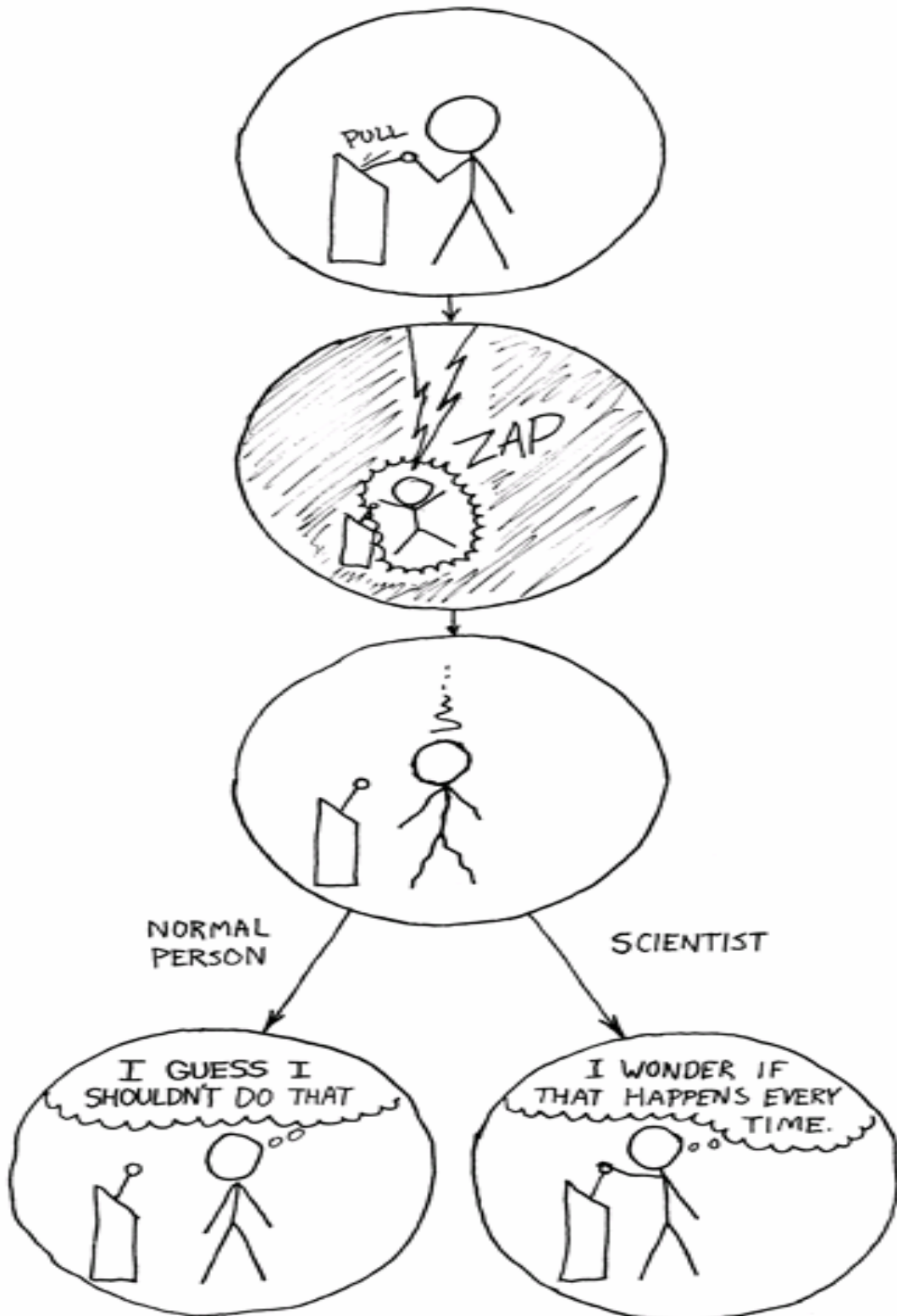


The Role of Simplicity in Theory Choice

Forster, Sober

What does it mean to be a scientist?



