematical skill, hypotheses involving multiplication became psychologically simpler for you. (And your concept of multiplication underwent a change.)[1]

Goodman's Grue Bleen Problem

We suggest further that the complexity of a hypothesis is relative to what you are interested in, that is, it depends on the complexity of the connections between the hypothesis and things in which you have an interest. In order to illustrate this suggestion, let us look at one aspect of Goodman's famous grue bleen puzzle (Goodman 1965), which is a variant of the curve fitting problem.

Suppose that you have examined a variety of emeralds and have determined in each case that the color of the emerald was green at the time of observation. Suppose that it is now January 1, 1990 and consider the following two hypotheses:

(H5) All emeralds are (always) green.

(H6) All emeralds *either* (a1) are first observed before the year 2000 and (a2) are (always) green *or* (b1) are not first observed before the year 2000 and (b2) are (always) blue.

These hypotheses agree about all emeralds first observed before the year 2000 but disagree about emeralds that do not get observed by then. (H5) implies that such emeralds are green. (H6) implies that such emeralds are blue.

Although both hypotheses would account for your evidence, you prefer (H5) over (H6) on grounds of simplicity. The measure of simplicity we have proposed seems to account for this, since the connection between any item of evidence and (H5) would seem to be immediate whereas the connection between an item of evidence and (H6) would seem to be mediated by several steps:

> All emeralds are either observed before the year 2000 and are (always) green or are not observed before the year 2000 and are (always) blue. So, this emerald is either observed before the year 2000 and is (always) green or is not observed before the year 2000 and is (always) blue. This emerald is observed before the year 2000 So, it is not

[1] In supposing that the equivalence between 4×7 and 28 is immediately obvious, representing a "single step," we do not deny that there may be a level of analysis at which the recognition of this equivalence involves several steps. For example, you could look up this product in a table in your memory. You could first locate the table, then locate the relevant column and row. It is possible that, at this deeper level of analysis, the recognition of the equivalence between 4×7 and 28 requires more steps and so more time than the recognition of the equivalence between 2×10 and 20, even though both of these equivalences are immediately obvious at a coarser level of analysis. [This issue is discussed in, e.g., Groen & Resnick (1977) and Ashcraft & Stazyk (1981).] We do not at this time know how to give a precise characterization of the relevant deeper level of analysis. All that concerns us at this stage is the difference between your situation before and after you have mastered the times tables. [Another complication is that a step of inference might be immediate even though the conclusion of that step is itself a complex argument-structure involving several steps of implication or explanation. See Harman et al. (1987).]

the case that this emerald is not observed before the year 2000 So, it is not the case that this emerald is not observed before the year 2000 and is (always) blue. So, this emerald is observed before the year 2000 and is (always) green. So, this emerald is green.

But more needs to be said about this case. Following Goodman, you can define a predicate "grue" as follows:

x is grue at t if and only if either (1) x is first observed before the year 2000 and is green at t or (2) x is not first observed before the year 2000 and is blue at t.[2]

Then you can represent this last hypothesis more briefly as follows:

(H6) All emeralds are (always) grue.

Since all emeralds observed so far are observed before the year 2000, all the observed emeralds are grue (at least when observed). And, while it is true that the hypothesis that all emeralds are green offers a simpler account of the fact that observed emeralds have so far been green, it is also true that the hypothesis that all emeralds are grue offers a simpler account of the fact that observed emeralds have so far been grue. Your situation with respect to the two hypotheses may therefore seem perfectly symmetrical, and it becomes unclear how you can take (H5) to be any simpler than (H6).

Here there is a temptation to return to counting symbols to measure complexity. Although you can use the term "grue" in order to express (H6) as compactly as (H5), it might be argued that this compact representation is an abbreviation of your original, much more complicated, representation of the hypothesis. In this view, "green" expresses a simple concept, whereas "grue" expresses a disjunctive concept.

But this attempt at a solution leads to the difficulties mentioned earlier. There is no objective way to determine when a representation should be counted as an abbreviation. Furthermore, Goodman points out that "green" and "blue" can be seen to express disjunctive concepts via the following definitions (where "bleen" is understood complementarily to the way in which "grue" is understood).

x is green at t if and only if either (1) x is first observed before the year 2000 and is grue at t or (2) x is not first observed before the year 2000 and is bleen at t.

x is blue at t if and only if either (1) x is first observed before the year 2000 and is bleen at t or (2) x is not first observed before the year 2000 and is grue at t.

It might be suggested that "green" is a term that can be applied purely on the basis of observation whereas "grue" is not, and that this has something to do with why we prefer (H5) to (H6). However, it is unclear how to make a general distinction between observational terms and other terms. (Is "emerald" an observational term?) Furthermore, it is an accident of this particular example that it involves a term that might count as observational. Many hypotheses in which we are interested do not make use of observational terms in this way and the same problem arises for them as for (H5). That is, we are not primarily concerned with hypotheses of the form, "All A's are B's," where "A" and "B" are observational terms. A distinction between what is directly observable and what is not cannot yield a general solution of Goodman's puzzle.

As we have already said, our proposal is that interests have a bearing on inferences. What conclusion you should reach depends in part on what question you are interested in answering.

---

[2] Notice that there are two time references, the time t at which we are considering what color x is and the usually quite different time at which x is first observed.

Normally, you are interested in what is blue or green, not in what is bleen or grue. (H5) is preferable to (H6) because (H5) gives a simpler route to the sorts of things you are interested in than (H6) does. Although (H6) provides a simpler route to conclusions about what things are grue or bleen, you are normally not directly interested in learning such things.