Questions about Simplicity

- Is a simpler theory more likely to be true?
Is a true theory more likely to be simple?
(Recall Whewell)
- Do scientists actually use one notion of simplicity as a criterion in theory choice?
What is simplicity? # of parameters?
- Does simplicity have an epistemic significance in its effects on theory choice?
- If a theory's simplicity does make it more probable in some sense, what is the reason for that? And how shall we argue for it?
- Can the correct role for simplicity in theory choice be captured in a Bayesian analysis?

Bayesianism

- 1) All scientific reasoning can be explained by reference to the axioms of probability. (Contra anti-methodists, like Kuhn, and methodists like Popper, Dempster-Shaferites, classical statistics)
- 2) The key to scientific inference is, in particular, Bayes' Thm. (Contra the above and classical statistics (Neyman-Pearson, Fisherian))
- 3) We assign probabilities to hypotheses. I.e. we have both prior and posterior probabilities of hypotheses. (Contra Popper, classical statistics)
- 4) Typically, though not exclusively, probabilities are degrees of belief of a rational subject. (Contra same as above, and objectivists about probability)

Some Forster Theses

- 1) Bayesianism works nicely as an account of curve-fitting (parameter estimation *given* a model).
- 2) Role of simplicity in theory choice is in model selection, and Bayesianism does not work at all for model selection.
- 3) Put simplicity in priors? Doesn't work, plus priors are junk anyway.
- 4) Bayesian dilemma: either admit only fit to data matters (can't solve model-selection problem regardless of simplicity), or face the reparametrization problem (i.e. language variance of likelihoods of families).
- 5) Continuum-many hypotheses \rightarrow all have zero probability, or else at least some can never be confirmed.

Compare to comparative likelihood method of theory choice, or to classical statistics.

Curve-fitting, Parameter Estimation

Suppose we are given a model,

$$y = cx^2 + bx + a,$$

and data in the form of ordered pairs, (x, y) .

You update on the data by calculating likelihoods for the various hypotheses with the different possible values of a , b , c , and find the most likely hypothesis in the model. Multiply $P(e/h)$ by $P(h)/pP(e)$ for each hypothesis and you have all the posterior probabilities of the various specific hypotheses. *Bayesianism good.*

Terminology: model = family of hypotheses

An individual hypothesis with particular parameters is a member of the model/family.

Model Selection

In science the role of simplicity is in choosing *between models*, e.g. epicyclic theory v. Kepler alternative, Kepler/Galileo v. Newton

Forster claim: Epicyclic theory has same fit as Kepler, but with a hysterical number of parameters (hence less simplicity). Simplicity is what decided this case.

Notice: Model/Family vs. Hierarchy

“Epicyclic Theory” is a hierarchy (unspecified number of parameters), whereas Kepler’s Laws give a model/family (specified number of parameters). If we set epicyclic theory to number of parameters of Kepler, it would have way worse fit.

→ Be careful to compare a model to a model, not a model to a hierarchy.

Model Selection

What is a model?

What do we choose when choose a model?

Consider the polynomial hierarchy:

⋮

$$H_4: \quad y = ex^4 + dx^3 + cx^2 + bx + a$$

$$H_3: \quad y = dx^3 + cx^2 + bx + a$$

$$H_2: \quad y = cx^2 + bx + a$$

$$H_1: \quad y = bx + a$$

$$H_0: \quad y = a$$

Each line identifies a model. The model is the family of hypotheses that take the specified form. The model/family is all of the hypotheses of that form with all possible values for $a, b, c, d, e \dots$

To assert the model is to assert the *disjunction* of all of those hypotheses that are members of the model. Model H_n says: $h_1 \vee h_2 \vee \dots \vee h_m \vee \dots$

Individual hypothesis h_m is also sometimes ⁸ called a “fitted model”.

Model Selection and Simplicity

⋮

$$H_4: y = ex^4 + dx^3 + cx^2 + bx + a$$

$$H_3: y = dx^3 + cx^2 + bx + a$$

$$H_2: y = cx^2 + bx + a$$

$$H_1: y = bx + a$$

$$H_0: y = a$$

Proposal for Bayesian understanding of role of simplicity: its role comes from scientists' preferences going into the inference.

Therefore, add a general **Simplicity Postulate** to constrain the prior probabilities of hypotheses.

Try: If H_1 is simpler than H_2 , then $P(H_1) \geq P(H_2)$.

H_1 above *is* simpler than H_2 .

But $P(H_2) > P(H_1)$ because all H_1 hypotheses are disjuncts of H_2 (with c set to zero), but not vice versa. Thus H_1 is strictly logically stronger.

Model Selection and Simplicity

Conclusion: If simplicity means fewer parameters, then there can be no Bayesian simplicity postulate because a *model* with fewer parameters is always less probable. Note that this is not generally true for single hypotheses.

(What if simplicity means something else?)

Why should we think simplicity means fewer parameters anyway?

Examples:

To fit the same data, Kepler needed fewer parameters than Ptolemaic astronomers.

Newton's theory was simpler in part because he had only one parameter for the earth's mass instead of two.

Einstein's theory of gravity was simpler in part because it used only one parameter for both inertial and gravitational mass (explained the empirical fact that they were always the same).

Model Selection and Simplicity

Bayesian: Oh, a simplicity postulate about the priors was a bad idea anyway.

1. It gives *no rationale* for scientists' prior preference for simplicity.
2. Just says one model has a head start going in, not why evidence confirms simpler model better.

So: try finding how simpler models might get *better confirmation* by the same data. Look not for higher posterior probability but that simpler model *increases* its probability more.

Suppose H_1 simpler than H_2 . Compare:
 $P(H_1/E)/P(H_1)$ to $P(H_2/E)/P(H_2)$.

This is the same as comparing $P(E/H_1)/P(E)$ to $P(E/H_2)/P(E)$. (Bayes Theorem)

Since $P(E)$ same in two cases, this is same as comparing:

$P(E/H_1)$ and $P(E/H_2)$, i.e. compare likelihoods of the models.

Likelihoods of Models are Fickle.

We want to know the fit to E, i.e. likelihood, $P(E/H)$, for a model H. This is a weighted average of the fit to E of each hypothesis in the model.

$$P(E/H) = P(E/h_1)P(h_1/H) + P(E/h_2)P(h_2/H) + \dots + P(E/h_n)P(h_n/H)$$

Note: $P(h_1/H) + P(h_2/H) + \dots + P(h_n/H) = 1$, so any probability taken from one must be given to others and vice versa.

Fact: we can reparametrize H so that h_1 , h_2 , etc. have different weights. This implies that H will have different likelihoods with respect to E depending on our choice of parametrization (language).

Likelihoods of Models are Language-dependent.

Model: the phenomenon is a circle. I.e.,

$$z = x^2 + y^2$$

Suppose we know that the circle's radius z lies between 3 and 100.

We can characterize any given circle by its radius z , or by its area πz^2 .

Consider the hypothesis $z = 15$ under the two parametrizations:

$$z = 15 \quad \text{vs.} \quad A = \pi(15)^2 = 706.5$$

Consider range of hypotheses in the model under each description:

$$\text{I.e., } 3 - 100 \quad \text{vs.} \quad \pi 3^2 - \pi(100)^2$$

In the first case 15 is one among a range of possibilities 98 units long, so its weight, $P(h_m/H)$, is $1/98$.

In the second case $\pi(15)^2$ is one among a range of possibilities $\sim 31,400$ units long, so its weight is $\sim 1/31,400$.

(Of course, the units are different in the two parametrizations, one of radius the other area.

Since this hypothesis's fit to the data is the same but its contribution to the average fit of the model is smaller, the likelihood of the model changes.

Replies to Language-Dependence of Models

- Exactly who *doesn't* have the problem that likelihoods of models are language-dependent?
- You have shown that there is no fit *per se* of a model H to data, but only fit relative to a parametrization of the model. What you need to show is how this affects the simplicity debate. Parametrizations are simpler or more complex (e.g., linear or polar for a circle). What if the likelihood of the model varied in the right way with the simplicity of its parametrization? (What would the “right way” be?)
- related idea: could there be a reason to pick the *simplest* parametrization? Notice: that kind of simplicity does not compete with fit.
- Ok, likelihoods of models aren't language-invariant. Why don't we Bayesians just use likelihoods of individual hypotheses? These *are* language-invariant. (They depend on the graph only, not the equation you write to express it.)
- Confirmation isn't language-invariant either, we know independently.
- Can we say: oh, use the same parametrization for every model you're comparing?