So, we suggest that the simplicity of a hypothesis for you is determined by the simplicity of the connection between that hypothesis and the data in which you are interested, as measured by the number of intermediate steps you need to consider in order to see the connection. Since you will sometimes accept a hypothesis just because it is the simplest of a group of hypotheses that equivalently account for the data, this suggestion implies that your interests can influence what conclusion you come to accept. A natural objection is that this must be a case of irrational wishful thinking. What can we say about that objection?

Wishful Thinking and the Relevance of Interests to Inference

Here it is important to distinguish reasoning that is aimed at what to believe, which (following Aristotle) we can call "theoretical reasoning," from reasoning that is aimed at what to plan to do, which we can call "practical reasoning." Clearly, your interests can legitimately help to determine what practical conclusions you should reach about what to do, so that is an obvious way in which your interest are relevant to your reasoning. But theoretical reasoning is not practical reasoning and we are now concerned with how your interests might affect what theoretical conclusions you are justified in reaching.

Consider a related way in which your interests can be relevant to your theoretical reasoning. Your interests help to determine what questions you have reasons to answer. In that way, your interests can legitimately affect which conclusions you will draw. At any moment, a vast number of conclusions follow trivially from your beliefs. But you are not equally justified in drawing each of those conclusions, since at best you will be interested in the truth of only a small number of them. Reasoning is subject to a principle of clutter avoidance. You should not clutter your mind with the trivial consequences of your beliefs, at least if there are certain questions you might be resolving in which you have an interest. This is not a general warrant for wishful thinking. The fact that you want a certain result to be true is not a reason to believe that it *is* true. Your interests can give you a reason to try to answer a particular question but they are irrelevant to what the answer is (except in special cases, for example, in which the question concerns your interests).

This is relevant to our proposal for resolving the grue-bleen problem. Our resolution appeals to your interests in order to determine what questions you want to answer, not what the answers are. We suggest that it is legitimate for you to accept the simplest account of the data in which you are interested, where simplicity is measured by the number of steps needed to get from hypothesis to data. You tend to be interested in whether certain things are blue or green, not in whether they are grue or bleen. You are normally interested in why observed emeralds are green in a way in which you are not so interested in why observed emeralds are grue. Even philosophers who are interested in why certain emeralds are grue are interested only because of an ultimate interest in whether unobserved emeralds are green. It is their ultimate interest in answering questions about green and blue that leads philosophers and others to accept (H5) rather than (H6). You do not accept (H5) because you prefer the answers that (H5) gives to other answers. Whether you want emeralds to be green or blue is irrelevant.

But what about the appeal to *simplicity* in believing (H5) rather than (H6)? Is that a case of wishful thinking? It is true that simpler hypotheses have pragmatic advantages over more complex hypotheses in that they are easier to use in accounting for data and in making predictions. So, you have a practical reason deriving from your interests to prefer believing simpler

hypotheses over believing more complex hypotheses. We (the authors of this paper) are divided as to whether believing a simpler hypothesis for this sort of practical reason is to engage in wishful thinking. A standard case of wishful thinking involves believing one hypothesis rather than another *because you want the first hypothesis rather than the second to be true*. Now, to prefer the simpler hypothesis because it is easier to use need not involve wanting that hypothesis to be true. A preference for believing *X* over believing *Y* is not the same as a preference for *X*'s being true over *Y*'s being true. Although this is not a standard case of wishful thinking, it may be just as bad.

So, we are still left with the second of the two questions about simplicity with which we began. The first question was "What makes one hypothesis simpler than another?" We have made a proposal about that. The second question is "Why, given that way of measuring simplicity, should you take the simplicity of a hypothesis to be any sort of indication of the truth of the hypothesis?" That is a deep question and we do not have the space for a full-scale discussion. All we can say is that we are unclear as to whether there is an independent source of information about what is likely to be true over and above the principles of reasoning we actually follow. Since people use simplicity to decide among hypotheses that are otherwise equally satisfactory, when we reflect on particular cases of this sort, the simpler hypothesis is likely to seem the most reasonable conclusion (unless we are temporarily skeptical).

Given a set of hypotheses that all account for the data, we do in fact take the simplicity of a hypothesis as making that hypothesis more likely than less simple alternatives. That is what we do, and we do not seem to have any reason to stop. Perhaps we are justified in continuing to use simplicity in this way—in the absence of a serious difficulty with our current practice and the absence of any reasonable alternative.

Bibliography

Ashcraft, Mark H., & Stazyk, Edmund H. (1981). "Mental addition: a test of three verification models." *Memory & Cognition, 9*, 185-196.

Goodman, Nelson (1965). *Fact, Fiction, and Forecast*, 2nd edition. Indianapolis: Bobbs, Merrill.

Groen, Guy, and Resnick, Lauren B. (1977). "Can preschool children invent addition algorithms?" *Journal of Educational Psychology, 69*, 645-652.

Harman, Gilbert (1986). *Change in View: Principles of Reasoning*. Cambridge, Massachusetts; M.I.T./Bradford Books.

Harman, Gilbert (1987). "(Nonsolipsistic) conceptual role semantics." In Ernest LePore (ed.), *New Directions in Semantics*. London: Academic Press, 55-81.

Harman, Gilbert, Bienkowski, Marie A., Salem, Ken, & Pratt, Ian (1987). "Measuring change and coherence in evaluating potential change in view." *Ninth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ.: Erlbaum, 203-209.

Sober, Elliot (1975). *Simplicity*. London: Oxford University Press.