Likelihoods of Models not Invariant under Language* Transformations

Ok, ok. So, why don't we Bayesians just use likelihoods of individual hypotheses? These *are* language-invariant. (They depend on the graph only, not the equation you write to express it.)

- Where does that leave the Bayesian in accounting for the role of simplicity in theory choice?
- Hypothesis choice is not model selection. Model selection, supposedly, is where simplicity plays a role. (True?)
- Comparative likelihoods of hypotheses is not distinctively Bayesian. Problem?
- Hypothesis-choice becomes exclusively a matter of fit: there is no simplicity element anymore!

Maybe this means simplicity is *not* a properly epistemic part of theory choice after all?

Who exactly *doesn't* have this problem?

Confirmation is not Language-Invariant either.

A model for data that is roughly circular:

$$H_1: z = ax^2 + by^2$$

A model for data that is green emeralds:

M_1 : All emeralds are _____.

Consider models H_2, M_2 :

$$H_2: z = ax^2 + by^2 + cy^{14}$$

M_2 : All emeralds are _____ or _____.

Best-fit H_1 to the roughly circular data.

Reparametrize H_2 so that its best-fit fits the data to the same degree as H_1 .

Best-fit M_1 to the green-emerald data to get:

h_1 : All emeralds are green.

Fill in M_2 to fit to the same degree:

m_2 : All emeralds are either observed before t and green or not observed until after t and blue.

x is *grue* if and only if x is either observed before t and green or not so observed and blue.

Thus: M_1 is "All emeralds are green." M_2 is "All emeralds are grue."

Intuition says that the green emeralds confirm "All emeralds are green" and not "All emeralds are grue."
But where's the asymmetry?

Confirmation is not Language-Invariant either.

$$H_1: z = ax^2 + by^2$$

$$H_2: z = ax^2 + by^2 + cy^{14}$$

M_1 : All emeralds are _____.

M_2 : All emeralds are _____ or _____.

All emeralds are green. vs.

All emeralds are grue.

Green and grue are syntactically symmetric. (If you had started with grue, then green would be the one broken.)

We find green natural and grue not. We think green is confirmed by the data, grue not. But the fit and syntax are symmetric. The difference is *language*.

→ Confirmation is not language-invariant. The strength of confirmation between h and e varies with your choice of language.

This match suggests the Bayesian is modeling confirmation *right*. Language must be fixed before fit has anything meaningful to say about confirmation. Language will be a preference given by the priors.

Confirmation is not language-invariant either--simplicity

We might also have reacted to grue and green by saying that green is obviously SIMPLER, and thereby better confirmed.

But that impression also comes from the image of grue being “broken”, with two parts. If you start with grue, green is the predicate that’s broken with two parts.

Intuition tells you that green is simpler than grue, but syntax does not show that difference.

What simplicity is for confirmation cannot be captured syntactically.

But notice on what basis we were saying H_1 is simpler than H_2 , and M_1 simpler than M_2 : syntax (counting parameters). The fact that grue and green are syntactically symmetric says that syntax is not where simplicity lives.

$$H_1: \quad z = ax^2 + by^2$$

$$H_2: \quad z = ax^2 + by^2 + cy^{14}$$

M_1 : All emeralds are _____.

M_2 : All emeralds are _____ or _____.

Revised Bayesian View of Simplicity

We wanted to see whether the simpler models were better confirmed by looking to see if the simpler models had higher likelihoods, i.e. better fit to the data. (Not likely, but let's ignore that.)

We found that the values for the likelihoods of models could be made whatever we wanted by changing the language in which the models are expressed (the parametrization). *Bad.*

We looked back at confirmation (probability raising) of individual hypotheses, and remembered that that is *also* language dependent (not semantically invariant). *Hmm.*

We saw in the grue/green phenomenon that green is only simpler than grue, even syntactically, if you *start* with the green language. If you start with a grue language, green is syntactically more complicated. So, it's not just the confirmation (likelihood) that depends on language, it's the simplicity of the model itself. *Hmm.*

→ You have to choose a language before you get determinate answers as to fit *or* comparative simplicity of a model.

Revised Bayesian View of Simplicity

The Bayesian now seems to be able to turn this language-variance to her favor. She can say that simplicity judgments are determined by peoples' prior probabilities, but not by prior probabilities of claims *about* simplicity.

Rather, the subject simply comes into the model selection scenario with language preferences. The subject comes in already speaking a particular language.

So, the language preferences you come in with (grue vs. green, polar vs. cartesian coordinates) tell you which language to use, and that determines what counts as simple.* The language also fixes likelihoods of models uniquely, so you can't make them whatever you want.

Revised Bayesian View of Simplicity

Ok, but can this view explain our simplicity intuitions?

On this view it's not simplicity per se that acts as a criterion in model choice. Rather, e.g., having started with the green language makes green the model with one parameter and grue the model with two.

Say those two models fit the data the same. You *think* that you choose the green model because it's simpler, meaning it has only one parameter. But what's really happening is that you're living in your language. Since who's simpler (by # of parameters) is determined by the language choice, to say that green is simpler, is just to say that you will use the language you started with (the green language).

This would suggest that simplicity doesn't have a real role in theory/model choice. It's an epiphenomenon we mistake for a criterion.

Will choice of language itself be able to be a choice based on fit of the language with our experience?

Revised Bayesian View of Simplicity

It is easier to see that this view lacks a role for simplicity per se if we ask what we do when two models of different simplicities have different degrees of fit. It looks like the Bayesian has to say that the choice is determined by fit alone. → pressure to choose the most complex model, which is not what we do.

It looks like the two things the language choice determines (which model is simpler and what the fit level of a model is), are independent of each other, so that the situation above can arise, and a model's ending up counting as simpler doesn't mean it will have better fit. It would have been nice if it had meant that—both factors would have had a role in theory choice.

But is this a problem, or just the way it really is? Could this account be a discovery that we are wrong to think simplicity has an epistemic role in that final choice? That depends in part on whether this account gives a plausible error theory, that is, explanation of why we thought simplicity had that role.

Can the Bayesian still allow choices between models that are based on simplicity? (Yes. Expected utility.)

AIC on simplicity, unification, adhockitude

If the goal is predictive accuracy (PA), then both fit and simplicity play a role:

$$\text{PA}[M] = (1/N)[\log\text{-likelihood}(L(M)) - k]$$

$L(M)$ is the likeliest member of M . Which member that is, and likelihood of that member is language invariant. So in evaluating fit, they avoid the problem the Bayesians had.

$(1/N)$ makes it a per-datum quantity so that different sized data sets can be compared.

Here fit and simplicity are in competition. Complexity will give you better fit, but go too complex and you'll have less predictive power. (fitting to noise)

You should maximize simplicity as far as possible as per this equation to the extent that you want predictive accuracy.

And that's the epistemic role simplicity plays in theory choice. They apply this to unification, Ockham's Razor, ad hoc hypotheses, the kitchen sink, etc.

AIC applied to unification

Intuition: Unified theories are epistemically preferable to non-unified.

AIC: Yes, that's because these models have fewer parameters, which avoids overfitting, so they are more predictively accurate.

So, all we need to show is that unification is plausibly understood as a matter of fewer parameters.

Example, pp. 13-14: D_1 , D_2 , M_1 ,
 M_2 , M_a

AIC on simplicity, unification, Ockham's Razor, adhocitude

We explain simplicity, and unification, etc.
here by our desire to:

maximize **predictive accuracy**
via minimizing **overfitting**
via **penalizing** models for **more**
parameters.

of parameters is a syntactic criterion.

There are cases (grue) where simplicity
cannot be captured syntactically.

How can *counting* work here to capture
simplicity?

Why is the criterion not arbitrary?

Is the # of parameters of a model
manipulable?

Does the number of parameters of a model
vary with language?

AIC and the Arbitrariness of k

The Subfamily Problem – Recipe for Disaster

Using AIC, choose the best-fitting curve $L(M)$ from the model/family with the highest estimated predictive accuracy.

Construct new families of curves by step-by-step fixing parameters at the values in $L(M)$.

With each step you construct a family with one less adjustable parameter. Finally you end up with a zero-parameter model with one member. I.e., The new model has zero penalty, so it becomes more choiceworthy than other models that are really simpler. This cannot be good when we didn't change the curve (graph) at all.

We can make the curve in this model win when the AIC did not intend it to win.

Maybe there is an answer to the subfamily problem, pp. 19-21, but there is a bigger₂₆ problem. →

of Parameters is *Language-Dependent*

By shifting language, we can make a model have fewer parameters. Illustration:

Consider a model,

$$z = ax^2 + by^2$$

This model has three parameters, z , a , and b . There is a way of expressing three numbers as one number, a computable function, f , that takes an ordered sequence of numbers (z, a, b) to the one number: $2^z \cdot 3^a \cdot 5^b = c$. Since every number has a unique prime factorization, the inverse function, f^{-1} , also exists (also computable). It takes c to (z, a, b) . Define f^{-1}_1 as f^{-1} composed with a function that takes that ordered triple and spits out its first member, etc.

Now we can rewrite the model as a model with only one parameter:

$$f^{-1}_1(c) = f^{-1}_2(c)x^2 + f^{-1}_3(c)y^2$$

What does this mean for the use of AIC to understand the role of simplicity in model selection?

Why would arbitrariness in the # of parameters matter?

But let us not despair.

The problem: how could these criteria—Bayesian or Akaikian—tell us anything when which model they tell us to choose is relative to language? Change our language and we should change our choice.

So the question is just pushed back: *why should we choose this or that language?*

We should view both the way everyone views Bayesianism: constraints of rationality telling you what to do given where you start. Note that Akaike IC is just an *estimator*. It cannot possibly tell you what is true, only what would be an unbiased estimate. What we have discovered is only that what you should estimate depends on what language you speak. Since no estimation, even an unbiased one, is guaranteed to be true, the fact that coming from different languages they give different answers doesn't make a contradiction.

Goals of science: probability of truth vs. estimated predictive accuracy

Bayesian goal: that the chosen hypothesis be the most probably true.

Akaike goal: that the chosen hypothesis be closest to the truth.

How they diverge: “average number of children in American family is 1.6”. But no family has 1.6 children.

Forster and Sober: no curve-fitting procedure will give us exactly the true curve. (Why?) Better to try for closeness to truth.

Empiricism, Rationalism, and Realism

Empiricism: all of our knowledge comes from experience.

Observed dependence of theory choice on simplicity a problem for empiricists: it's not part of fit to data so it's not evidential, so whatever it contributes to our choices between theories does not bring us knowledge.

Rationalist can say: there is an a priori truth that we know by insight that simple theories are more likely to be true.

Note: simplicity isn't a factor *only* in choices scientists make between big theories. (If it were, what could we say?) It's there in choices between curves in fitting to data. Those choices we make aren't about what's true either if simplicity has no evidential significance.

Dilemma: Either accept rationalism ("realism") and simplicity as an a priori sign of truth, "or embrace anti-realism" (p. 28)

Anti-realism: Our theories aren't getting at the truth, or *we have no reason to believe* our theories are getting at the truth.

Why would we have to be anti-realists on the second horn of the dilemma?

AIC to the Rescue

Simplicity tends to be a sign of predictive accuracy.

Choices that promote predictive accuracy are epistemic in so far as they promote closeness to truth *of our future predictions of data*.

Therefore, simplicity tends to be a sign of something epistemic.

Nevertheless, simplicity is not *per se* an epistemic criterion. Why?

Also, this gives us no reason to think that simplicity is a sign of *truth* of a theory/model. ³¹